A Tutorial in Generating Synthetic Data to Mitigate Disclosure Risk in Microdata: An Expository Review of Taylor, Zhou, and Rise (2017)

Abstract

In a world increasingly powered by data aggregation and analysis, the secure release of information has become a cornerstone of privacy protection. Traditionally, agencies have relied on microdata risk assessment procedures based on checklist criteria, ad hoc rules, and data-based summary measures. However, the growing demand for quantitative risk measures calls for a more objective criteria for data release (Taylor, Zhou, and Rise 2017). Taylor, Zhou, and Rise (2017) focus on the risk analysis stage within the disclosure control framework, defining the existing measures and how to estimate them via software. This paper goes a step further, investigating a practical solution for datasets deemed high-risk: synthetic data generation as a means to balance data utility with privacy. Using the National Health and Nutrition Examination Survey (NHANES) as a case study, this paper explores the R package `synthpop` and evaluates its effectiveness in preserving key statistical properties of the original data while reducing disclosure risks.

Keywords: Microdata, Statistical Disclosure Control, Risk Analysis, Synthetic Data, `synthpop`

Introduction

"Getting information from the Internet is like taking a drink from a hydrant," says Mitchell Kapor, personal computing pioneer and investor (Buttice 2022). In today's era of data abundance, the challenge is no longer just uncovering data, but also safeguarding the privacy of individuals whose identities are embedded within it. National statistical agencies face increasing pressure to release microdata, or data that contains record-level or detailed information about individual entities, on a number of variables. This data is often used to facilitate breakthrough discoveries or research that help inform policy decisions, whether it be in the realm of public health, economics, or the social sciences (Taylor, Zhou, and Rise 2017).

With the growing momentum around increased access to microdata, it is essential to understand how to safely release such information. Disclosure, or reidentification, occurs when a person or organization uncovers new and often confidential information about an entity due to the release of data. For instance, as demonstrated by Emam, Dankar, and Jonker (2013), combining public obituary records with a deidentified clinical trial dataset can reveal the identities of deceased subjects. In cases like this, privacy is compromised, stripping individuals – and sometimes their friends and families – of control over sensitive personal information.

Disclosure even has ramifications for the agencies that release data. If a breach occurs, the agency may face legal consequences, a loss of trust, and a decrease in the quality of data that they collect (Taylor, Zhou, and Rise 2017).

To reduce the risk of reidentification, organizations deidentify data before its release. The United States' Health Insurance Portability and Accountability Act (HIPAA) offers a deidentification framework, which involves removing 18 categories of identifiers from a dataset. However, research by Benitez and Malin (2010) shows that even when datasets comply with HIPAA standards, the risk of reidentification remains significant. This is because demographic variables – such as gender and state of residency – can still be linked with publicly accessible databases (Taylor, Zhou, and Rise 2017).

In their article, Taylor, Zhou, and Rise (2017) focus on statistical methods to assess disclosure risk, where the microdata are obtained as a sample of the original population dataset. They propose three approaches to measuring disclosure risk and outline different measures and modeling methods. These measures strive to detect high-risk datasets as an entirety as well as which records within the dataset may be of high risk. Taylor, Zhou, and Rise (2017) thus investigate Stage 3 of the statistical disclosure control process:

- Stage 1: Determine whether confidentiality protection is needed
- Stage 2: Identify key data characteristics (e.g. whether the data comes from a sample or census) and how the data will be used
- Stage 3: Define and estimate measures of disclosure risk
- Stage 4: Choose appropriate disclosure control methods (DCMs) if high risk was detected in Stage 3

- Stage 5: Implement the risk analysis and subsequent DCMs to produce the final dataset for release

This paper builds on the work of Taylor, Zhou, and Rise (2017) by focusing on Stage 4. Given that high risk was identified in Stage 3, a possible next step involves generating synthetic data that retains the essential statistical properties of the original data while reducing reidentification risk. This aims to strike a balance between safeguarding privacy and maximizing the utility of the released data.

The report will begin with a brief overview of a set of statistical measures used to assess disclosure risk, as outlined by Taylor, Zhou, and Rise (2017). This background will provide a foundation for understanding how these measures can be applied to a given dataset. Next, the true data, drawn from the National Health and Nutrition Examination Survey (NHANES), will be introduced. NHANES is a comprehensive program designed to assess the health and nutritional status of adults and children in the United States. The dataset contains a wealth of information on demographics, health conditions, and lifestyle factors (Pruim 2015).

Following the introduction of the NHANES data, the report will discuss the use of the `synthpop` package in R, which is employed to generate two synthetic datasets. The primary focus of the analysis will be on the first synthetic dataset, while the second synthetic dataset will be included in the multivariate analysis to emphasize the stochastic nature of the synthesis process. The `synthpop` package replaces sensitive records with values simulated from probability distributions specified to preserve key features of the actual observed data. Additionally, `synthpop` can assess the quality of these synthetic datasets and provide additional means of confidentiality protections (Nowok, Raab, and Dibben 2016).

To guide the analysis, this paper will outline methods for assessing whether the generated synthetic datasets maintain the important relationships and dependencies within the original data. Further, the paper will outline techniques for evaluating the reduction in disclosure risk when employing such synthetic processes.

Methods

Expository Review

Taylor, Zhou, and Rise (2017) assume that an intruder aims to identify an individual in the microdata by matching records to known individuals in the population. This process involves using key variables – identifying variables whose values are known for both the microdata and the population. An individual is at risk of disclosure if they can be easily distinguished from others, making disclosure analysis a measure of how uniquely identifiable a unit is within the sample or population. Taylor, Zhou, and Rise (2017) focus on statistical modeling to measure disclosure risk, making assumptions about the nature of the external (population) file within a modeling framework and estimating the probability of disclosure within this framework. The

approach enables file and record level risk estimation without the need to physically construct the external file.

The following risk estimates are considered conservative as they are built around a "worst case" scenario:

- The intruder's external file contains individual identifiers and categorical key variables that overlap with the sample microdata
- The intruder's external data covers the entire population so that each record in the sample microdata can be matched to a record in the external dataset
- There are no error or missing values in the key variables

Consider microdata to be released that contains a set of records, each corresponding to a subject in a sample s selected from a finite population U. Let n and N represent the number of subjects in sample s and population U, respectively. The microdata contains a set of categorical variables, k, which the intruder uses to match records with the intruder's external dataset. Risk measures are based on the categorical variable formed by cross-classifying all key variables, X, with values $j = 1, ..., J$. A hypothetical value for individual i might be $X_i =$ (57, *male*, *doctor*, *married*). The number of cells, J, is expected to be very large. Frequencies in the population are defined as

$$F_j = \sum_{i \in U} I_{[X_i = j]}, j = 1, ..., J$$

and the observed frequencies in the microdata are defined as

$$f_j = \sum_{i \in s} I_{[X_i = j]}, j = 1, ..., J$$

where *I()* is the indicator function. If there is only one subject in the population having value j for X, a population unique, then $F_j = 1$. Similarly, a value j of X is sample unique if $f_j = 1$.

Disclosure risk literature commonly references five key measures of disclosure risk, all of which are primarily global, file-level measures, though some have significant record-level variants.

Measure 1 is the expected number of population uniques:

$$\sum_j Pr(F_j = 1).$$

If a subject is unique in the population, and is found in the released data, then the subject is reidentifiable by linking such records.

Measure 2 concentrates on those that are sample unique and measures the expected number of sample uniques that are population unique:

$$\sum_{j:f_j=1} Pr(F_j = 1|f_j = 1).$$

In contrast to Measures 1 and 2, Measures 3-5 consider risk arising from records that are not population unique (e.g. pairs, triplets, etc.). If the population frequency F_j is known for combination j, risk is simply measured by $\frac{1}{F_j}$ for each record in the jth combination. For instance, if there are 4 records in the population with the same values of key variables, there is a $\frac{1}{4}$ probability of a correct match to a record in the microdata. Measure 3 is the expected number of correct matches among sample uniques:

$$\sum_{j:f_j=1} E(\frac{1}{F_j}|f_j = 1)$$

with a record-level form of

$$E(\frac{1}{F_j}|f_j = 1).$$

Measure 4 represents the probability of a correct match given a unique match:

$$\frac{\sum_j I(f_j = 1)}{\sum_j F_j I(f_j = 1)}.$$

Lastly, Measure 5, or the "Benedetti-Franconi risk measure," estimates the probability of a correct match even among records that are not sample unique:

$$\sum_{j=1}^{J} E(\frac{1}{F_j}|f_j) = \sum_{j=1}^{J}(\sum_{r \geq f_j} \frac{1}{r} Pr(F_j = r|f_j))$$

with a record level form of

$$E(\frac{1}{F_j}|f_j) = \sum_{r \geq f_j} \frac{1}{r} Pr(F_j = r|f_j).$$

Although beyond the scope of this paper, Taylor, Zhou, and Rise (2017) discuss existing methods for modeling file- and record-level measures. Estimating these measures requires specific modeling assumptions and estimation techniques, which can be categorized into parametric, semiparametric, and nonparametric approaches.

The R package `sdcMicro` contains a `modRisk` function that estimates global risk Measures 2 and 3 using log-linear models and estimation methods developed in Rinott and Shlomo

(2006) and can account for hierarchy in the data structure. The `indivRisk` function estimates individual risk Measure 4 by Bayesian methods (Templ, Kowarik, and Meindl 2015). The package is intuitive and user-friendly, making it simple to implement in practice:

```
# example implementation of statistical disclosure control (sdc)
sdc <- createSdcObj(data,
                    # key categorical variables
                    keyVars = c('Profession', 'Marital'),
                    # quantitative variables
                    numVars = c('Age', 'Income', 'Height', 'Weight'),)

modRisk(sdc, method = "default",
        formula = ~ Profession + Marital + Age + Income + Height + Weight)

# how to view results
sdc@risk
```

Application to NHANES Data

Imagine we collected the NHANES dataset and were asked to publicly share it. Due to the complexity of these methods and limited available computer memory, let's assume that utilizing `sdcMicro` indicates very dangerous risk levels. What can we do?

A solution mentioned in Taylor, Zhou, and Rise (2017) entails generating synthetic data that preserves important statistical properties of the original data.

Let's take a random sample of 500 individuals from the NHANES dataset to use as the true data. The key variables, or variables that the intruder can theoretically use to merge the deidentified dataset with the identified dataset, consist of gender (Gender), age in years (Age), race (Race1), ratio of family income to poverty guidelines (Poverty), weight in kilograms (Weight), and body mass index (BMI).

For this true, or usually unobservable, data, Table 1 shows there are 58 missing values across the six variables – 41 in Poverty, 13 in BMI, and four in Weight. The patterns of this missingness are also intended to be synthesized.

Table 1: Distribution of Missing Values Across True-Data Variables: The missingness patterns, with a notable concentration in the Poverty variable, are intended to be synthesized.

Variable	# Missing Values
Poverty	41
BMI	13
Weight	4

Gender	0
Age	0
Race1	0

The frequency of distribution counts, f_j, for cross-classified categories is shown in Table 2. This table refers to the number of categories j that contain only one person, two people, three people, etc. Using the variables of Gender, Age, Race1, Poverty, Weight, and BMI, there are 467 uniques, 13 pairs, one triplet, and one quartet.

Table 2: True Data Frequency of Distribution Counts: Disclosure analysis evaluates the uniqueness of individual units within a dataset. In this context, the 467 unique cases in the true data underscore the critical importance of deidentification.

Count(fj)	Frequency	% of Records
1	467	93.4
2	13	5.2
3	1	0.6
4	1	0.8

Figure 1 illustrates the univariate and bivariate relationships among the quantitative variables in the true data. While the correlation between Weight and BMI is very strong, which is plausible as BMI is calculated using weight, BMI shares no correlation with Poverty. Apart from BMI and Poverty, all other bivariate relationships are statistically significant, meaning the relationship between the variables is not due to random chance. While the distributions of Weight and BMI are right skewed, the distribution of Age appears approximately normal and the distribution of Poverty appears bimodal. As shown in the scatterplots, outliers appear to exist in the data, particularly in the top right corner of the Weight-BMI plot and the top of the Poverty-BMI and Age-BMI plot. These specific points may pose a disclosure risk due to their uniqueness, accentuating the synthetic dataset's essential nature in anonymizing these individuals.

A contingency table for the categorical variables presents the joint distribution of Gender and Race1. As illustrated in Table 3, the distribution of males and females is nearly balanced, with the majority of individuals identifying as White. Furthermore, the Gender distribution within each racial group is also almost equal.

Table 3: 2x5 Contingency Table of True Data: The overall Gender distribution is nearly equal, with a balanced gender split within each race and a high volume of White observations.

	Black	Hispanic	Mexican	White	Other	Sum
female	24	18	26	158	23	249

male	30	15	26	159	21	251
Sum	54	33	52	317	44	500

Synthetic Data Generation

In real-world applications, specifying the exact joint distribution of all variables in a dataset is challenging. To address this, `synthpop` approximates the joint distribution using a series of conditional distributions. Synthesis occurs on a variable-by-variable basis: each variable is modeled sequentially by fitting a sequence of regression models and drawing synthetic values from the corresponding predictive distributions. Each model conditions on variables processed earlier in the sequence, resulting in a cumulative increase in covariates, with the final variable conditioned on all preceding variables. For missing data, `synthpop` synthesizes on the observed patterns of missingness, preserving the structure of missing values in the true data (Nowok).

The `syn()` function creates synthetic versions of a dataset provided as its argument, with the process being largely automated when using default settings. These default settings apply a "cart" method, which uses classification and regression trees, to all variables except the first in the visit sequence (Nowok). Given that the true data contains both quantitative and categorical variables, the synthesis method has been manually specified for each type: "norm" is used for quantitative variables and "cart" for categorical variables, with the exception of the first in the visit sequence (Gender), which is retained as sampling with replacement. "norm" is favored for quantitative variables because continuous data is often symmetrically distributed or transformable to approximate normality. "cart" is better suited for categorical variables, splitting the data based on decision rules and segmenting the data into categories and subcategories based on patterns without presuming a specific distribution. Additional post processing is conducted via the `sdc()` function to strengthen the protection of information confidentiality. Through `sdc()`, *all* unique cases in the synthetic data that are identical to unique individuals in the real dataset are removed by setting the `rm.replicated.uniques` command to `TRUE` (Nowok). This approach represents the most conservative scenario, where all actual values are substituted.

To evaluate how well the synthetic datasets preserve relationships from the true data, various bivariate and multivariate analyses are conducted. A scatterplot matrix will illustrate the relationships among quantitative variables across the first synthetic dataset. This matrix, displaying each variable's distribution as well as its pairwise relationship with others, allows for visual comparison between the true and synthetic data. A 2x5 contingency table is used to summarize the joint distribution of the two categorical variables, Gender and Race1, offering another direct comparison between the actual and synthetic data. The `compare()` function in `synthpop` analyzes the relative frequency of distributions of each variable, providing tabular and graphical measures of alignment.

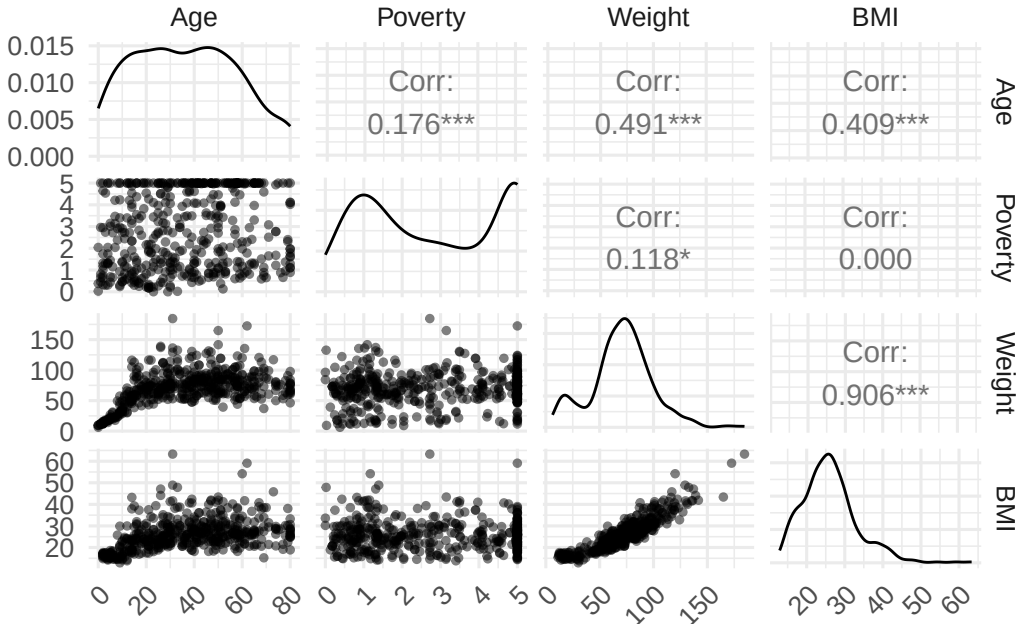

Figure 1: Scatterplot Matrix of True Data: Notice the non-normal univariate distributions and the presence of outliers in the bivariate relationships, which will be crucial to address during the synthetic data generation process.

In the multivariate analysis, a linear regression model is fitted to the true dataset as well as the two synthetic datasets. The model is defined as follows:

$$\hat{BMI} = \beta_0 + \beta_1 Gender + \beta_2 Age + \beta_3 Race1 + \beta_4 Poverty + \beta_5 Weight$$

where β_0 is the intercept, β_1 to β_5 are the coefficients for each predictor variable, and female and Black are the reference categories. Comparing these coefficients across the true and synthetic data will indicate the extent to which the synthetic data captures the multivariate relationships present in the true data. By incorporating `lm_synds()`, estimates based on the synthesized data are compared to those based on the true data, and 95% confidence intervals for the Z statistics for observed and synthetic data are calculated and plotted (Nowok).

Finally, the disclosure risk associated with the first synthetic dataset is assessed. Specific measures will be taken to quantify the potential risk, such as applying the synthetic model to the true data to see if it can reidentify observations. This will involve predicting BMI and calculating the mean squared error (MSE) and mean absolute error (MAE). A small MSE and MAE would indicate that the synthetic model is accurately predicting true data BMI values, which poses disclosure risk. Another approach involves comparing outliers between the true and synthetic data. A unit is defined as an outlier if its value is two standard deviations above the mean. Comparing outliers across quantitative and categorical variables can also reveal potential disclosure risks.

Results

After generating the first synthetic dataset, the scatterplot matrix in Figure 2 visualizes the relationships among the quantitative variables, showing correlations that closely mirror those in the true data in Figure 1. The relationship between Weight and BMI remains notably strong, while the correlation between Poverty and BMI remains weak. The distributions of each quantitative variable in the synthetic data approximate normality – a distinct contrast to the true data, where certain variables displayed non-normal patterns, such as Poverty with its bimodal distribution. Perhaps most importantly, the noticeable outliers in the true data are not present in Figure 2. The scatterplots are more clustered and cloud-like, suggesting that the synthetic data has effectively anonymized these individuals.

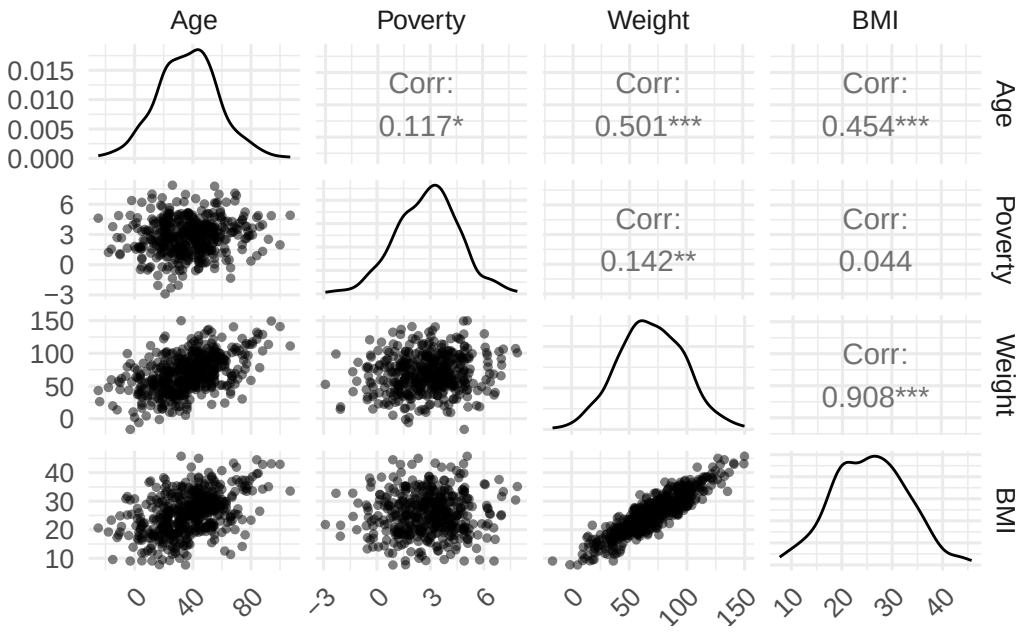

Figure 2: Scatterplot Matrix of Synthetic Data: Notice the normal univariate relationships and the absence of outliers in the bivariate relationships, with distinct clouds forming.

In terms of the contingency table for the synthetic data, Table 4 indicates that the distribution across Gender almost perfectly resembles the distribution in the true data, with only a one person increase in males and a one person decrease in females. The total number of individuals falling under each race is also consistent, as well as the distribution of Gender within each racial group. Two exceptions are the Mexican and White groups. Although the true data had equal numbers of male and female individuals in the Mexican group and only a one person difference in the White group, the synthetic data shows more of a discrepancy: there are 20 more Mexican females than males and 17 more White males than females.

Table 4: 2x5 Contingency Table of Synthetic Data: Similar to the true data, the overall Gender distribution is nearly equal, with a balanced gender split within each race, except for the Mexican and White groups. The distribution of each race also closely matches the true data.

	Black	Hispanic	Mexican	White	Other	Sum
female	24	14	40	149	21	248
male	28	20	20	166	18	252
Sum	52	34	60	315	39	500

Relative frequency distributions for the true and synthetic data are depicted in Figure 3 and Figure 4. These graphical displays, provided by the `compare` function, emphasize that the synthetic data decently emulates the true data's distribution patterns. The synthetic data appears to have more prevalence in the (unreasonable) extremes, meaning that some synthetic values have negative Poverty, negative or extremely old Age, or negative Weight. This is a common issue with synthetic data – in its effort to capture the true data's complexity, it is susceptible to generating unrealistic values.

Propensity scores represent probabilities of group memberships, where small distinguishability relates to high distributional similarity between the original and masked data. The set of predictors is specified/calculated for the original and synthetic datasets, which are then combined with the addition of an indicator variable I to denote the dataset (0 for original, 1 for altered). A propensity score is estimated for each of the rows of the combined data as a probability of classification for the indicator variable. By taking the mean-squared difference between the estimated probabilities and the true proportion of records from the masked data in the combined data, the propensity score mean-squared error is found (pMSE). The desired result is poor classification, and thus a lower pMSE. The standardized propensity mean-squared error (S_pMSE) is designed to have an expectation of zero and a standard deviation of 1 under the null hypothesis, where the synthetic data is generated from a model that mirrors the true data's distribution. This measure adjusts for the expected value and the variability of pMSE under the null, and is expected to increase if correct synthesis does not hold (Snoke et al. 2018).

Figure 3 and Figure 4 reveal that these standardized propensity scores are much lower for the categorical variables, indicating high distributional similarity. The standardized propensity scores for the quantitative variables are notably higher, with a particularly concerning increase for Poverty. This increase may be partially influenced by the high concentration of observed Poverty values around 4.5.

Measuring the synthetic data's ability to preserve multivariate relationships starts by fitting a linear model to predict BMI based on the true data, incorporating all other variables as predictors. Observing the Residual vs. Fitted plot in Figure 5, there is a relatively scattered distribution of points. Lower fitted values have positive residuals and higher fitted values lead

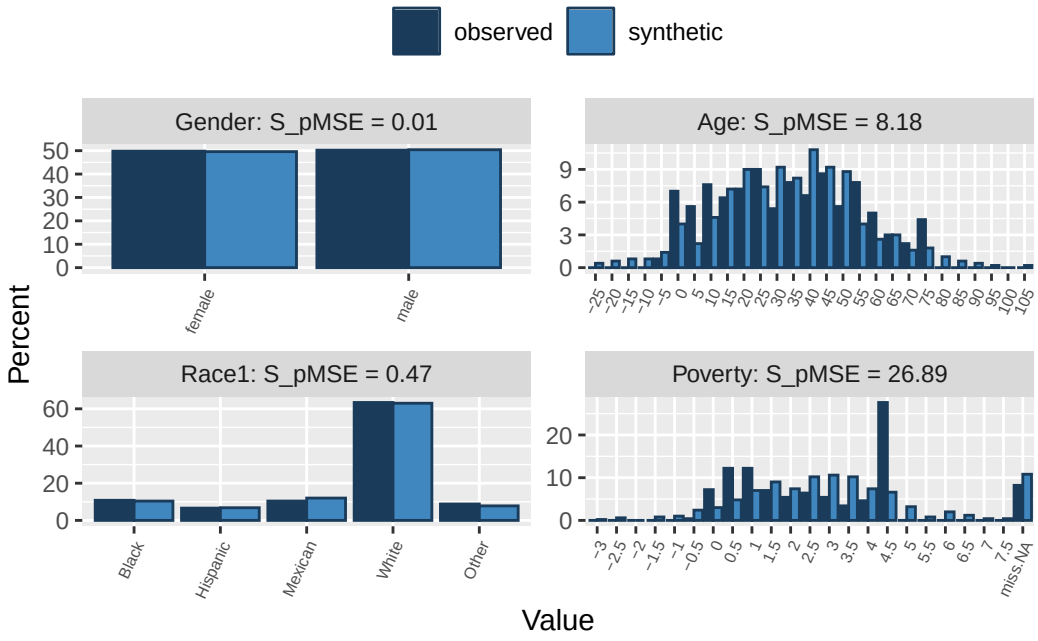

Figure 3: Relative Frequency Distribution for True and Synthetic Data: The synthetic data closely mirrors the true data's distribution patterns, especially among the categorical variables (which have lower S_pMSE). The quantitative variables have higher S_pMSE and synthetic values in unrealistic extremes.

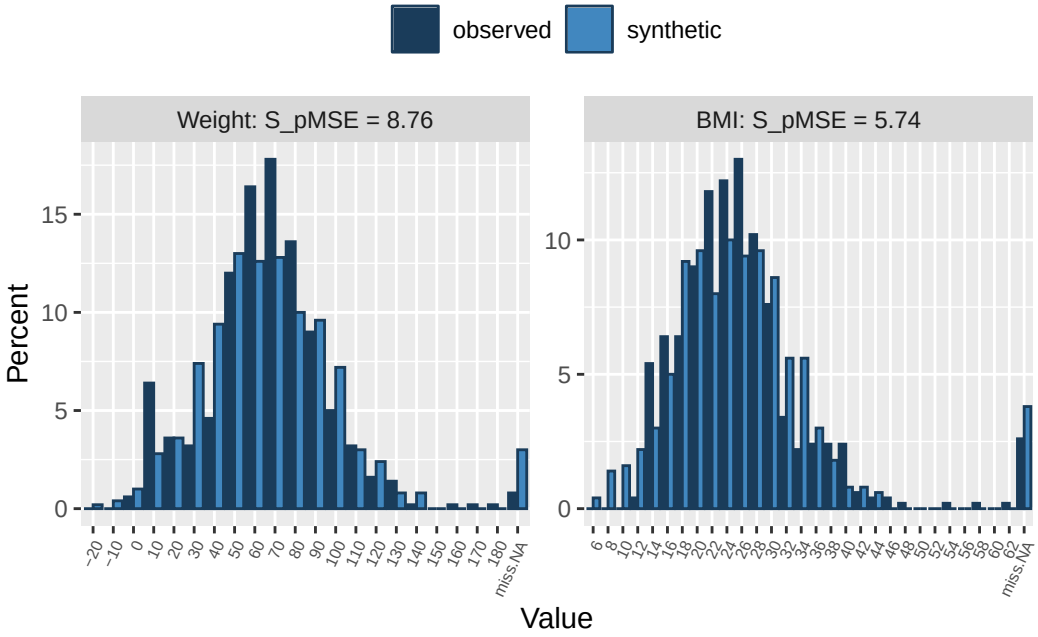

Figure 4: Relative Frequency Distribution for True and Synthetic Data: This second set of plots echo the quantitiative findings from the above plot - comparable distributions are observed across data type (observed/synthetic), with higher S_pMSE for the quantitative variables and synthetic values extending into unrealistic extremes.

to a small increase in the spread of residuals. The red line, or scatterplot smoother, displays a convex shape, warranting analysis/interpretation with caution. Almost all points on the Normal Q-Q plot fall on the line, conveying that the linearity condition of the model – that the errors are normally distributed – is met.

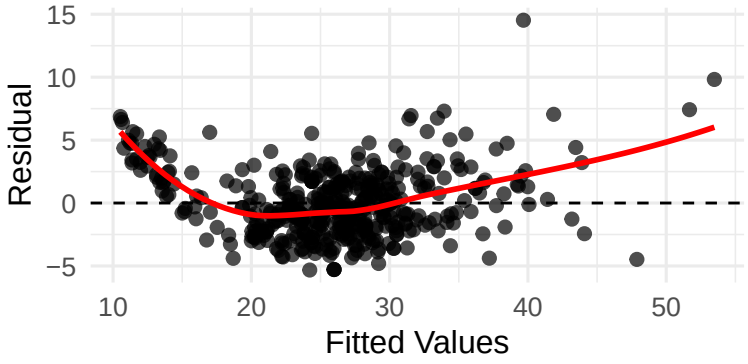

Figure 5: Residual vs. Fitted Plot for Linear Regression of BMI using True Data: There is some heteroscedasticity, particularly as fitted values deviate from the center of the plot. The analysis will proceed with caution.

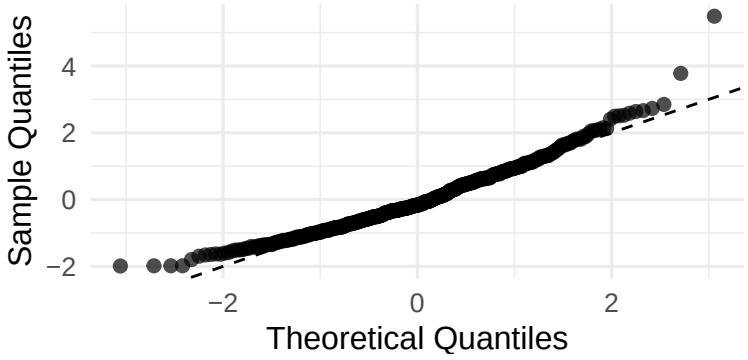

Figure 6: Normal Q-Q Plot for Linear Regression of BMI using True Data: The residuals closely follow the line, indicating that the errors are normally distributed and the condition is met.

By applying `lm_synds()` to the function `compare()` as its `object` argument, the estimates derived from the synthetic data and an additional synthetic dataset are compared against the original data. Table 5 and Table 6 illustrate that the difference in coefficients between the data are minimal, never exceeding one unit (disregarding the Intercept). The standardized coefficient difference, which accounts for the scale of the variables involved, is also small, with the largest difference being just -2.16 in the first synthetic dataset and 2.05 in the second. The final column presents the percentage of overlap between the estimated synthetic confidence intervals and the original sample confidence intervals for each parameter at the 95% confidence level (Nowok, Raab, and Dibben 2016). When considering both tables, only three coefficients

– Age, Race1White, and Intercept – have confidence interval overlaps below 0.5, at 0.48, 0.45, and 0.48 respectively. All overlap values suggest that while the model exhibits some differences, the estimates share a significant degree of similarity.

Table 5: Comparison of Coefficients and Confidence Interval Overlaps for Linear Regression of BMI (Observed vs. Synthetic Dataset 1): The differences in coefficients are minimal and the confidence intervals overlap to a substantial degree.

	Observed	Synthetic	Difference	Std. Coef. Diff	CI Overlap
(Intercept)	10.49	10.93	0.43	0.80	0.80
Gendermale	-2.21	-2.35	-0.15	-0.57	0.86
Age	-0.01	0.00	0.01	2.04	0.48
Race1Hispanic	1.40	1.31	-0.09	-0.14	0.96
Race1Mexican	0.16	-0.83	-0.99	-1.75	0.55
Race1White	-0.36	-1.29	-0.93	-2.16	0.45
Race1Other	0.53	-0.04	-0.57	-0.97	0.75
Poverty	-0.29	-0.21	0.08	1.03	0.74
Weight	0.25	0.25	-0.01	-1.00	0.74

Table 6: Comparison of Coefficients and Confidence Interval Overlaps for Linear Regression of BMI (Observed vs. Synthetic Dataset 2): Note the change in estimates due to the stochastic nature of the synthesis process. Nonetheless, the differences in coefficients remain minimal and the confidence intervals still overlap to a substantial degree.

	Observed	Synthetic	Difference	Std. Coef. Diff	CI Overlap
(Intercept)	10.49	11.60	1.11	2.05	0.48
Gendermale	-2.21	-2.21	0.00	0.00	1.00
Age	-0.01	0.00	0.01	0.94	0.76
Race1Hispanic	1.40	1.14	-0.26	-0.42	0.89
Race1Mexican	0.16	-0.44	-0.60	-1.05	0.73
Race1White	-0.36	-0.87	-0.51	-1.19	0.70
Race1Other	0.53	0.08	-0.45	-0.76	0.81
Poverty	-0.29	-0.39	-0.10	-1.24	0.68
Weight	0.25	0.24	-0.01	-1.40	0.64

A more holistic overview of the linear models is provided in Table 7, displaying a substantial mean CI overlap and a minimal mean absolute standardized difference in both synthetic datasets. The lack-of-fit test is applied to determine how well the synthetic data linear regression model fits the observed data model. In the test, the vector of mean differences between the coefficients calculated from the synthetic and original data provides a standardized lack-of-fit.

This value follows a chi-squared distribution with nine degrees of freedom, corresponding to the number of parameters in the fitted model. The p-value for the lack-of-fit test evaluates the null hypothesis that the method used for synthesis retains all relationships between variables that influence the parameters of the fit (Nowok, Raab, and Dibben 2016). Since the p-values for both synthetic datasets are above any reasonable significance level, we fail to reject the null that the synthesis method retains these relationships.

Table 7: Linear Model Comparison Metrics Between True and Synthetic Datasets: The high mean confidence interval overlap, low mean absolute standardized difference, and non-significant lack-of-fit test p-values illustrate the synthetic datasets' capacity to preserve the true data's multivariate relationships.

Synthetic	Mean CI Overlap	Mean Abs Std Diff	L.O.F. Test Stat	L.O.F. P-value
Synthetic 1	0.70	1.16	11.63	0.23
Synthetic 2	0.74	1.01	6.74	0.66

Figure 7 and Figure 8 accentuate the synthetic datasets' capacity to preserve the true data's multivariate relationships when predicting BMI. The Z-values for the coefficients for each predictor in these synthetic datasets strongly resemble those in the true data, with the confidence intervals for each coefficient overlapping, as expressed above. Despite establishing synthetic preservation, how well has the synthetic data decreased disclosure risk?

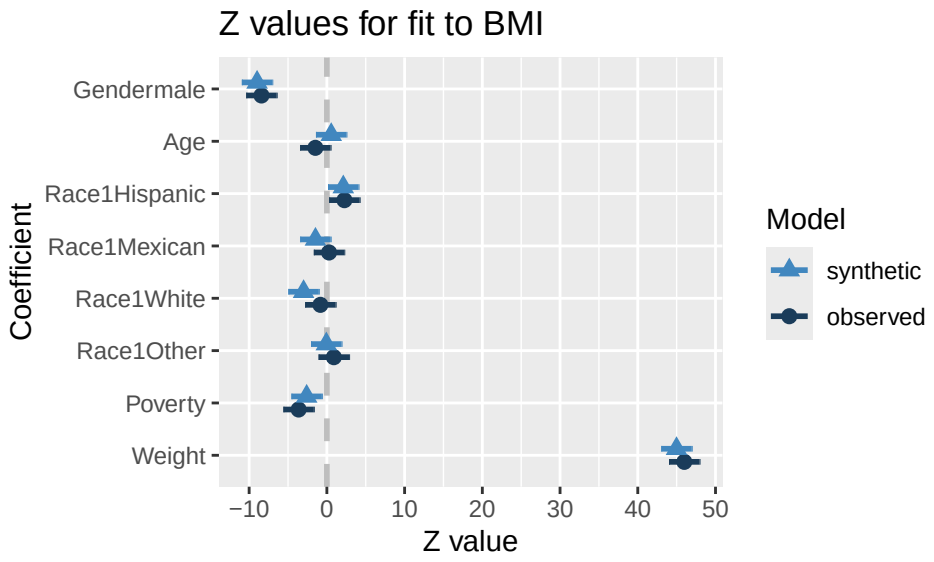

Figure 7: Estimates and 95% Confidence Intervals for Z Statistics from a Linear Regression of BMI for Observed and Synthetic Dataset 1: These Z values among the observed and synthetic data are very similar, accentuating the findings from Table 5.

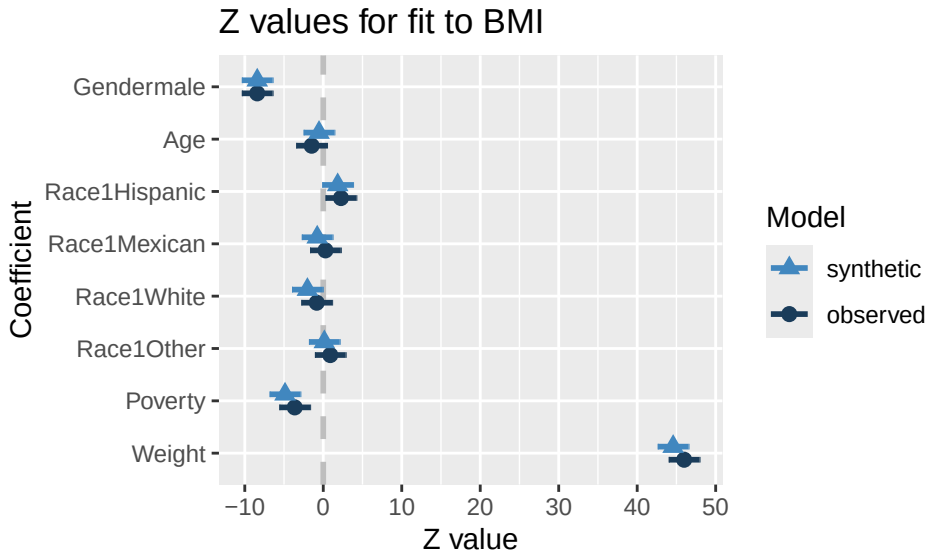

Figure 8: Estimates and 95% Confidence Intervals for Z Statistics from a Linear Regression of BMI for Observed and Synthetic Dataset 2: These Z values among the observed and synthetic data are very similar, accentuating the findings from Table 6.

The coefficients from the linear model of the first synthetic dataset are applied to the true data to predict individuals' actual BMI values. A small MSE and MAE would indicate that the synthetic model predicts BMI values accurately. Conversely, larger MSE and MAE values would signify greater deviation between the synthetic model's predictions and the actual data, making it more difficult for an intruder to reverse-engineer the data or infer sensitive details about the true dataset.

Focusing on MAE, Table 8 highlights that the absolute difference between predicted and observed BMI values averages 2.11 units. For MSE, which is more sensitive to outliers, the average squared difference between the predicted and observed BMI values is 7.26. Considering that BMI values in the true data range from 12.90 to 63.30, these errors are moderately small. This magnitude signals that the true data's relationships are preserved while simultaneously complicating any effort towards reidentification.

Table 8: Error Metrics for Predicted BMI in True Data Using Synthetic Model 1: The moderately small MSE and MAE highlight the synthetic data's ability to balance two critical objectives: preserving key relationships and complicating reidentification efforts. This contrasts very small MSE and MAE values, where the synthetic data is overly accurate and potentially compromising, or much larger values, where the synthetic data would bear little resemblance to the original dataset.

Metric	Value
Mean Squared Error	7.26

The distribution of outliers in the true data is visualized in Figure 9. The race categories with the highest concentration of outliers are White and Black, whereas Hispanic and Other categories exhibit comparatively low counts. Interestingly, the distribution of outliers does not vary significantly across Gender categories, and no outliers are observed for the Poverty variable.

If the synthetic data were to precisely replicate these sensitive patterns, it could inadvertently expose private information about that group (or individual). Consider an example from the healthcare domain: suppose a rare genetic condition, uncommon in the general population but more prevalent in a specific ethnic group, appears as outliers in the true data. If the synthetic data reproduces this outlier pattern exactly, it could lead to the identification of the affected group or individuals within that group, revealing their health condition. While it is important for the synthetic data to reflect the overall distribution of the true data for analytical utility, it must carefully balance this realism with safeguards to prevent the tracing of such patterns back to individuals.

The distribution of outliers in the first synthetic dataset is shown in Figure 10. Apart from White, which retains a higher volume of outliers, there are no substantial differences among the other races, as the magnitude of outlier disparities between them is less pronounced. The absence of substantial differences among these race categories suggests that the synthetic data attempts to balance the distribution of outliers more evenly across the groups, reducing disproportionate representation. The Poverty variable, which had no outliers in the true data, now contains at least one in every Gender and Race1 combination except Mexican males.

Conclusion

This report began with a thorough exploration of the risk analysis landscape within the statistical disclosure control framework. By providing background information and defining existing disclosure risk measures in accordance with Taylor, Zhou, and Rise (2017), it stressed the key methodologies and considerations vital for protecting sensitive data. The five distinct risk measures, derived from modeling methods detailed in Taylor, Zhou, and Rise (2017), are crucial for statistical agencies worldwide to adopt when assessing disclosure risk.

The subsequent sections built upon Taylor, Zhou, and Rise (2017)'s work, focusing specifically on one action that can be taken when high risk is identified in a dataset – generating synthetic data. The `synthpop` package was introduced and applied to a real-world dataset, the National Health and Nutrition Examination Survey, to generate synthetic data that aimed at preserving the true data's key characteristics while minimizing disclosure risk. Univariate, bivariate, and multivariate analyses were conducted to assess the synthetic data's capacity to maintain these relationships. Results from the scatterplots, contingency tables, frequency distributions, and linear model coefficient comparisons demonstrated that the synthetic data effectively imitated

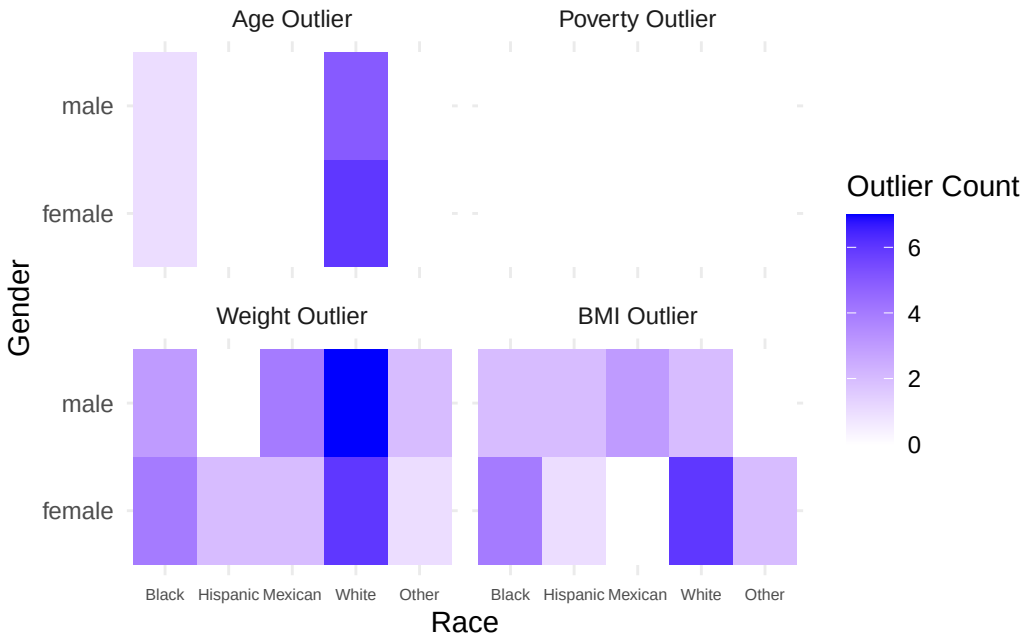

Figure 9: Distribution of Outliers in True Data across Gender and Race: There is a noticeable concentration of outliers in the White and Black groups compared to the Hispanic and Other groups. Ideally, the synthetic data should obscure these patterns to prevent the tracing of sensitive information back to such groups/individuals.

Figure 10: Distribution of Outliers in Synthetic Data across Gender and Race: Other than White, the differences in outlier concentration among racial groups is less pronounced.

the true data, with the scatterplots also depicting the reduced presence of outliers. Despite the absence of overlapping observations between the true and synthetic datasets, reduction in disclosure risk was still analyzed by applying the synthetic model to the true data to predict BMI values. The moderately small MSE and MAE values illustrated the synthetic data's ability to balance preserving important statistical properties with complicating reidentification efforts. Furthermore, comparing the heatmap of outliers in the true and synthetic data revealed that the synthetic data muted the concentration of outliers among racial groups (with the exception of White) – a pivotal characteristic in safeguarding personal, private information.

Limitations

The results presented above relied on linear regression to explore multivariate relationships, which may not fully capture the complexity inherent in the data. The Residual vs. Fitted plot revealed signs of heteroscedasticity, suggesting a potential violation of the assumption of constant variance in the error terms. Alternative modeling approaches, such as generalized linear models or machine learning algorithms, could provide a more comprehensive understanding of the true data's underlying relationships.

Within the linear regression framework, only BMI was considered as the outcome variable. To gain better insight into the synthetic data's ability to maintain key statistical properties,

additional outcome variables could be explored.

The synthetic data generation process was conducted assuming the disclosure risk was high in the true data, which may have not been the case. Although this assumption is plausible given the large number of unique counts in the true data, if the disclosure risk was not as high as assumed, the generation process might have been overly cautious and could have unnecessarily compromised the utility of the synthetic data.

There was also the assumption that outlier patterns across races were dangerous and needed to be obscured. However, in some cases, outlier patterns may be essential for answering specific research questions. The disclosure control methods/synthetic data generation process should try to tailor to the specific needs of the research questions that are being addressed.

Future Directions

Future work could consider the optimal MAE and MSE thresholds for determining the synthetic data's utility and disclosure risk. The current analysis holds that moderate errors are ideal, symbolizing the middle ground between emulating the true data and preventing reidentification. Discussions on the acceptable range of these errors could offer a more concrete framework for evaluating synthetic data quality. How small is too small, and how large is too large? Is this an appropriate statistic to consider, or are there other measures that could be computationally equivalent and more informative?

Future research could also take into account several other methods to alter data prior to release. Taylor, Zhou, and Rise (2017) mention nonpertrubative masking methods that do not distort the data, like reducing detail by categorizing variables or suppressing some variables. Economic Statistics (2024) reports on the implications of new privacy protection methods for economic research, and includes differential privacy as a growing statistical disclosure limitation method. Chapter 3 of Hundepool et al. (2012) covers disclosure control methods in intricate detail, which could be a valuable resource for future research and implementation in this domain. As the desire for digital information expands and computing firepower advances, disclosure risk and disclosure control methods will be a critical area of development – one that will be exciting to follow.

References

Benitez, Kathleen, and Bradley Malin. 2010. "Evaluating Re-Identification Risks with Respect to the HIPAA Privacy Rule." *J Am Med Inform Assoc* 17 (2): 169–77. https://doi.org/10.1136/jamia.2009.000026.

Buttice, Claudio. 2022. ""10 Quotes about Tech Privacy That'll Make You Think"." 2022. https://www.techopedia.com/10-quotes-about-tech-privacy-thatll-make-you-think/2/33713.

Economic Statistics, Committee on. 2024. ""AEAStat Statement on Implications of New Privacy Protection Methods for Economic Research"." 2024. https://www.aeaweb.org/content/file?id=20449.

Emam, Khlaed El, Fida K Dankar, and Elizabeth Jonker. 2013. "Evaluating the Risk of Patient Re-Identification from Adverse Drug Event Reports." *BMC Medical Informatics and Decision Making* 13 (114). https://doi.org/10.1186/1472-6947-13-114.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul de Wolf. 2012. *Statistical Disclosure Control*. West Sussex, UK: John Wiley & Sons.

Nowok, Beata. "Synthpop: An r Package for Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control." https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_24_bnowok_synthpop.pdf.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. "synthpop: Bespoke Creation of Synthetic Data in R." *Journal of Statistical Software* 74 (11): 1–26. https://doi.org/10.18637/jss.v074.i11.

Pruim, Randall. 2015. *NHANES: Data from the US National Health and Nutrition Examination Study*. https://CRAN.R-project.org/package=NHANES.

Rinott, Yosef, and Natalie Shlomo. 2006. "A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation." *Privacy in Statistical Databases*, 82–93. https://doi.org/10.1007/11930242_8.

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society Series A* 181: 663–88. https://academic.oup.com/jrsssa/article/181/3/663/7072005.

Taylor, Leslie, Xiao-Hua Zhou, and Peter Rise. 2017. "A Tutorial in Assessing Disclosure Risk in Microdata." *Statistics in Medicine* 37 (25): 3693–3706. https://doi.org/10.1002/sim.7667.

Templ, Matthias, Alexander Kowarik, and Bernhard Meindl. 2015. "Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro." *Journal of Statistical Software* 67 (4): 1–36. https://doi.org/10.18637/jss.v067.i04.

Acknowledgements

I acknowledge the usage of generative AI tools like GitHub Copilot and ChatGPT. Copilot was used for minor code suggestions and ChatGPT assisted with the development of pivot functions, the frequency of distribution counts, the Residual vs. Fitted and Normal Q-Q plots, and the heatmaps. Specific notes and prompts can be found in the code (.qmd) as comments.