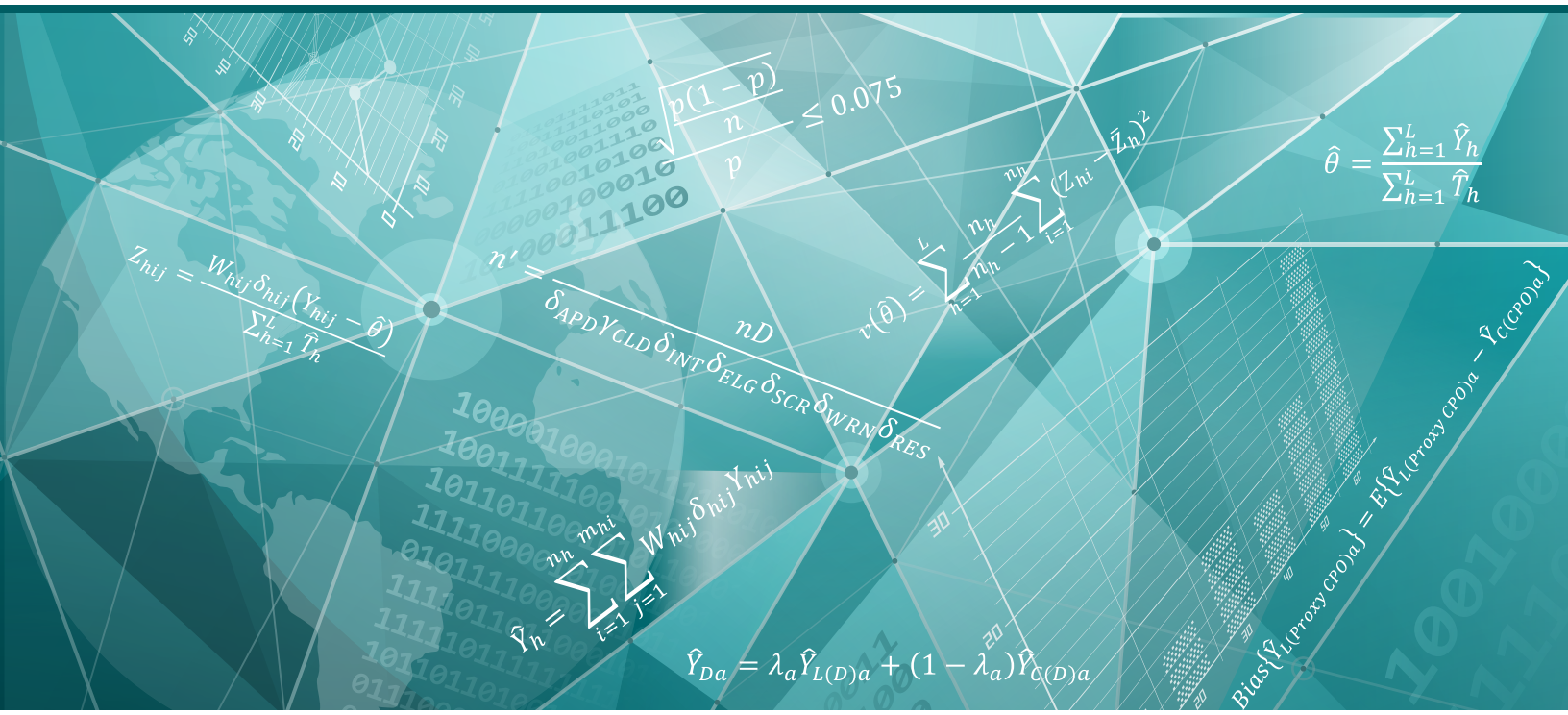
NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 212

April 2025



Creating Synthetic Data for Complex Surveys Using the Research and Development Survey: A Comparison Study

Data Evaluation and Methods Research

U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION

NCHS reports can be downloaded from: https://www.cdc.gov/nchs/products/index.htm.

Suggested citation

Zhang G, He Y, Oganian A, Cai B. Creating synthetic data for complex surveys using the Research and Development Survey: A comparison study. Vital Health Stat 2. 2025 Apr;(212):1–17. DOI: https://dx.doi.org/10.15620/cdc/174586.

For sale by the U.S. Government Publishing Office
Superintendent of Documents
Mail Stop: SSOP
Washington, DC 20401–0001
Printed on acid-free paper.

NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 212 April 2025

Creating Synthetic Data for Complex Surveys Using the Research and Development Survey: A Comparison Study

Data Evaluation and Methods Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
April 2025

National Center for Health Statistics

Brian C. Moyer, Ph.D., *Director*

Amy M. Branum, Ph.D., *Associate Director for Science*

Division of Research and Methodology

Morgan S. Earp, Ph.D., *Acting Director*

John R. Pleis, Ph.D., *Associate Director for Science*

Contents

Detailed Tables

Creating Synthetic Data for Complex Surveys Using the Research and Development Survey: A Comparison Study

by Guangyu Zhang, Ph.D., Yulei He, Ph.D., Anna Oganian, Ph.D., and Bill Cai, M.Sc.

Abstract

Background

Synthetic data has been gaining popularity in many fields as an approach to retain data utility (the validity of inference using synthetic data) and protect confidentiality. However, creating synthetic data for complex surveys remains a challenge.

Methods

This research compared three approaches to incorporate survey design information (stratification, clustering, and sampling weights) during the synthetic data-generating process using the Research and Development Survey (RANDS), a series of primarily web surveys conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. Both parametric (logistic and linear regression models) and nonparametric (classification and regression trees [CART]) methods were used to create synthetic data. Data utility and disclosure risk were evaluated via confidence interval overlap, propensity score measurement, and average matching probability for re-identification.

Results

Using the original survey design information as predictors during the synthesis process improved data utility for the parametric method. However, the nonparametric method yielded results with better data utility but slightly higher disclosure risk.

Keywords: survey design information • parametric and nonparametric methods • Research and Development Survey (RANDS)

Introduction

Federal agencies, such as the National Center for Health Statistics (NCHS), collect, analyze, and disseminate information for statistical purposes. They conduct national surveys and release microdata (data containing information about individuals) for public use. To protect survey participants' confidentiality, statistical disclosure limitation (SDL) techniques have been used to de-identify data (1–4). Some commonly used SDL methods include deletion of data, data coarsening, top (or bottom) coding, adding noise to data, data swapping, and using synthetic data (5–11).

Deleting records of individuals vulnerable to disclosure protects their confidentiality; however, information about these individuals is lost. For example, individuals with very high income or uncommon characteristics, such as rare diseases, are removed from the public-use data. As a result, the released data may lead to biased estimates of specific characteristics of the study population. Data coarsening creates coarsened categories on variables with sensitive information such that detailed information is not released. For example, income categories are released for public use instead of exact income to protect sensitive income information. This prevents disclosure but also precludes detailed statistical analysis on these variables. Top (or bottom) coding truncates data at certain cut-off values; for example, the age of a person 85 years or older is reported as 85 years old. Adding noise to the data adds a random number to the original data values, so the data are not as accurate as the original data (5,12). However, this increases variance estimates, which may lead to an increased type II error. Data swapping methods exchange data values for pairs of individuals (6,13). As a result, the original data values are released for public use. However, because data values are switched between individuals, the marginal distribution of the swapped variable(s) (such as survey data with sampling weights associated with different subjects) and conditional inferences of the swapped variable(s) with respect to other variables may change.

Use of synthetic data may offer another option. Partially synthetic data are generated by replacing some of the observed values, whereas fully synthetic data are derived by replacing all of the observed values with sampled values from the approximate probability distributions of variables estimated from the original data (1,2,7,8,11). The goal of synthetic data is to preserve essential statistical features and

variable relationships of the original data such that statistical inference based on the synthetic data is close to that of the original data. In certain circumstances, this method may be preferable for protection because only some or none of the original data are released. However, sophisticated statistical methods are required to model the joint distribution of all variables. Departure from the original distributions may lead to misleading results. Moreover, it could produce records with identical or almost identical values as records in the original data if the model to create synthetic data is saturated; that is, the model fits the data perfectly and consequently yields predictions very close to the truth. As a result, synthetic data are not guaranteed to be risk-free.

The SDL techniques mentioned above can be used alone or together. Combining SDL techniques may lead to better data utility and better protection (14). In practice, a trade-off always exists between data utility (the validity of inference using synthetic data) and data confidentiality (4,15). A good SDL technique retains data utility and, in the meantime, protects confidentiality. Yet ill-intentioned intruders may still identify an individual's identity and violate participants' privacy. Usually, individuals with a unique combination of certain observable traits are vulnerable to identification. In recent years, the synthetic data approach has been gaining popularity in many fields as an approach to retaining data utility and protecting confidentiality. However, creating synthetic data for complex surveys remains challenging due to the complex survey design features, the high dimensionality of the survey data with different distributional forms, and unique data structures such as skip patterns and missing data.

This report studies the impact of incorporating survey design information (strata and primary sampling units [PSUs], referred to in this report as design variables; and sampling weights) when synthesizing sample survey data using parametric and nonparametric methods. The Research and Development Survey (RANDS) (https://www.cdc.gov/nchs/rands/) (16–18) was used to create synthetic data. RANDS is a series of surveys conducted by NCHS that was designed to explore the use of recruited web panels to collect information on national health outcomes and to supplement NCHS's question-response evaluation and statistical research. Nine rounds of cross-sectional surveys (RANDS 1 to RANDS 9) and three rounds of RANDS During COVID-19 surveys were completed by the spring of 2024. RANDS During COVID-19 surveys, a special iteration of RANDS, collected timely information on COVID-19-related health outcomes from U.S. adults (19). COVID-19 Survey Round 3, conducted in the summer of 2021 and the most recent RANDS at the time of the study, was used for this report.

Methods

A General Procedure to Create Synthetic Data

To create synthetic data, the joint distribution of the variables to be synthesized can be estimated using a series of sequential conditional distributions. Let $Y_1, Y_2, \ldots, Y_k$ denote k variables to be synthesized, and let X denote a vector of variables not to be synthesized but to be used as predictors when synthesizing $Y_1, Y_2, \ldots, Y_k$. The synthesis process for $Y_1, Y_2, \ldots, Y_k$ is based on the following conditional distributions: $f_1(Y_1|X), f_2(Y_2|Y_1, X), \ldots, f_k(Y_k|Y_1, Y_2, \ldots, Y_{k-1}, X)$. If X is empty, the variable to be synthesized first, Y_1, can be generated by random sampling with replacement from the marginal distribution of $Y_1, f_1(Y_1)$, estimated from the original data.

Incorporate Survey Design Information to Create Synthetic RANDS Data

RANDS During COVID-19 collected data on loss of work, use of telemedicine, and access to health care during the COVID-19 pandemic (for more information, see: https://www.cdc.gov/nchs/covid19). Table 1 shows the selected demographic and health variables from RANDS During COVID-19 Round 3 for data synthesis. Variables that were common across all rounds of RANDS and had lower percentages of missing data were selected. The statistical inference of these variables (analytical variables) is used to evaluate different synthetic methods. For sample survey data such as RANDS, the survey design information, which includes strata, PSUs, and sampling weights, is an essential part of the survey. The design variables (strata and PSUs) are used to create meaningful subgroups and select representative units from the population. The sampling weights account for differences in selection probabilities and nonresponse rates and adjust for under- or over-representing specific groups within the sample. Using the sampling weights ensures that the estimates accurately represent the characteristics of the population. Both design variables and weights must be used for statistical analyses to derive valid variance and point estimates.

This research compared three approaches to incorporate survey design information in creating synthetic data. Approach 1 included design variables and sampling weights during the data synthesis process (variables were synthesized along with the analytical variables) (Table 1). Approach 2 synthesized the sampling weights along with the analytical variables during the synthesis process. While the design variables were not synthesized, they were used as predictors when synthesizing other variables. Approach 3 synthesized only the analytical variables without synthesizing any design variables or weights; however, the design variables and weights were used as predictors when synthesizing the analytic variables. Table 2 shows whether the design

information was synthesized, and for which components, during the data synthesis process for Approaches 1–3.

The joint distribution of the variables to be synthesized (the analytical variables in Table 1 and the survey design information in Table 2) can be estimated using a series of sequential conditional distributions described above. Both parametric and nonparametric methods can be used to estimate these conditional distributions. In this report, the performances of both methods were compared. For this analysis, the parametric method included a logistic or multinomial logistic model for categorical variables and a linear regression for continuous variables, implemented using the IVEware SYNTHESIZE module (20) and the R synthpop package (21). The nonparametric method included classification and regression trees (CART) methods, implemented using the R synthpop package (21–23). In CART, data are grouped into smaller homogeneous clusters by binary recursive partitioning of the predictors. At each partitioning step, a predictor and a split point are chosen, and all or a portion of the data set is split into two groups based on the split point of the predictor. This process repeats until a predetermined best-fit criterion, such as the residual sum of squares or deviance, is achieved. Then an observed value is drawn using simple random sampling from the final clusters as the synthetic value.

Both the R synthpop package and the IVEware SYNTHESIZE module were created for data synthesis purposes, and they both have multiple modeling options to create synthetic data. In this report, these two software packages were used for the parametric method so that the results from the two packages can be compared. For example, when the two packages used the same regression models, the results from both were expected to be similar. Consequently, a comparison of the results can serve as a model check. In addition, different modeling options can be used for the continuous sampling weights. In this study, the square root transformation was used for R, and the bounds function, a feature only available in IVEware that draws values from truncated predictive distributions, was used for IVEware; both options guaranteed that the synthesized sampling weights would be positive values.

The order of the variables to be synthesized affects data synthesis. Because the categorical variable strata had 71 levels, it needed to be the first to be synthesized by simple random sampling with replacement from the marginal distributional of strata, estimated from the original data. IVEware does not have this option. Instead, it synthesizes strata from a multinomial logistic regression model, which does not perform well for categorical variables with many categories, so IVEware was not applied to Approach 1. Thus, this study synthesized the design variables before the analytical variables (Approach 1 using R and Approach 2 using R and IVEware), and the order to synthesize the design variables is strata, PSUs (a variable nested within strata), and then sampling weights. All analytical variables included in this study were categorical variables with 2 to 6 levels, where logistic and multinomial logistic regression models perform well for the parametric approach. Without special patterns among these analytical variables (such as nested data structures or skip patterns), the storage order of the variables in the data set was used to synthesize them.

Analysis of the Synthesized Data

For each approach described above, five synthetic data sets were created. Marginal proportions of all the analytic variables were derived for each synthetic data set, and survey design features were incorporated for variance estimation. For Approach 1, the synthesized design variables and sampling weights were used in the statistical analysis; for Approach 2, the original design variables and synthesized sampling weights were used in the analysis; and for Approach 3, the original design variables and sampling weights were used in the analysis. Results from each of the five synthetic data sets were combined using the combining rules of multiple synthetic data sets described later in this report (2,8,24). Statistical analyses were conducted using SAS (25).

The combining rules of multiple synthetic data sets

Let Q be a parameter of interest (for example, population mean, proportion, regression coefficients, etc.), let q_i be the point estimate of Q, and u_i be its estimated variance from the ith synthetic data, where i = 1, 2, …, m, where m is the total number of synthetic data sets. The combined point estimate from the multiple synthetic data sets, $\hat{q}$, is

$$\hat{q} = \frac{1}{m}\sum_{i=1}^{m} q_i$$

The estimated variance of this point estimate consists of two components. The first component, "within synthetic variance," is the mean of variances across multiple synthetic data sets:

$$\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i$$

The second component, "between-synthetic variance,"

$$\bar{b} = \frac{1}{m-1}\sum_{i=1}^{m} (q_i - \hat{q})^2$$

is the variation due to differences across m sets of synthetic data. The total estimated variance of the point estimate $\hat{q}$ is

$$\hat{V} = \bar{u} + \bar{b} / m$$

Furthermore, it was shown in Reiter (26) that, approximately,

$$(\hat{q} - Q) \sim t_{df}(0, \hat{V})$$

where the degrees of freedom (df) is

$$df = (m-1)\,(1 + \bar{u} / (\bar{b} / m))^2$$

Confidence intervals (CIs) of $\hat{q}$ could be calculated based on the *t* distribution for the synthetic data.

CI overlap

CI overlap measures the overlap of the CIs derived from the original data and the synthetic data (27,28). The overlap (or lack thereof) of CIs does not imply or test statistical significance. Instead, the higher the overlap, the better the data utility. Let U_{ori} and L_{ori} be the upper and lower bounds of the confidence interval for an estimate derived from the original data, U_{syn} and L_{syn} be the upper and lower bounds of the CI for the same estimate derived from the synthetic data, and U_{over} and L_{over} denote the upper and lower bounds of the overlap of the CIs derived from the original and synthetic data sets for the estimate. The CI overlap is

$$\text{CI overlap} = \left(\frac{U_{over} - L_{over}}{U_{ori} - L_{ori}} + \frac{U_{over} - L_{over}}{U_{syn} - L_{syn}}\right) / 2$$

When there is no overlap, CI overlap = 0.

Propensity score measurement

Propensity score measurement is a method used to distinguish the synthetic from the original data (28,29). The main steps of deriving the propensity measurements are: 1) stack the original data and the synthetic data and for each record include an indicator for the data source; 2) for each record in the stacked data set, calculate the propensity of being in the original data from a logistic regression model that includes the synthesized variables as predictors; 3) compare distributions of the estimated propensity scores. Assuming the original and synthetic data have the same number of records, if the two data have similar distributions, the chance for any chosen record to belong to the original or the synthetic data would be similar (around 50%), and the estimated propensity scores would be close to 0.5 for all records. On the other hand, if the two data sets are completely distinguishable, the estimated propensity score for each record would move away from 0.5. For example, the estimated propensity scores for records from one data set tend to be close to 1, and the estimated propensity scores for records from the other data set tend to be close to zero.

Because the survey design information was synthesized in this study (Approaches 1 and 2), the survey design variables and sampling weights were used to estimate the propensity scores from a weighted logistic regression model using the SAS surveylogistic procedure (25), and the variables synthesized (Table 1) were included as predictors. Because multiple synthetic data sets were created, the first copy of the synthetic data set was used for this analysis.

Analysis of sampling weights from Approach 2

For Approach 2, using the parametric method, both the R synthpop package and the IVEware synthesized the sampling weights from a linear regression model; however, the square root transformation of the sampling weights was used for the R synthpop package, while the IVEware SYNTHESIZE module used the bounds function for the sampling weights. For the nonparametric method (CART), the sampling weights were synthesized using random draws of the observed values from the homogeneous clusters created from the CART model. The synthesized sampling weights from different modeling options may lead to differences between the synthesized data. Thus, quantiles and means of synthetic sampling weights were compared with the original sampling weights. Departure from the original sampling weights suggests low data utility.

Disclosure risk analysis via average matching probability

The synthetic data approach applied to complex survey data can protect an individual's confidentiality while preserving data utility. The previously described analyses for this report focused on data utility. This analysis focused on disclosure risk using the average matching probability (30–32). Synthetic data are generated based on statistical models. When the synthetic data set is very similar to the original data set (for example, the synthetic data set contains records with identical or almost identical values as records in the original data), the chance of disclosure is greater. Suppose an intruder knew specific information about a target person who participated in a survey. These variables could be linked to the synthetic data sets to potentially identify the target individual. Let P_match be the average matching probability, which can be derived as

$$\text{P\_match} = \frac{\sum_{i=1}^{m} n_i / Nmatch_i}{m}$$

where n_i is the number of correct matches for synthetic data *i* (a correct match means the match is a true match [all variables of a matched case are the target person's real values]). For high-dimensional data, n_i would most likely be 0 (not a match) or 1 (a true match). The chance of n_i being greater than 1 is small as the likelihood of two records in synthetic data with identical measurements across all variables is expected to be small. However, in this study, a subset of the survey variables was used for data synthesis, so having common measurements across all 12 variables is possible; consequently, n_i could be greater than 1. $Nmatch_i$ is the number of records fulfilling the matching criteria in the synthetic data *i* (the total number of records in the synthetic data *i* with identical information the intruder possessed); *m* is the total number of synthetic data sets.

For example, suppose the intruder knows the target person is a 50-year-old White female with a Ph.D. Checking the synthetic data, if the intruder finds two White 50-year-old women with a Ph.D., then $Nmatch_i = 2$. Suppose neither of the two subjects is the target person (that is, the remaining variables are not the real values of the target person); then it is a false match and $n_i = 0$, and the intruder would have a

0% chance of identifying the target person. If one of the two subjects is the target person (that is, the remaining variables for this record are the same as those of the target person), then it is a true match and n_i = 1, and the intruder would have a 50% chance of identifying the target person. If all of these two matched records have identical measurements for all 12 variables like the target person, then n_i = 2, and the intruder would have a 100% chance of identifying the target person. The overall matching probability can be derived after calculating n_i and $Nmatch_i$ for all synthetic data sets.

Results

Results of CI Overlap

Table 3 shows the means of CI overlap across all synthesized analytical variables. Synthesizing survey design information during data synthesis reduced data utility, but the impact was only shown in the parametric method. For Approach 1, where all survey design information was synthesized, the parametric method implemented using the R synthpop package yielded a mean CI overlap of 0.73, while the nonparametric method yielded a mean CI overlap of 0.88. For Approach 2, where design variables were not synthesized but the sampling weights were synthesized, using the square root transformation on the sampling weights (R synthpop) or using the bounds function for the sampling weights (IVEware) yielded similar means of CI overlap (0.71 and 0.70, respectively); the nonparametric method yielded a mean CI overlap of 0.89. For Approach 3, where none of the survey design information was synthesized, the two parametric methods yielded results (0.88 using the R synthpop package, 0.85 using IVEware) close to the nonparametric method (0.88 using the R synthpop package).

Results of Propensity Score Measurement

The means of the estimated propensity scores and the range of the propensity scores are in Table 4. The means of the propensity scores for all three approaches were the same (0.50) using either the nonparametric method or the parametric method using the R synthpop package. The means of the propensity scores from the parametric method using IVEware differed more from 0.50 compared with the other methods (0.56 for Approach 2 and 0.51 for Approach 3). The estimated propensity scores would be close to 0.50 for all records if the original and synthetic data had similar distributions. Thus, a narrower range around 0.50 suggests the two data sets are closer. In this analysis, the ranges of the propensity scores were narrower for the nonparametric method compared with the parametric methods (IVEware was not applied to Approach 1). For example, for Approach 1, the range of the propensity scores was 0.42 to 0.61 for the nonparametric method (CART in R), while the range of the propensity scores was 0.34 to 0.70 for the parametric method (regression in R). Similar patterns were shown in Approaches 2 and 3. The results suggested that the distributions of the synthetic data created from the nonparametric method were closer to the distributions of original data than those created from the parametric methods, especially in Approaches 1 and 2.

Results of Sampling Weight Comparisons From Approach 2

For Approach 2, with the parametric method, the only difference between the R synthpop package and the IVEware was how the continuous variable, the sampling weights, was synthesized. If the original and synthetic data were similar, the distributions of the original and synthesized sampling weights would be close to each other. In this analysis, quantiles and means of the original and synthetic sampling weights were compared (Table 5). The range of the original sampling weights was from 0.01 to 17.65. The parametric methods yielded a narrower range of the sampling weights using both the R synthpop package and the IVEware SYNTHESIZE module (0.00 [rounded] to 5.25 and 0.00 [rounded] to 5.34 for R and IVEware, respectively). In contrast, the range of the synthesized sampling weights using the nonparametric method was about the same as that of the original data. The quantiles of the sampling weights using the parametric methods departed from the original data, while the quantiles of sampling weights using the nonparametric method were closer to those of the original data. The differences in the synthesized sampling weights led to better performance of the nonparametric method (R CART), including a higher mean of CI overlap and a narrower range of propensity score measurements, compared with the parametric methods (Tables 3 and 4). For the parametric methods, although the ranges of the synthesized sampling weights were similar using IVEware and R, the mean and median of the synthesized sampling weights using R (regression) were closer to those of the original data than those using IVEware (mean: 1.27, 1.00, and 1.00 for IVEware, R, and the original data, respectively; median: 1.15, 0.83, 0.69 for IVEware, R, and the original data, respectively), which corresponded to a mean of propensity scores higher than 0.50 using IVEware for Approach 2 (Table 4).

Disclosure Risk Analysis via Average Matching Probability

Five target cases were selected based on the number of records with identical information (records with the same age, sex, education, etc., across all 12 analytical variables) (Table 6). For example, the first target case was a sample unique case (no other record with the same values across all 12 variables as the target case was seen in the original data); and the last target case is a case where there were 20 subjects in the original data with identical measurements across the 12 analytical variables. For each target case,

the intruder is assumed to know only the values of four variables: age, race and ethnicity, education, and sex. Linking these four variables to the synthetic data, the intruder would be able to identify cases with identical values for the four variables, although the linked cases may or may not be a true match.

Table 6 shows the matching probability of each target case and the mean matching probability (expressed as percentages) across the five target cases for Approach 1. The parametric method implemented using R synthpop had a low matching probability when the target case was a unique sample record (0.29% for target case 1) or when there was a small number of records with identical values (1.11% for target case 2). When the number of records with values identical to the target case increased (10 or more records with identical values in the original data), the matching probability of the parametric method also increased (4.16%, 2.62%, and 4.53% for target cases 3–5, respectively). Similar patterns were shown in the nonparametric method. However, the increasing trend was not monotone, because when $Nmatch_i$ increased in the synthetic data, "n_i" also tended to increase. Thus, the matching probability decreased for target case 4 (15 records with identical information in the original data) and then increased for target case 5 (20 records with identical information in the original data) for both the parametric and nonparametric methods. The mean matching probability was slightly higher for the nonparametric method than the parametric method (2.94 compared with 2.54). Approaches 2 and 3 showed a similar pattern as Approach 1 (Tables 7 and 8).

Although IVEware performed worse than the R synthpop package regarding data utility and propensity score measurement, the overall matching probability based on IVEware was the lowest for Approaches 2 and 3 (2.60% and 2.40%, respectively, for IVEware; 2.82% and 3.01%, respectively, for the R synthpop package using the parametric method; and 3.11% and 3.24%, respectively, for the R synthpop package using the nonparametric method) (Tables 7 and 8). The results of the disclosure risk analysis were in alignment with the trade-off between data utility and disclosure risk.

Discussion

Synthetic data is a promising method to balance the need for keeping data utility and protecting privacy for public-use data. However, generating synthetic data for national surveys is challenging because of complex survey designs, high dimensionality of the original data with different distributional forms, and complicated data structures. This report focused on incorporating survey design information during the synthetic data-generating process.

Three approaches incorporating survey design information to create synthetic data using both parametric and nonparametric methods were compared. The nonparametric method (CART) yielded better data utility than the parametric methods in Approaches 1 and 2, where all or part of the survey design information was synthesized along with the analytical variables. When none of the design information was synthesized (Approach 3), both the parametric and nonparametric methods yielded similar data utility. The poorer performance of Approaches 1 and 2 using the parametric method might be related to the unusual data format of the survey design information, the sampling weights, which was a truncated continuous variable. The parametric method was sensitive to the underlying model assumptions, and violation of model assumptions would create synthetic values dissimilar to the original data, leading to low CI overlap. On the other hand, the nonparametric method (CART) was more robust for unusual data types, and synthesizing the survey design information did not lead to low data utility. Keeping original survey design variables while using them as covariates to synthesize the analytical variables (Approach 3) helped data utility for the parametric approach, with data utility measurements improved to the level of the nonparametric method.

Although the nonparametric method yielded better data utility, it also increased disclosure risk. The disclosure risk was related to the amount of information to be synthesized (or not synthesized). In this study, it was related to whether the survey design information was used as predictors (instead of being synthesized). The more the survey design information was used as predictors, the higher the chance of creating synthesized records identical to the original data, leading to a higher disclosure risk. The mean matching probabilities were higher in Approach 3 than in Approaches 1 and 2 for both the nonparametric method (CART) and the parametric method using the R synthpop package. However, unlike the parametric method, synthesizing survey design information using the nonparametric method impacted more on disclosure risk but less on data utility because the nonparametric method was less sensitive to unusual data types of survey design information. Thus, synthesizing survey design information using the nonparametric method may be an option to balance the trade-off between data utility and disclosure risk. Moreover, applying parametric and nonparametric methods according to different data types should be used to achieve optimal data utility and disclosure risk.

References

1. Rubin DB. Discussion: Statistical disclosure limitation. J Off Stat. 1993;9:461–8.
2. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. J Off Stat. 2003;19(1):1–16.
3. Matthews GJ, Harel O. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Stat Surv. 2011;5:1–29.

4. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, de Wolf P. Statistical disclosure control. John Wiley & Sons. 2012.

5. Fuller W. Masking procedures for microdata disclosure limitation. J Off Stat. 1993;9:383–406.

6. Dalenius T, Reiss SP. Data-swapping: A technique for disclosure control. J Stat Plan Inference. 1982;6:73–85.

7. Reiter JP. Satisfying disclosure restriction with synthetic data sets. J Off Stat. 2002;18(4):531–43.

8. Reiter JP. Inference for partially synthetic, public use microdata sets. Surv Methodol. 2003;29(2):181–8.

9. Little RJA. Statistical analysis of masked data. J Off Stat. 1993;9:407–26.

10. Fienberg SE, Steele RJ. Disclosure limitation using perturbation and related methods for categorical data. J Off Stat. 1998:14(4):485–502.

11. Drechsler J. Synthetic datasets for statistical disclosure control: Theory and implementation. Vol. 201. Springer Science & Business media. 2011.

12. Gouweleeuw JM, Kooiman P, Wolf P. Post randomisation for statistical disclosure control: Theory and implementation. J Off Stat. 1998;14(4):463.

13. Moore R Jr. Controlled data-swapping techniques for masking public use microdata. Census Tech Report. 1996. Available from: https://www.census.gov/content/dam/Census/library/working-papers/1996/adrm/rr96-4.pdf.

14. Oganian A, Karr AF. Combinations of SDC methods for microdata protection. In: Domingo-Ferrer J, Franconi L, editors. Privacy in statistical databases. Springer Berlin Heidelberg. 2006. p. 102–113.

15. Reiter JP. Estimating risks of identification disclosure in microdata. J Am Stat Assoc. 2005;100:1103–12.

16. Irimata KE, He Y, Cai B, Shin H-C, Parsons VL, Parker JD. Comparison of quarterly and yearly calibration data for propensity score adjusted web survey estimates. Surv Methods Insights Field. Special issue: Advancements in online and mobile survey methods. 2020. DOI: https://doi.org/10.13094/SMIF-2020-00018.

17. Parker J, Miller K, He Y, Scanlon P, Cai B, Shin H-C, et al. Overview and initial results of the National Center for Health Statistics' Research and Development Survey. Stat J IAOS. 2020;6(4):1199–1211. DOI: https://doi.org/10.3233/SJI-200678.

18. He Y, Cai B, Shin H-C, Beresovsky V, Parsons V, Irimata K, et al. The National Center for Health Statistics' 2015 and 2016 Research and Development Surveys. Vital Health Stat 1(64). 2020. Available from: https://www.cdc.gov/nchs/data/series/sr_01/sr01-64-508.pdf.

19. National Center for Health Statistics. RANDS during COVID-19 round 3 technical documentation. Hyattsville, Maryland. 2022. Available from: https://archive.cdc.gov/www_cdc_gov/nchs/rands/files/RANDS_COVID_3_technical_documentation.pdf.

20. Raghunathan TE, Solenberger P, Berglund P, Hoewyk JV. IVEware: Imputation and variance estimation software (Version 0.3). 2016.

21. Nowok B, Raab GM, Dibben C. Synthpop: Bespoke creation of synthetic data in R. J Stat Softw. 2016;74(11):1–26. DOI: https://doi.org/10.18637/jss.v074.i11.

22. Reiter JP. Using CART to generate partially synthetic public use microdata. J Off Stat. 2005;21(3):441–462.

23. Drechsler J, Reiter J. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput Stat Data Anal. 2011;55:3232–3243.

24. Reiter JP. Simultaneous use of multiple imputation for missing data and disclosure limitation. Surv Methodol. 2004;30(2):235–242.

25. SAS Institute Inc. SAS® 9.4 language reference: Concepts, sixth edition. Cary, NC: SAS Institute Inc. 2016.

26. Reiter JP. Significance tests for multi-component estimands from multiply imputed, synthetic microdata. J Stat Plan Inference. 2005;131:365–77.

27. Karr A, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. Am Stat. 2006;60(3):224–32.

28. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. J R Stat Soc Ser A Stat Soc. 2018;181(3):663–88. DOI: https://doi.org/10.1111/rssa.12358.

29. Woo MJ, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. J Priv Confid. 2009;1(1). DOI: https://doi.org/10.29012/jpc.v1i1.568.

30. Reiter JP, Wang Q, Zhang B. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. J Priv Confid. 2014;6(1). DOI: https://doi.org/10.29012/jpc.v6i1.635.

31. Paiva T, Chakraborty A, Reiter J, Gelfand A. Imputation of confidential data sets with spatial locations using disease mapping models. Stat Med. 2014;33(11):1928–45. DOI: https://doi.org/10.1002/sim.6078.

32. Reiter JP and Mitra R. Estimating risks of identification disclosure in partially synthetic data. J Priv Confid. 2009;1(1). DOI: https://doi.org/10.29012/jpc.v1i1.567.

Table 1. Analytical variables selected for data synthesis: Research and Development Survey During COVID-19, Round 3

RANDS variable name	Definition	Category
AGE	Age	1 is 18–44 2 is 45–64 3 is 65 and older
RACETHNICITY	Race and ethnicity	1 is White, non-Hispanic 2 is Black, non-Hispanic 3 is Other, non-Hispanic 4 is Hispanic
EDUC	Education	1 is High school graduate or less 2 is Some college 3 is Bachelor's degree or more
Sex	Sex	1 is Male 2 is Female
MARITAL	Marital status	1 is Married 2 is Widowed 3 is Divorced 4 is Separated 5 is Never married 6 is Living with partner
REGION4	Census region	1 is Northeast 2 is Midwest 3 is South 4 is West
INCOME	Household income from the last year	1 is Less than $50,000 2 is $50,000 to $99,999 3 is $100,000 or more
PHSTAT	Self-reported health status	1 is Excellent, very good, good 2 is Fair, poor
HYPEV	Ever had doctor-diagnosed hypertension	1 is Yes 2 is No
HICOV	Are you covered by any kind of health insurance or some other kind of health care plan?	1 is Yes 2 is No
EMPLASTWK	Last week, did you work for pay at a job or business?	1 is Yes 2 is No
SMKEV/SMKNOW	Current smoking status	1 is Yes 2 is No

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19 Round 3, 2021.

Table 2. Role of design variables during the data synthesis process for Approaches 1–3: Research and Development Survey During COVID-19, Round 3

	Are any of the design variables (strata, PSU[1]) or sampling weights synthesized?	Which components of design information were synthesized?	Is all design information used as covariates when synthesizing the analytical variables?
Approach 1[2]	Yes	Strata, PSU, sampling weights	Yes
Approach 2[3]	Yes	Sampling weights	Yes
Approach 3[4]	No	None	Yes

[1]Primary sampling unit.
[2]All variables synthesized.
[3]Design variables not synthesized.
[4]Design variables and weights not synthesized.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19 Round 3, 2021.

Table 3. Mean of confidence interval overlap across all synthesized analytical variables: Research and Development Survey During COVID-19, Round 3

Method	Software used for implementation	Approach 1[1]	Approach 2[2]	Approach 3[3]
Parametric	R synthpop (regression)	0.73	0.71	0.88
	IVEware (SYNTHESIZE)	…	0.70	0.85
Nonparametric	R synthpop (CART[4])	0.88	0.89	0.88

… Category not applicable.
[1]All variables synthesized.
[2]Design variables not synthesized, weights synthesized.
[3]Design variables and weights not synthesized.
[4]Classification and regression trees.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19 Round 3, 2021.

Table 4. Propensity scores estimated from a weighted logistic regression model: Research and Development Survey During COVID-19, Round 3

Method	Software used for implementation	Approach 1[1]	Approach 2[2]	Approach 3[3]
			Mean (range)	
Parametric	R synthpop (regression)	0.50 (0.34–0.70)	0.50 (0.35–0.64)	0.50 (0.42–0.64)
	IVEware (SYNTHESIZE)	…	0.56 (0.45–0.68)	0.51 (0.40–0.64)
Nonparametric	R synthpop (CART[4])	0.50 (0.42–0.61)	0.50 (0.41–0.56)	0.50 (0.42–0.57)

… Category not applicable.
[1]All variables synthesized.
[2]Design variables not synthesized, weights synthesized.
[3]Design variables and weights not synthesized.
[4]Classification and regression trees.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19 Round 3, 2021.

Table 5. Quantiles and means of the original and synthesized sampling weights for selected parametric and nonparametric synthesis methods applied to Approach 2: Research and Development Survey During COVID-19, Round 3

		Parametric method		Nonparametric method
Quantile	Original data	R	IVEware	R
100% maximum	17.65	5.25	5.34	17.65
99%	5.12	3.60	3.52	4.70
95%	2.86	2.57	2.80	2.86
90%	2.12	2.07	2.39	2.05
75%	1.24	1.42	1.79	1.23
50% median	0.69	0.83	1.15	0.68
25%	0.37	0.39	0.61	0.36
10%	0.22	0.14	0.26	0.22
5%	0.14	0.05	0.14	0.14
1%	0.05	0.00[1]	0.03	0.05
0% minimum	0.01	0.00[2]	0.00[3]	0.01
Mean	1.00	1.00	1.27	0.98

[1]Rounded to 0.00 from 0.002.
[2]Rounded to 0.00 from 0.00000002.
[3]Rounded to 0.00 from 0.001.

NOTE: Approach 2 is design variables not synthesized.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19, Round 3.

Table 6. Average matching probability of Approach 1, by selected synthesis method and number of records with matching original values: Research and Development Survey During COVID-19, Round 3

Target case	Records with the same values in the original data	Parametric		Nonparametric
		R (regression)	IVEware (regression)	R (CART[1])
	Number	Percent		
1.	1 (unique record in the original data)	0.29	…	–
2.	5	1.11	…	0.94
3.	10	4.16	…	4.29
4.	15	2.62	…	3.70
5.	20 (high frequency record in the original data)	4.53	…	5.79
Across all target cases	Mean	2.54	…	2.94

… Category not applicable.
– Quantity zero.
[1]Classification and regression trees.

NOTE: Approach 1 is all variables synthesized.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19, Round 3.

Table 7. Average matching probability of Approach 2, by selected synthesis method and number of records with matching original values: Research and Development Survey During COVID-19, Round 3

Target case	Records with the same values in the original data	Parametric		Nonparametric
		R (regression)	IVEware (regression)	R (CART[1])
	Number	Percent		
1.	1 (unique record in the original data)	0.31	0.52	–
2.	5	1.42	1.16	0.68
3.	10	3.44	3.75	5.06
4.	15	3.72	2.72	4.45
5.	20 (high frequency record in the original data)	5.22	4.83	5.38
Across all target cases	Mean	2.82	2.60	3.11

– Quantity zero.
[1]Classification and regression trees.

NOTE: Approach 2 is design variables not synthesized.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19, Round 3.

Table 8. Average matching probability of Approach 3, by selected synthesis method and number of records with matching original values: Research and Development Survey During COVID-19, Round 3

Target case	Records with the same values in the original data	Parametric		Nonparametric
		R (regression)	IVEware (regression)	R (CART[1])
	Number	Percent		
1.	1 (unique record in the original data)	0.82	–	–
2.	5	0.94	1.09	1.78
3.	10	4.37	3.78	4.29
4.	15	3.25	3.14	3.89
5.	20 (high frequency record in the original data)	5.66	4.01	6.22
Across all target cases	Mean	3.01	2.40	3.24

– Quantity zero.
[1]Classification and regression trees.

NOTE: Approach 3 is design variables and weights not synthesized.

SOURCE: National Center for Health Statistics, Research and Development Survey During COVID-19, Round 3.

Vital and Health Statistics Series Descriptions

Active Series

Series 1. **Programs and Collection Procedures**
Reports describe the programs and data systems of the National Center for Health Statistics, and the data collection and survey methods used. Series 1 reports also include definitions, survey design, estimation, and other material necessary for understanding and analyzing the data.

Series 2. **Data Evaluation and Methods Research**
Reports present new statistical methodology including experimental tests of new survey methods, studies of vital and health statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. Reports also include comparison of U.S. methodology with those of other countries.

Series 3. **Analytical and Epidemiological Studies**
Reports present data analyses, epidemiological studies, and descriptive statistics based on national surveys and data systems. As of 2015, Series 3 includes reports that would have previously been published in Series 5, 10–15, and 20–23.

Discontinued Series

Series 4. **Documents and Committee Reports**
Reports contain findings of major committees concerned with vital and health statistics and documents. The last Series 4 report was published in 2002; these are now included in Series 2 or another appropriate series.

Series 5. **International Vital and Health Statistics Reports**
Reports present analytical and descriptive comparisons of U.S. vital and health statistics with those of other countries. The last Series 5 report was published in 2003; these are now included in Series 3 or another appropriate series.

Series 6. **Cognition and Survey Measurement**
Reports use methods of cognitive science to design, evaluate, and test survey instruments. The last Series 6 report was published in 1999; these are now included in Series 2.

Series 10. **Data From the National Health Interview Survey**
Reports present statistics on illness; accidental injuries; disability; use of hospital, medical, dental, and other services; and other health-related topics. As of 2015, these are included in Series 3.

Series 11. **Data From the National Health Examination Survey, the National Health and Nutrition Examination Surveys, and the Hispanic Health and Nutrition Examination Survey**
Reports present 1) estimates of the medically defined prevalence of specific diseases in the United States and the distribution of the population with respect to physical, physiological, and psychological characteristics and 2) analysis of relationships among the various measurements. As of 2015, these are included in Series 3.

Series 12. **Data From the Institutionalized Population Surveys**
The last Series 12 report was published in 1974; these reports were included in Series 13, and as of 2015 are in Series 3.

Series 13. **Data From the National Health Care Survey**
Reports present statistics on health resources and use of health care resources based on data collected from health care providers and provider records. As of 2015, these reports are included in Series 3.

Series 14. **Data on Health Resources: Manpower and Facilities**
The last Series 14 report was published in 1989; these reports were included in Series 13, and are now included in Series 3.

Series 15. **Data From Special Surveys**
Reports contain statistics on health and health-related topics from surveys that are not a part of the continuing data systems of the National Center for Health Statistics. The last Series 15 report was published in 2002; these reports are now included in Series 3.

Series 16. **Compilations of Advance Data From Vital and Health Statistics**
The last Series 16 report was published in 1996. All reports are available online; compilations are no longer needed.

Series 20. **Data on Mortality**
Reports include analyses by cause of death and demographic variables, and geographic and trend analyses. The last Series 20 report was published in 2007; these reports are now included in Series 3.

Series 21. **Data on Natality, Marriage, and Divorce**
Reports include analyses by health and demographic variables, and geographic and trend analyses. The last Series 21 report was published in 2006; these reports are now included in Series 3.

Series 22. **Data From the National Mortality and Natality Surveys**
The last Series 22 report was published in 1973. Reports from sample surveys of vital records were included in Series 20 or 21, and are now included in Series 3.

Series 23. **Data From the National Survey of Family Growth**
Reports contain statistics on factors that affect birth rates, factors affecting the formation and dissolution of families, and behavior related to the risk of HIV and other sexually transmitted diseases. The last Series 23 report was published in 2011; these reports are now included in Series 3.

Series 24. **Compilations of Data on Natality, Mortality, Marriage, and Divorce**
The last Series 24 report was published in 1996. All reports are available online; compilations are no longer needed.

For answers to questions about this report or for a list of reports published in these series, contact:

Information Dissemination Staff
National Center for Health Statistics
Centers for Disease Control and Prevention
3311 Toledo Road, Room 4551, MS P08
Hyattsville, MD 20782

Tel: 1–800–CDC–INFO (1–800–232–4636)
TTY: 1–888–232–6348
Internet: https://www.cdc.gov/nchs
Online request form: https://www.cdc.gov/info
For e-mail updates on NCHS publication releases, subscribe online at: https://www.cdc.gov/nchs/updates/.

U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES

Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 4551, MS P08
Hyattsville, MD 20782–2064

For more NCHS Series Reports, visit:
https://www.cdc.gov/nchs/products/series.htm

Series 2, No. 212
CS357967