Utrecht University

Master's Thesis

MICE to Synthesize

Generating Synthetic Data Sets with MICE for the Analysis of Private Data

Author:
Mirthe Hendriks

First supervisor:
Dr. Gerko Vink

Second supervisor:
Prof. Dr. Stef van Buuren

July 2, 2021

Universiteit Utrecht

A Thesis submitted to the Board of Examiners in partial fulfilment of the requirements of the degree of Master of Science in Applied Data Science

Abstract

The aim of this paper is to present a proof-of-concept for generating high-quality partially synthetic data with multiple imputations in MICE for the analysis of privacy-sensitive data. The synthetic data framework finds its foundation in the multiple imputation of missing data, originally proposed by Rubin (1987; 1993) and Little (1993). We conduct a Monte Carlo simulation to generate synthetic data, by overimputing observed values with multiple draws from the posterior predictive distribution.

We found that synthetic data generated with CART in MICE yields unbiased and confidence valid estimates in different scenarios, given suitable pooling rules. The simulation study shows that, with the use of CART as an appropriate imputation method, the synthetic data (1) preserves the univariate statistical properties, and (2) yields the same multivariate statistical inferences as the observed data. Moreover, the synthetic data seems to be indistinguishable from the observed data. We therefore conclude that MICE might be used to generate high-quality partially synthetic data. The results are promising, especially with regard to the practical implications of easily accessible synthetic data on privacy-preservation in data ownership and sharing.

1 Introduction

In this digital age, privacy-preservation in data sharing continues to be a topic of undeniable importance. Digitized information has grown explosively, and with it have privacy and confidentiality concerns (Donaldson and Lohr, 1994). While the protection of personal data is necessary, it restricts the public distribution of information and thus complicates the process of research and policy making. There is an increasing aspiration to share real world data, of which the open science movement is an example (Gewin, 2016; Utrecht University, 2020). However, it is challenging to use data to its fullest potential without compromising confidentiality. In addressing this challenge, synthetic data may present a solution.

The synthetic data framework finds its foundation in the multiple imputation of missing data, originally proposed by Rubin (1993) and Little (1993). To generate the synthetic data, observed values are replaced with multiple imputations. The resulting synthetic units contain no real personal information, and thus protect confidentiality. The underlying idea is that an appropriate imputation model should be able to account for the process that created the observed data, preserve the statistical properties and relations in the data, and accurately capture the uncertainty about these relations (Rubin, 1993; Reiter, 2003; Van Buuren, 2018). Nevertheless, the process of generating high-quality synthetic data remains challenging and laborious. This simulations study examines the performance of the R-package MICE in generating high-quality partially synthetic data.

1.1 Historical overview of synthetic data

In past decades and to this day, there have been a number of techniques to deal with privacy-sensitive data. These statistical disclosure techniques have offered some safeguards, but they all have certain limitations (Drechsler, 2011). Information-reducing techniques, like redaction, aggregation and grouping, are disadvantageous because the loss of information

limits further statistical analysis (Drechsler, 2011; Little, 1993; Rubin, 1993). Encryption and secure multiparty computation techniques are computationally complex and require communication overhead (Kantarcolu and Vaidya, 2018; Yin, Kaku, Tang, and Zhu, 2011). Alternatively, perturbation techniques that adjust values in the data at the micro-level to protect confidentiality, are prone to lower the data quality because complicated modifications may distort relationships among variables (Reiter, 2002; Reiter and Drechsler, 2010). These issues stressed the need for a privacy-preserving technique without such limitations.

It gave rise to a new stream of research: multiply imputed synthetic data sets, which originated in the *Journal of Official Statistics* in 1993 with the articles from Rubin (1993) and Little (1993). Rubin (1993) proposed a research effort to release fully synthetic data generated with multiple imputation, the successful methodology he initially developed to handle missing data (Rubin, 1987). Conceptually and in terms of procedure the approach is similar, but rather than imputing missing values, observed values are replaced with multiple imputations. The synthetic data sets for release would comprise of synthetic values that no longer contain personal or sensitive information. Rubin (1993, p. 463) argued that multiple-imputed synthetic data "looks just like replicated actual microdata", and thus preserves the users' ability to obtain valid inferences for legitimate estimands with standard statistical methods.

Little (1993) introduced a similar but less radical approach for partially synthetic data. While the fully synthetic data approach treats the entire unsampled population as missing data, and subsequently draws new samples out of the population based on an imputation model constructed on the observed sample (Rubin, 1993), the partially synthetic data approach *overimputes* the values from the observed sample (Reiter, 2003). Little (1993) further discusses the possibility to synthesize only part of the data, specifically sensitive values at high risk of disclosure or values of key identifiers. Nowadays, these approaches comprise the theoretical foundation of multiply imputed synthetic data.

It took another ten years for the complete theory of synthetic data to develop (Drechsler, 2011). Raghunathan, Reiter and Rubin (2003) developed the correct procedures for deriving valid inferences for fully synthetic data sets under different sampling scenarios (Reiter, 2002; 2005a), and Reiter (2003; 2005b) published the procedures for partially synthetic data. In addition, Fienberg, Makov, and Steele (1998) generated synthetic categorical data using bootstrapping and the cumulative distribution function. Further developments followed, including nonparametric imputation methods based on CART (Reiter, 2005c) and random forests (Caiola and Reiter, 2010), and methods to measure disclosure risk (Reiter and Drechsler, 2010).

While to all appearances synthetic data offers a valuable and valid approach for circumventing issues of privacy and data ownership, synthetic data sets are not yet widely used in practice. Why is this the case? Generating high-quality synthetic data can be laborious and complex. An appropriate imputation model is imperative for synthetic data to yield valid statistical inferences and obtaining such high-quality synthetic data requires considerable effort on behalf of the imputer. Therefore, this paper inquires into the performance of accessible software to more easily generate synthetic data.

1.2 Aim of this research

This paper contributes to the current synthetic data research by exploring the performance of easy-to-use software: the R-package MICE for Multivariate Imputation by Chained Equations, in generating partially synthetic data (Van Buuren and Groothuis-Oudshoorn, 2011). The MICE package is widely-used and popular for its multiple imputation method to deal with missing data. The functionality of MICE allows for generating multiple imputations by Fully Conditional Specification, and analyzing the imputed data (Van Buuren and Groothuis-Oudshoorn, 2011). Subsequently, Reiter's procedures (2003) and Volker's functions (2021) may be used to pool the synthetic estimates. Potentially, MICE might facilitate an efficient and accessible way of generating synthetic data to be released for public-use, while honoring confidentiality constraints.

The aim of this paper is to establish a proof-of-concept for generating high-quality partially synthetic data with multiple imputations in MICE for the analysis of privacy-sensitive data. We propose a Monte Carlo simulation to generate the synthetic data, based on actual observed data. The observed data comes from the publicly available Pima Indians Diabetes (PID) data set (National Institute of Diabetes and Digestive and Kidney Diseases, 2016). To assess the synthetic data we consider whether it preserves the univariate statistical properties of the observed data as well as the multivariate statistical inferences of a selected analysis model. Ideally, a prediction model should not be able to distinguish the synthetic from the observed data. Therefore, this paper will answer the following two research questions:

1. Do the synthetic data and the observed data yield the same statistical inferences, i.e. are the synthetic estimates unbiased and confidence valid?

2. Can we discriminate between the synthetic data and the observed data with a prediction model?

This simulation study is outlined in accordance with the ADEMP structure (i.e. defining aims, data-generating mechanisms, estimands, methods, and performance measures) (Morris, White, and Crowther, 2019). First, we report the data-generating mechanism for partially synthetic data in MICE. Second, we specify the statistical estimands of interest in analyzing and comparing the synthetic and observed data. Third, we reflect on appropriate imputation methods to obtain synthetic estimates, specifically predictive mean matching (PMM) and classification and regression trees (CART). Subsequently, we define the performance measures to assess the synthetic data. The results section reports the findings of this simulation study. In the discussion, we situate these findings in the context of current research, conclude with an answer to the research questions, and reflect on the strengths and limitations of this study. Finally, we formulate recommendations for practice and research.

2 Simulation

This simulation study generates and analyzes synthetic data with MICE (Van Buuren and Groothuis-Oudshoorn, 2011, version 3.13) through the process of multiple imputation in R (R Core Team, 2021, version 4.0.4). The set-up for this Monte Carlo experiment, with 1000 simulations ($nsim = 1000$) and different imputation methods, is specified below. The complete R script of this study is available from https://github.com/MMJHendriks/Synthetic_ADS/tree/main/SimulationStudy_MHendriks.

There are two main stages to simulating synthetic data: we (1) generate multiple synthetic data sets by *overimputing* the observed values using an imputation method, and (2) pool the estimates for each synthetic data set and obtain correct standard errors (Murray, 2018; Reiter, 2003). The intention is to generate synthetic data that yields estimates that are unbiased and confidence valid (Neyman, 1934) as well as indistinguishable from the observed data (Drechsler, 2011).

2.1 Data-generating mechanism: synthetic data in MICE

This section describes the process of generating partially synthetic data with multiple imputation in MICE. The MICE package was originally developed to handle missing data (Van Buuren and Groothuis-Oudshoorn, 2011), where $Y = (Y_{obs}, Y_{mis})$ denotes the data of the sample, and the observed data Y_{obs} is used to impute the missing data Y_{mis}. In the simulation, this approach is generalized to the synthetic data framework. As such, the focus is not on imputing missing values, but on taking draws from the posterior predictive distribution to replace, i.e. *overimpute* the observed values.

Let us first report some notations based on Reiter (2003) and Drechsler (2011). For simplicity, assume that the sample data are completely observed $Y = (Y_{obs})$. Let $Z_j = 1$ if unit j is selected to be replaced and $Z_j = 0$ if unit j is not to be replaced, with $Z_j = (Z_1, \ldots, Z_n)$ where n is the number of records in the observed data. Accordingly, the data $Y = (Y_{nrep}, Y_{rep})$ contains all the values to be replaced Y_{rep}, and all to remain unchanged Y_{nrep}. We use Y^i_{rep} for the replacement values in synthetic data set i. To generate Y^i_{rep}, we simulate values form the posterior predictive distribution $(Y^i_{rep}|Y, Z)$. Multiple imputation entails that we repeat this process m times, to generate $D^i = (Y_{nrep}, Y^i_{rep})$ for $i = 1, \ldots, m$. Accordingly, m denotes the number of versions of the imputed data. The result is the synthetic data $D = D^1, \ldots, D^m$ free to be released to the public.

The MICE algorithm allows for the generation of synthetic data with the `mice`-function. The function requires the specification of a variety of arguments; the `data`, the number of imputations `m`, what of the observed data to overimpute with `where`, what imputation `method` to use for each column in the data, and what predictor variables to use with `predictorMatrix` (Van Buuren, 2021, p. 73-75). More specifically, the `where` parameter can be used to overimpute either a subset or the observed data set as a whole. For a detailed report of the function consult https://cran.r-project.org/web/packages/mice/mice.pdf.

In this simulation, the publicly available Pima Indians Diabetes (PID) database, originally from the National Institute of Diabetes and Digestive and Kidney Diseases (2016), is

used to generate and evaluate the synthetic data. This data sets contains eight predictor variables; number of pregnancies, glucose concentration, blood pressure (in mm Hg), skin thickness (mm), insulin (mu U/ml), BMI, diabetes pedigree function, and age, and the binary outcome variable diabetes. The data contains $n = 768$ observations of female patients, 21 years old or above, of Pima Indian heritage. For a number of the predictor variables, the value of zero was recorded in place of missing data (Hayashi and Yukita, 2016).

For this reason, the simulation is performed for two scenarios. In the first scenario, the missing data is imputed once with MICE to create a complete data set. In the second scenario, the values of zero are not imputed, instead we consider these values to be part of the data. The predictor variables in question are taken to represent semicontinuous variables with point masses at zero and a continuous distribution among the remaining responses (Vink, Frank, Pannekoek, and Van Buuren, 2014). We run this scenario because it might be more complicated to synthesize the semicontinuous data.

In both scenarios, the objective is to synthesize the PID data set as a whole, so with the `where` argument we select all values to be overimputed. In accordance with the advice from Van Buuren (2018), we set the number of imputations to $m = 5$. `Method` and `predictorMatrix` are specified later on. Given the number of simulations, the `mice`-function outputs a list of 1000 synthetic data sets each with $m = 5$ imputations for analysis.

2.2 Scientific estimands

Once we have generated the synthetic data, we are interested in whether this synthetic data preserves the univariate statistical properties and multivariate inferences from the observed data. To assess this we compare the scientific estimands. Consider an analyst interested in scientific estimand Q, for example a regression coefficient. In the data analysis process, the inferences for this estimand from the synthetic data should be similar to the inferences from the original data. First, to see if the data sets have similar univariate statistical properties we compute the mean, standard deviation, median, skew, kurtosis, and standard error for each variable in the data.

Second, we conduct a multivariate analysis to examine whether the synthetic data preserves the relationships between the variables. We select an analysis model to compare the statistical inferences for the synthetic and the observed data. Specifically, we consider the regression coefficients for the following logistic regression equation:

$$\text{Diabetes} = \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{Glucose} + \beta_3 \text{Pregnancies}$$

To generate synthetic data that yields unbiased scientific estimands, an appropriate imputation method is needed.

2.3 Method

2.3.1 Imputation method

Selecting an adequate imputation method is particularly relevant to ensure that synthetic estimates are of high quality (Drechsler, 2011; Murray, 2018). In this simulation, we test two different methods: (1) predictive mean matching (PMM) for the continuous variables and logistic regression for the binary outcome, and (2) classification and regression trees (CART) for each variable in the data. These imputation methods are applied to the observed data (later on referred to as 'complete PID') as well as to a bootstrapped sample of this data.

First, PMM is a widely used method for generating hot-deck imputations (Little, 1988). The method replaces the observed values by means of a nearest-neighbor donor, with distance based on the expected values of a variable conditional on the other observed covariates (Vink et al., 2014). For the procedure of PMM in MICE we refer to Van Buuren (2018, Algorithm 3.1; Rubin, 1987, p. 167). Heeringa, Little and Raghunathan (2002) and Yu, Burton and Rivero-Arias (2007) found that, for the multiple imputation of missing data, PMM yields acceptable estimates and preserves the original data distributions. Additionally, Vink et al. (2014) found that PMM is an appropriate method for imputing semicontinuous data. However, a disadvantage of PMM is that it might not be well suited to model nonlinear relations, such as interaction effects, in the data.

The second method, CART is able to capture more complex patterns in the data, because it does not assume a parametric distribution (Doove, Van Buuren, and Dusseldorp, 2014). In essence, CART sequentially splits the data into non-overlapping homogeneous subsets. Recursive partitioning selects the split that is most predictive of the outcome variable conditional on the predictor variables. The flexibility of CART is a benefit, as earlier studies found that this imputation method yields the least biased parameter estimates in case of interaction effects in the data (Burgette and Reiter, 2010). We refer to Doove et al. (2014) for a detailed discussion.

Additional parameters to specify for the use of CART are: the complexity parameter `cp` and the minimum number of observations in any terminal node `minbucket`. Increased complexity in the imputation model might limit the bias in the synthetic estimates (Doove et al., 2014). In the simulation, we test multiple parameter specifications: `cp` fixed at $1e-4$ and $1e-32$, `minbucket` set to 3, and the number of iterations `maxit` at 1 and default 5 in the bootstrapped sample.

Furthermore, the `predictorMatrix` specifies the set of predictor variables to be used for each target column in the imputation method. For each variable Y_j, all remaining variables Y_{-j} are used as predictors; let $Y_j|Y_{-j}$ be the distribution of Y_j conditional on all columns of Y except Y_j.

2.3.2 Pooling method

After generating the synthetic data sets we pool the m parameter estimates into a final point estimator $\bar{Q}$ and its associated variance. Let $\hat{Q}^i$ be the estimate of Q from the i-th synthetic data set D^i, for $i = 1, ..., m$, with $\bar{U}^i$ the estimated variance-covariance matrix of

$\hat{Q}^i$ (Rubin, 1987). We then combine the m estimates into a single pooled estimate $\bar{Q}$.

Given partially synthetic approaches assume the observed data sample as a finite population, adopted pooling rules are required to obtain appropriate variance estimates (Vink and Van Buuren, 2014). Reiter (2003; Reiter and Raghunathan, 2007) developed the procedures to obtain valid inferences from multiply-imputed partially synthetic data sets. We refer to Rubin (1987, p. 76) who defined the needed quantities. Then, for the partially synthetic data, Reiter (2003, p. 5; Drechsler, 2011, p. 54) estimates the variance of $\bar{Q}_m$ as:

$$T_p = B_m/m + \bar{U}_m,$$

where $\bar{U}_m$ is the average within-imputation variance, and B_m the average between-imputation variance for a finite number of m imputations. Additionally, B_m/m represent the variance because the number of imputations is finite. In this simulation, we use Volker's (2021) partially synthetic data pool function, which is bases on this estimation of variance.

Since some uncertainty remains about what pooling rules yield the correct inferences in practice, we might also consider Rubin's rules (1987, p. 76) to pool the estimates. In the missing data context, these are the appropriate pooling rules. According to these rules, variance is estimates as:

$$T_p = \bar{U}_m + B_m + B_m/m,$$

where an additional B_m is used to capture the between-imputation variance. The pooling methods assume that, under repeated sampling, the estimates $\hat{Q}$ are normally distributed around the population value Q (Rubin, 1987). After obtaining these pooled synthetic estimates and their variances, we assess their performance.

2.4 Performance measures

For this simulation, we can evaluate the synthetic data because we have the 'true' observed data. The following two performance measures are used to assess the quality of the estimates: unbiasedness and confidence validity (Neyman, 1934). First, bias is calculated as the average difference between the estimates and true value of the parameter (Doove et al., 2014; Van Buuren, 2018). Ideally, the bias is close to zero. Bias is calculated for each aforementioned scientific estimand, univariate and multivariate.

Second, to quantify the confidence validity of the synthetic data, we evaluate the confidence interval coverage (CIC) and confidence interval width (CIW) (Neyman, 1934). CIC is the proportion of simulation repetitions in which the 95% confidence interval around pooled synthetic estimate $\bar{Q}$ contains the true estimand Q (i.e. the estimate from the observed PID data set) (Morris et al., 2019; Van Buuren, 2018). A synthetic estimate is considered confidence valid if it yields a nominal coverage of 95% or higher (Neyman, 1934).

The CIW is calculated by subtracting the lower bound of the 95% confidence interval from the upper bound and can be seen as an indicator of statistical efficiency (Van Buuren, 2018). Ideally, the CIW would be as small as possible, provided that the CIC does not fall below the nominal level of 95%. In that scenario, a smaller CIW would correspond to higher statistical power. The CIC and CIW are calculated for the regression coefficients.

Subsequently, we assess the ability to discriminate between the observed and synthetic data using a prediction model. The data sets are merged by combining a sample of half the synthetic data randomly with half of the observed data, thereby adding the indicator column Synth_or_True. In turn, a generalized linear model is trained on 50% of this data and tested on the remaining 50%. Ideally, an accuracy of 0.5 would indicate the prediction performs no better than random and the synthetic data is indistinguishable from the observed data.

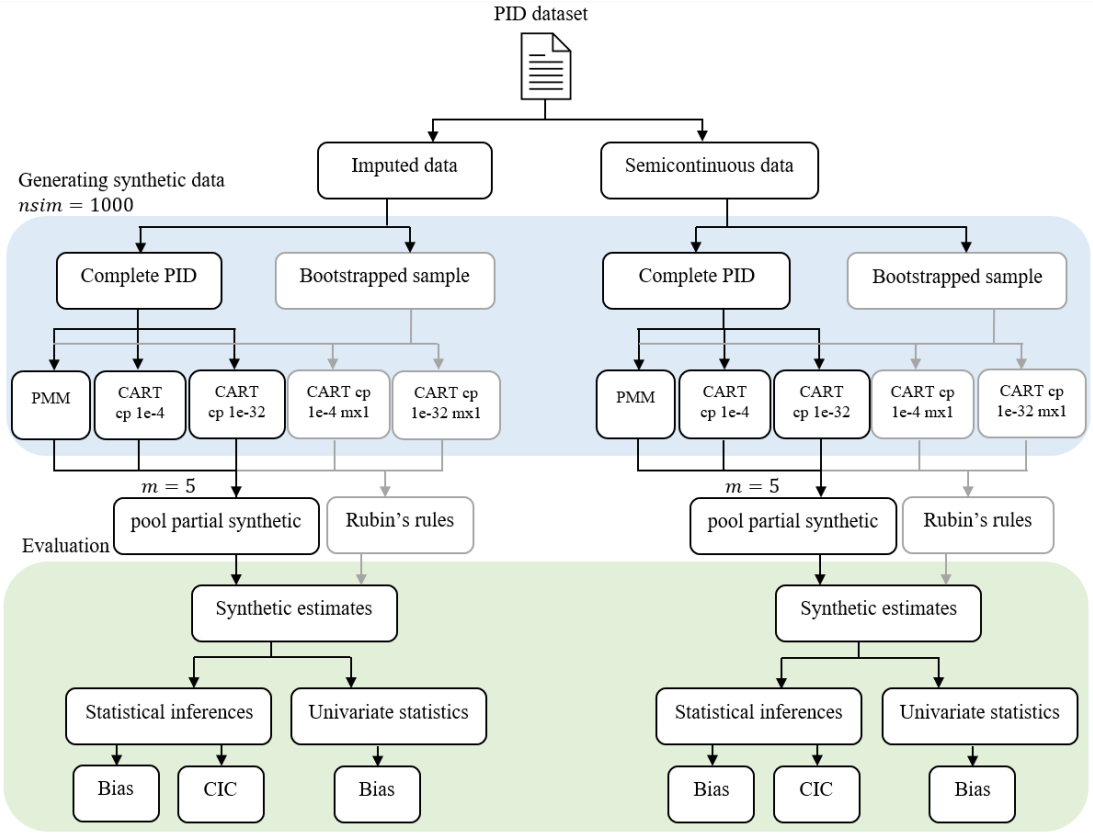

Figure 1: Flowchart simulation study

3 Results

The results show the performance of the MICE algorithm in generating partially synthetic data under different conditions. Firstly, we discuss whether the synthetic data preserves the statistical inferences and univariate statistical properties of the observed PID data. This section is structured in accordance with the flowchart in Figure 1 and reports the two simulation scenarios; the imputed data and the non-imputed semicontinuous data. For both scenarios, we reflect on the results from the complete PID data set and on a bootstrapped

sample of this data. The results for PMM are not discussed, because the synthetic estimates are highly negatively biased and CART outperforms PMM in every aspect. Tables 1 to 5 show the results for the CART-generated synthetic data; the pooled synthetic estimates for $m = 5$ averaged over 1000 simulations for the differ parameter specifications.

Secondly, we report on the distinguishability between the synthetic and observed data.

3.1 Statistical inferences

3.1.1 Imputed data

First, for the scenario where the missing data is imputed to form the complete PID data set, CART as an imputation method yields unbiased and confidence valid synthetic estimates. Table 1 shows the coefficients in the analysis model, their true estimates from the observed data, and the pooled synthetic estimates averaged over 1000 simulations. Bias is the difference between these two estimates, where low scores indicate that the synthetic regression coefficients are on average not very different from the observed coefficients. Based on the results in Table 1, we find that in this scenario, CART yields synthetic estimates that are unbiased for both complexity parameter specifications (`cp` set to $1e-4$ and $1e-32$).

Table 1: Bias and confidence validity of the synthetic data generated with CART for scenario 1 imputed data based on the complete PID data set and partial synthetic data pooling rules

	term	True est	Syn est	Bias	CIW	CIC
cp 1e-4	(Intercept)	-8.864	-8.302	0.562	2.820	1.00
	BMI	0.090	0.092	0.001	0.060	1.00
	Glucose	0.037	0.033	-0.004	0.014	0.98
	Pregnancies	0.137	0.123	-0.015	0.112	1.00
cp 1e-32	(Intercept)	-8.864	-8.312	0.552	2.821	1.00
	BMI	0.090	0.092	0.001	0.061	1.00
	Glucose	0.037	0.033	-0.004	0.014	0.98
	Pregnancies	0.137	0.123	-0.014	0.112	1.00

Additionally, Table 1 shows that the synthetic estimates for intercept, β_{BMI}, and $\beta_{\text{pregnancies}}$ yield a CIC of 100% and of 98% for β_{glucose}. Essentially, this CIC score implies that, for BMI for example, 100% of the 1000 synthetic estimates have a 95% confidence interval that covers the true estimate. Given the situation where the sample happens to be the finite population, coverages of 100% can be expected. This is because when drawing an estimate from a population, instead of a sample of that population, all units are recorded, so there is no sampling variation and the true estimate is known. Nominal coverages higher than 95% imply that the synthetic estimates are confidence valid (Neyman, 1934), but that variance estimates are inefficient, resulting in lower statistical power.

Likewise, for the bootstrapped sample of the PID data set, CART also yields unbiased synthetic estimates, displayed in Table 2. The left column in the table reports the different parameter specifications for the use of CART on the bootstrapped sample: both levels of

complexity `cp` and the number of iterations `maxit` set to 1 or default 5. The bias scores and coverage rates per regression coefficient are close to identical under the different conditions.

The estimates pooled with the partially synthetic data rules (Reiter, 2003; Volker, 2021), however, are not confidence valid, as indicated by the CIC rates ranging from 88% to 92% for the regression coefficients. The CICs for the intercepts are even lower. In Table 2 these coverages are displayed in the left CIC column under 'Partially synthetic'. Undercoverage in case of unbiased estimates might suggest that the estimation of variance is inaccurate and that this partially synthetic data pooling method (Reiter, 2003; Volker, 2021) might not be suitable in this case. Since the pooling method assumes the estimates to be normally distributed, we made sure to check this assumption, and found that it holds.

Table 2: Bias and confidence validity of the synthetic data generated with CART for scenario 1 imputed data based on a bootstrapped sample of the PID data set pooled with partial synthetic data and Rubin's rules

					Partially synthetic		Rubin's rules	
	term	True est	Syn est	Bias	CIW	CIC	CIW	CIC
cp 1e-4	(Intercept)	-8.864	-8.174	0.690	2.736	0.82	9.657	0.91
	BMI	0.090	0.081	-0.010	0.059	0.90	0.193	0.96
	Glucose	0.037	0.035	-0.002	0.014	0.89	0.054	0.96
	Pregnancies	0.137	0.122	-0.015	0.112	0.90	0.380	0.96
cp 1e-32	(Intercept)	-8.864	-8.186	0.679	2.739	0.82	9.885	0.93
	BMI	0.090	0.081	-0.009	0.059	0.90	0.195	0.96
	Glucose	0.037	0.035	-0.002	0.014	0.88	0.054	0.96
	Pregnancies	0.137	0.122	-0.015	0.112	0.92	0.376	0.96
cp 1e-4 maxit 1	(Intercept)	-8.864	-8.184	0.680	2.751	0.83	10.755	0.94
	BMI	0.090	0.081	-0.009	0.059	0.90	0.198	0.97
	Glucose	0.037	0.035	-0.002	0.014	0.88	0.056	0.96
	Pregnancies	0.137	0.122	-0.015	0.111	0.91	0.381	0.97
cp 1e-32 maxit 1	(Intercept)	-8.864	-8.183	0.681	2.742	0.83	9.852	0.93
	BMI	0.090	0.081	-0.009	0.059	0.89	0.187	0.95
	Glucose	0.037	0.035	-0.002	0.014	0.88	0.054	0.96
	Pregnancies	0.137	0.122	-0.015	0.112	0.91	0.380	0.96

For this reason, the synthetic estimates were also pooled using Rubin's rules (1987), which increases the CIC scores as reported in Table 2 under 'Rubin's rules'. These pooling rules do not change the synthetic estimates nor the bias, but the estimation of variance is different. As mentioned in the method section, Rubin's pooling rules capture additional between-imputation variance (Rubin, 1987). Accordingly the CIW is wider, and therefore the CIC scores are higher. Specifically, it yields coverage rates between 95% and 97% for the regression coefficients, resulting in confidence valid synthetic estimates.

3.1.2 Semicontinuous data

This subsection reports the results for the second scenario, where synthetic data is generated with CART based on the semicontinuous PID data set. In the first scenario, the CART-generated synthetic data yields unbiased estimates. This second scenario functions as a check to see whether the performance of CART is relatively consistent in the case of non-imputed, semicontinuous data. Again, we discuss the synthetic data generated for the complete PID data set and a bootstrapped sample of this data (see Figure 1).

When ran on the complete PID data set, CART performs less well compared to the first scenario. As Table 3 shows, the synthetic estimates for β_{glucose} and $\beta_{\text{pregnancies}}$ are slightly biased at both complexity parameter specifications. This is noticeable, especially when the bias is considered relative to the true estimates; the synthetic estimate for glucose has a bias of approximately 13%, and pregnancies has a bias of approximately 23%. As might be expected given the slight bias in the synthetic data, the CIC scores for intercept 89% and β_{glucose} 90% indicate a small undercoverage and confidence invalidity. Nevertheless, CART still yields confidence valid synthetic estimates for the other two regression coefficients.

Again, the difference between `cp` set to $1e-4$ and $1e-32$ is neglectable, suggesting that more complexity does not reduce bias in this simulation.

Table 3: Bias and confidence validity of the synthetic data generated with CART for scenario 2 semicontinuous data based on the complete PID data set and partial synthetic data pooling rules

	term	True est	Syn est	Bias	CIW	CIC
cp 1e-4	(Intercept)	-8.864	-7.159	0.965	2.606	0.89
	BMI	0.090	0.076	-0.006	0.057	1.00
	Glucose	0.037	0.029	-0.005	0.014	0.90
	Pregnancies	0.137	0.106	-0.031	0.108	0.99
cp 1e-32	(Intercept)	-8.864	-7.146	0.979	2.594	0.90
	BMI	0.090	0.076	-0.006	0.057	1.00
	Glucose	0.037	0.029	-0.005	0.014	0.88
	Pregnancies	0.137	0.105	-0.032	0.108	0.99

In contrast, for the bootstrapped sample of the PID data set, the CART-generated synthetic data yields estimates that are unbiased, at both levels of complexity and maxit set to default 5 or 1 (see Table 4). Comparing the results for the bootstrapped sample in both scenarios, the performance of CART as an imputation model seems to be consistent.

Moreover, Table 4 shows CICs that are rather similar to those for the imputed data as reported in Table 2. With CIC scores ranging between 82% and 92%, these synthetic estimates are likewise confidence invalid when pooled with the partially synthetic data rules (Reiter, 2003; Volker, 2021). Whereas Rubin's rules (1987) yield coverage rates between 95% and 97% for the regression coefficients. Therefore, these pooled synthetic estimates are considered confidence valid, provided wider CIWs are necessary to obtain such nominal coverage rates.

Table 4: Bias and confidence validity of the synthetic data generated with CART for scenario 2 semicontinuous data based on a bootstrapped sample of the PID data set pooled with partial synthetic data and Rubin's rules

							Partially synthetic Rubin's rules	
	term	True est	Syn est	Bias	CIW	CIC	CIW	CIC
cp 1e-4	(Intercept)	-8.864	-7.519	0.605	2.608	0.82	11.115	0.94
	BMI	0.090	0.075	-0.007	0.056	0.89	0.234	0.96
	Glucose	0.037	0.032	-0.003	0.014	0.85	0.063	0.95
	Pregnancies	0.137	0.122	-0.015	0.110	0.91	0.379	0.97
cp 1e-32	(Intercept)	-8.864	-7.521	0.604	2.608	0.83	11.117	0.94
	BMI	0.090	0.075	-0.007	0.056	0.89	0.234	0.97
	Glucose	0.037	0.032	-0.003	0.014	0.85	0.064	0.95
	Pregnancies	0.137	0.123	-0.014	0.110	0.92	0.364	0.97
cp 1e-4 maxit 1	(Intercept)	-8.864	-7.534	0.590	2.615	0.83	11.708	0.95
	BMI	0.090	0.075	-0.007	0.057	0.90	0.236	0.96
	Glucose	0.037	0.032	-0.002	0.014	0.85	0.065	0.96
	Pregnancies	0.137	0.123	-0.014	0.110	0.91	0.368	0.96
cp 1e-32 maxit 1	(Intercept)	-8.864	-7.522	0.602	2.618	0.83	11.745	0.94
	BMI	0.090	0.075	-0.007	0.057	0.89	0.238	0.96
	Glucose	0.037	0.032	-0.003	0.014	0.85	0.061	0.95
	Pregnancies	0.137	0.122	-0.015	0.110	0.90	0.371	0.96

Overall, these results show that CART is able to generate synthetic data that yields estimates that are unbiased and confidence valid when pooled with Rubin's rules. Therefore, we might say that in this simulation, the CART-generated synthetic data preserves the multivariate statistical inferences of the observed data.

3.2 Univariate statistical properties

Additionally, we consider whether the synthetic data preserves the univariate statistical properties of the observed data. The univariate statistical properties; the mean, standard deviation, median, skew, kurtosis, and standard error of the synthetic data appear largely unaffected by bias for all variables in the data set. Table 5 reports these properties for the 'true' imputed PID data set and shows the bias in the synthetic data with regard to these properties. Specifically, Table 5 shows the results for the CART-generated synthetic data based on scenario one and the complete PID data set with `cp` set to $1e-4$.

We did not include the tables for the other conditions in the simulation study, because these results are highly similar to the results reported in Table 5. First, the results for the bootstrapped sample are equally unbiased. Likewise, the bias scores for the univariate statistical properties in the synthetic data in scenario two (i.e. the semicontinuous data)

are also highly similar. One exception to this general trend is the variable insulin in the semicontinuous data, where the output indicates a slight bias for the standard deviation (-0.16) and median (-2.59) in the synthetic data. As the magnitude of the bias in the univariate statistical properties is negligible for all other variables in all conditions, these results are not further discussed for reasons of brevity.

Table 5: Univariate statistical properties of the true imputed PID data and the bias of the synthetic data generated with CART for the complete PID with cp set to 1e-4

	term	Mean	SD	Median	Skew	Kurtosis	SE
True data	Pregnancies	3.85	3.37	3	0.90	0.14	0.12
	Glucose	121.72	30.58	117	0.53	-0.30	1.10
	BloodPressure	72.24	12.30	72	0.14	0.89	0.44
	SkinThickness	29.17	10.67	29	0.83	3.77	0.39
	Insulin	146.67	109.82	117.5	2.05	5.76	3.96
	BMI	32.42	6.89	32.3	0.60	0.87	0.25
	DiabetesPedigreeFunction	0.47	0.33	0.37	1.91	5.53	0.01
	Age	33.24	11.76	29	1.13	0.62	0.42
	Outcome	1.35	0.48	1	0.63	-1.60	0.02
Bias in synthetic data	Pregnancies	0.00	-0.00	0.00	-0.00	-0.01	-0.00
	Glucose	0.00	-0.00	0.09	-0.00	-0.00	-0.00
	BloodPressure	-0.01	-0.00	0.01	-0.00	-0.02	-0.00
	SkinThickness	0.00	-0.00	0.25	-0.03	-0.27	-0.00
	Insulin	0.01	-0.08	-0.04	-0.02	-0.15	-0.00
	BMI	-0.00	-0.01	-0.08	-0.01	-0.05	-0.00
	DiabetesPedigreeFunction	0.00	-0.00	0.00	-0.02	-0.16	-0.00
	Age	-0.00	-0.00	0.05	-0.00	-0.01	-0.00
	Outcome	-0.00	-0.00	0.00	0.00	0.00	-0.00

3.3 Distinguishability synthetic and observed data

Furthermore, we define the quality of the synthetic data by whether it is indistinguishable from the observed data. Ideally, people will be unable to discriminate the synthetic values from the observed values. For both simulation scenarios and under the different conditions, the prediction model yields an accuracy of 0.50, which suggest that the model's performs no better than random chance in distinguishing true data from synthetic data. These results are displayed in Table 6. In addition, the model yields a Cohen's kappa (i.e. a measure of inter-rater reliability) of 0.00, which implies consistency in the prediction performance, so the model consistently performs no better than random chance. All of this would suggest that we cannot discriminate between the synthetic and observed data.

Table 6: Prediction performance of a logistic regression model on whether rows in the merged data set are synthetic or true, for both the imputed data and semicontinuous data scenarios

	Accuracy	Kappa	Sensitivity	Specificity	Prevalence	Balanced Accuracy
Imputed	0.50	0.00	0.72	0.28	0.50	0.50
Semicontinuous	0.50	0.00	0.73	0.27	0.50	0.50

In sum, CART in MICE seems to perform well with regard to the generation of partially synthetic data. CART appears to be an appropriate imputation method that yields unbiased and confidence valid synthetic estimates given suitable pooling rules. While the semicontinuous data has a more challenging point mass, the synthetic estimates for the bootstrapped sample are unbiased in both scenarios.

Besides, the results suggest that, in this simulation, the degree of bias does not depend on the level of complexity specified for CART. Furthermore, the tables seem to confirm that the differences in synthetic estimates for the number of iterations set to default 5 or 1 are neglectable, suggesting that algorithmic convergence might be reached with as little as one iteration. Moreover, the synthetic and observed data appear to be indistinguishable. The results seem promising, but let us further discuss the strengths, limitations and implications of this simulation study.

4 Discussion

In the interest of protecting confidentiality, Rubin (1993) and Little (1993) envisioned the use of multiple imputation to release synthetic data, and initial results seemed promising. However, one challenged remained; the process of creating high-quality synthetic data was laborious and complex. The aim of this paper was to present a proof-of-concept for generating high-quality partially synthetic data with multiple imputations in MICE. In this simulation study, we conducted a Monte Carlo experiment to generate synthetic data, by *overimputing* observed values with multiple draws from the posterior predictive distribution.

The results of the simulation study show us that, as an imputation method in MICE, CART is able to generate synthetic data that yields unbiased and confidence valid estimates, given suitable pooling rules. We conclude that the CART-generated synthetic data preserves (1) the univariate statistical properties and (2) the multivariate statistical inferences of the observed data. Moreover, we cannot discriminate between the synthetic data and the observed data with a prediction model.

This brings us to the limitations of this study. The quality of the synthetic data is contingent on the quality of the imputation method, where an inappropriate method for one variable might result in bias over the whole imputation process (Drechsler, 2011). In this simulation, the synthetic data generated with PMM yielded highly biased estimates. PMM as an imputation method, might be less well suited to preserve multivariate relations and capture higher-order interaction effects in the data (Doove et al., 2014). For the average

user, we advise against the use of PMM in generating synthetic data, especially when the intention is to preserve relationships among multiple variables.

Our results suggest that CART as an imputation method can be used by data analysts and applied researchers to generate partially synthetic data. CART appears to performs well consistently; for univariate and multivariate estimates, when interaction effects are present in the data, and when the observed data are semicontinuous. This recommendation is further supported by earlier findings on the performance of CART (Burgette and Reiter, 2010; Doove et al., 2014; Reiter, 2005c). Nevertheless, it should be noted that the parameters specified for the use of CART may affect the quality of the synthetic data, we recommend that data analysts test different values of the complexity parameter and of minbucket, i.e. the minimum number of observations in any terminal node.

Even though we have demonstrated the appropriateness of generating synthetic data with CART in MICE in this study, results may not extrapolate to all other scenarios. As a suggestion for future research, we propose an extension of this simulation by systematically varying the structure of the data and considering alternative imputation methods such as GUIDE (Loh, 2002) and STIMA (Dusseldorp, Conversano, and Van Os, 2010).

In addition, more research is needed with regard to the pooling rules and their estimation of variance for partially synthetic data in practical applications. Technically the partially synthetic data pooling rules developed by Reiter (2003) are the appropriate rules for combining the m synthetic estimates and estimating the associated variances. In this simulation, however, these pooling rules resulted in nominal coverage lower than 95% for unbiased synthetic estimates. Instead, synthetic estimates were confidence valid when pooled with Rubin's rules (1987), while theoretically these rules are inappropriate in this context. Evidently, the uncertainty surrounding the pooling rules is a limitation to this research.

Lastly, this simulation study does not explicitly evaluate the disclosure risks or the implications of the synthetic data for protecting confidentiality in case of public release. Nor does the study compare synthetic data with alternative statistical disclosure techniques. For further research, imputers might use the approach by Reiter and Mitra (2009) to compute the probabilities of identification.

In conclusion, the possibilities of synthetic data outweigh the limitations. There are several instance where high-quality synthetic data may be of practical interest, in domains including health sciences, social sciences, behavioral science, as well as any other field where confidentiality concerns present a challenge. The simulation study shows MICE, an easy-to-use software and popular approach for implementing multiple imputation, can be used to generate high-quality partially synthetic data. There are still some efforts necessary, foremost, the appropriate pooling rules require more research before they might be implemented in MICE. However, the framework to generate partially synthetic data using multiple imputation is available. The performance of MICE is promising, especially with regard to the practical implications of easily accessible synthetic data on privacy-preservation in data ownership and sharing.

References

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. https://doi.org/10.1093/aje/kwq260

Caiola, G., & Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, *3*(1), 27–42. https://dl.acm.org/doi/abs/10.5555/1747335.1747337

Donaldson, M. S., & Lohr, K. N. (1994). Confidentiality and privacy of personal data. In M. S. Donaldson & K. N. Lohr (Eds.), *Health data in the information age: Use, disclosure, and privacy* (pp. 136–226). National Academies Press.

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104. https://doi.org/10.1016/j.csda.2013.10.025

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-0326-5

Dusseldorp, E., Conversano, C., & Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics*, *19*(3), 514–530. https://doi.org/10.1198/jcgs.2010.06089

Fienberg, S. E., & Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, *14*(4), 485–502.

Gewin, V. (2016). Data sharing: An open mind on open data. *Nature*, *529*(7584), 117–119. https://doi.org/10.1038/nj7584-117a

Hayashi, Y., & Yukita, S. (2016). Rule extraction using recursive-rule extraction algorithm with j48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the pima indian dataset. *Informatics in Medicine Unlocked*, *2*, 92–104. https://doi.org/10.1016/j.imu.2016.02.001

Heeringa, S, Little, R. J. A., & Raghunathan, T. E. (2002). Multivariate imputation of coarsened survey data on household wealth. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. Little (Eds.), *Survey nonresponse* (357–371). Wiley.

Kantarcolu, M., & Vaidya, J. (2018). Secure multiparty computation methods. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 3352–3358). Springer. https://doi.org/10.1007/978-1-4614-8265-9_1388

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296. http://dx.doi.org/10.2307/1391878

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*(2), 407–426.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica sinica*, *12*(2), 361–386.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, *33*(2), 142–159. https://doi.org/10.1214/18-STS644

National Institute of Diabetes and Digestive and Kidney Diseases. (2016). Pima indians diabetes database: Predict the onset of diabetes based on diagnostic measure [Dataset, Accessed: 2021-04-28]. https://www.kaggle.com/uciml/pima-indians-diabetes-database

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625. https://doi.org/10.1007/978-1-4612-4380-9_12

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*(1), 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, *18*(4), 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, *29*(2), 181–188.

Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society: Series A*, *168*(1), 185–205. https://doi.org/10.1111/j.1467-985X.2004.00343.x

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, *131*(2), 365–377. https://doi.org/10.1016/j.jspi.2004.02.003

Reiter, J. P. (2005c). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.

Reiter, J. P., & Drechsler, J. (2010). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, *20*(1), 405–421. http://www.jstor.org/stable/24308998

Reiter, J. P., & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, *1*(1), 99–110. https://doi.org/10.29012/jpc.v1i1.567

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, *102*(480), 1462–1471. https://doi.org/10.1198/016214507000000932

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Utrecht University. (2020). Open science. https://www.uu.nl/sites/default/files/ubd_strategic_plan_utrecht_university_2016-2020.pdf

Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (Second Edition). CRC press: Taylor & Francis Group. https://doi.org/10.1201/9780429492259

Van Buuren, S. (2021). Package 'mice'. R package version 3.13.0. https://cran.r-project.org/web/packages/mice/mice.pdf

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–68. https://doi.org/10.18637/jss.v045.i03

Vink, G., Frank, L. E., Pannekoek, J., & Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61–90. https://doi.org/10.1111/stan.12023

Vink, G., & Van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv preprint arXiv:1409.8542*. http://arxiv.org/abs/1409.8542

Volker, T. B. (2021). Synthetic data with mice [source code]. https://github.com/amices/Federated_imputation/blob/master/mice_synthesizing/simulations/functions.R

Yin, Y., Kaku, I., Tang, J., & Zhu, J. (2011). Privacy-preserving data mining. In Y. Yin, I. Kaku, J. Tang, & J. Zhu (Eds.), *Data mining* (pp. 101–119). Springer. https://doi.org/10.1007/978-1-84996-338-1_6

Yu, L.-M., Burton, A., & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, *16*(3), 243–258. https://doi.org/10.1177/0962280206074464