Article

Anony*mice*d shareable data: Using *mice* to create and analyze multiply imputed synthetic data sets

Thom Volker[1,*], Gerko Vink[1,†]

[1] Utrecht University - Department of Methodology and Statistics Padualaan 14, 3584CH Utrecht, the Netherlands;

* Correspondence: t.b.volker@uu.nl.

† These authors contributed equally to this work.

Version October 4, 2021 submitted to Psych

Simple Summary: A Simple summary goes here.

Abstract: A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: 1) Background: Place the question addressed in a broad context and highlight the purpose of the study; 2) Methods: Describe briefly the main methods or treatments applied; 3) Results: Summarize the article's main findings; and 4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

1. Introduction

Open science, including open data, has been marked as the future of science [1], and the advantages of publicly available research data are numerous [2,3]. Collecting research data requires an enormous investment both in terms of time and monetary resources. Openly accessible research data bears the potential of increasing the scientific returns for the same data collection effort. Additionally, the fact that public funds are used for data collection results in increasing demand for the collected data. Nevertheless, the possibilities to distribute research data directly are often very limited due to restrictions on data privacy and data confidentiality. Although these regulations are much needed, privacy constraints are also ranked among the toughest challenges to overcome in the advancement of modern day social science research [4].

Anonymizing research data might seem a quick and appealing approach to limit the unique identification of participants. However, this approach is not sufficient to fulfil contemporary privacy and confidentiality requirements [5,6]. Over the years, several other techniques have been used to increase the confidentiality of research data, such as categorizing continuous variables, top coding values above an upper bound or adding random noise to the observed values [7]. However, these methods may distort the true data relation between variables, thereby reducing the data quality and the scientific returns for re-using the same data for further research.

An alternative solution has been proposed separately by Rubin [8] and Little [9]. Although their approaches differ to some extent, the overarching procedure is to use bonafide observed data to generate multiply imputed synthetic data sets that can be freely disclosed. While in practice, one could see this as replacing the observed data values by multiple draws from the posterior predictive distribution of the observed data, based on some imputation model, Rubin would argue that these synthetic data values are merely draws from the same true data generating model. In that sense, the observed data is never replaced, but the population is resampled from the information captured in the (incomplete) sample. Using this approach, the researcher could replace the observed data set as a whole with multiple synthetic versions. Alternatively, the researcher could opt to only replace a subset of the observed data. For example, one can choose to only replace dimensions in the data that could be compared with publicly available data sets or registers. Likewise, synthetisation could be limited to those values that are disclosive, such as high incomes or high turnovers.

Conceptually, the synthetic data framework is based upon the building blocks of multiple imputation of missing data, as proposed by Rubin [10]. Instead of replacing just the missing values with multiple draws from the posterior predictive distribution, one could easily *overimpute* any observed sensitive values. Similarly to multiple imputation of missing data, the multiple synthetic data sets allow for correct statistical inferences, despite the fact that the analyses do not use the "true" value. The analyses over multiple synthetic data sets should be pooled into a single inference, so that the researcher can draw valid conclusions from the pooled results. To that respect, the variance should reflect the added variability that is induced by the imputation procedure.

Potentially, this approach could fulfill the needs for openly accessibly data, without running into barriers with regard to privacy and confidentiality constraints. However, there is no such thing as a free lunch: data collectors have to put effort in creating high-quality synthetic data. Also, the quality of the synthetic data is highly dependent on the imputation models, and using flawed models to generate synthetic data might bias subsequent analyses. Conversely, if the models used to create the synthetic data are able to preserve the relationships between the variables as in the original data, the synthetic data can be nearly as informative as the observed data. Thus, to fully exploit the benefits of synthetic data, the effort to actually create these high-quality data sets should be kept at a minimum.

To mitigate additional effort of creating synthetic data sets on behalf of the researcher, software aimed at multiple imputation of missing data can be employed. Especially if researchers used this software at an earlier stage in the research process, or acquired familiarity with it during earlier projects, the additional burden of creating synthetic data sets is relatively small. The R-package `mice` [11] implements multiple imputation of missing data in a straightforward and user-friendly manner. However, the functionality of `mice` is not restricted to the imputation of missing data, but allows to impute any value in the data: even observed values. Consequently, `mice` can be utilized for the creation of multiply imputed synthetic data sets.

After creating the multiply imputed synthetic data sets, the goal is to obtain valid statistical inferences in the spirit of Rubin [10] and Neyman [12]. In the missing data framework, this is done by performing statistical analyses on all imputed data sets, and pooling the results of the analyses according to Rubin's rules [10, pp.76]. In the synthetic data framework, the same procedure is followed, but with a slight twist: there are no values that remain constant over the synthetic data sets. The procedure of drawing valid inferences from multiple synthetic data sets is therefore slightly different.

In this manuscript we detail a workflow for synthesizing data with `mice`. First, the `mice` algorithm for the creation of synthetic data will be shortly explained. The aim is to generate synthetic sets that reassure the privacy and confidentiality of the participants. Second, a straightforward workflow for imputation of synthetic data with `mice` will be demonstrated. Third, we demonstrate the validity of the procedure through statistical simulation.

2. Generating synthetic data with `mice`

The `mice` package [11] in `R` [13] has been developed for multiple imputation of missing data. In this context, the aim is to replace missing values due to nonresponse by plausible values from the posterior predictive distribution of the variable containing the missings. Doing so, `mice` makes use of fully conditional specification [FCS; Van Buuren *et al.* [14]], which breaks down the multivariate distribution of the data $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ into $j = 1, 2, \ldots, k$ univariate conditional densities, where k denotes the number of columns in the data. Using FCS, a model is constructed for every incomplete variable and the missing values $Y_{j,mis}$ are then imputed with draws from the posterior predictive distribution of $P(Y_{j,mis}|\mathbf{Y}_{obs}, \theta)$ on a variable-by-variable basis. Note that the predictor matrix Y_{-j} may contain yet imputed values from an earlier imputation step, and thus will be updated after every iteration. This procedure is applied m times, resulting in m completed data sets $\mathbf{D} = (\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \ldots, \mathbf{D}^{(m)})$, with $\mathbf{D}^{(l)} = (\mathbf{Y}_{obs}, Y^{(l)}_{mis})$.

In `mice`, the generation of multiply imputed data sets to solve for unobserved values is straightforward. The following pseudocode details the multiple imputation of the `mice::boys` data set [15] into the object `imp` with `m = 10` imputated sets and `maxit = 7` iterations for the algorithm to converge, using the default imputations methods for each column data class.

```
library(mice)
imp <- mice(boys,
            m = 10,
            maxit = 7)
```

It is straightforward to extended the imputation approach to generate synthetic values. Rather than imputing missing data, the observed values are then replaced by synthetic draws from the posterior predictive distribution. For simplicity, assume that the data is completely observed (i.e., $\mathbf{Y} = \mathbf{Y}_{obs}$). Following the notation of Reiter and Raghunathan [16], let for n units denote $Z_i = 1$ if any of the values of unit $i = 1, 2, \ldots, n$ are to be replaced by imputations, and $Z_i = 0$ otherwise, with $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$. Accordingly, the data consists of values that are to be replaced and values that are to be kept (i.e., $\mathbf{Y} = (\mathbf{Y}_{rep}, \mathbf{Y}_{nrep})$. Now, instead of imputing $\mathbf{Y}_{mis}$ with draws from the posterior predictive distribution of $P(Y_{j,mis}|\mathbf{Y}_{obs}, \theta)$ as in the missing data case, $\mathbf{Y}_{rep}$ is imputed from the posterior distribution of $P(Y^{(l)}_{j,rep}|\mathbf{Y}^{(l)}_{-j}, \mathbf{Z}, \theta)$, where l is an indicator for the synthetic data set ($l = 1, 2, \ldots, m$). Note that synthetic values that are imputed at an earlier step can be used for imputing variable j. This process results in the synthetic data $\mathbf{D} = (\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \ldots, \mathbf{D}^{(m)})$.

For example, overimputing synthetic values for both the observed and missing cells in the `mice::boys` data set into the object `syn`, given the same imputation parameters as before, can be realized by the following code execution.

```
syn <- mice(boys,
            m = 10,
            maxit = 7,
            where = matrix(TRUE,
                           nrow = nrow(boys),
                           ncol = ncol(boys)))
```

where the argument `where` requires a matrix of the same dimensions as the data, (i.e., a $n \times k$ matrix) containing logicals z_{ij} that indicate which cells are selected to have their values replaced by draws from the posterior predictive distribution. This approach allows to *overimpute* a subset of the observed data, or - as in the above example - the observed data as a whole, resulting in a data set that partly or completely consists of synthetic data values.

Choosing an adequate imputation model to impute the data is paramount, as a flawed imputation model may drastically impact the validity of inferences. Imputation models should be as flexible as possible to capture most of the patterns in the data, and to model possibly unanticipated data characteristics [17,18]. Parametric methods, albeit easy to implement in practice, may be too restrictive to capture generally complex patterns in the data, especially in the case of nonlinear relations and interactions between multiple variables. Classification and regression trees [CART; Breiman *et al.* [19]] allow to model more complex patterns in the data, and have therefore been suggested as an appropriate imputation method [20–22]. Loosely speaking, CART sequentially splits the predictor space into non-overlapping regions in such a way that the within-region variance is as small as possible after every split. As such, CART does not impose any parametric distribution on the data, making it a widely applicable method that allows for a large variety of relationships within the data [23]. Given these appealing characteristics and the call for the use of flexible methods when multiply imputing data, we will focus our illustrations and evaluations of `mice` to method `mice.impute.cart()`, realized by:

```
syn <- mice(boys,
            m = 10,
            maxit = 7,
            method = "cart",
            where = matrix(TRUE,
                           nrow = nrow(boys),
                           ncol = ncol(boys)))
```

In a nutshell, the above code shows the simplicity of creating $m = 10$ synthetic data sets using `mice`. In practice, however, one should take some additional complicating factors into account. For example, one should account for deterministic relations in the data. Additionally, relations between variables may be described best using a different model than `CART`. Such factors are data dependent and should be considered by the imputer. In the next section, we will describe how the `boys` data can be adequately imputed. Additionally, we will show through simulations that this approach yields valid inferences.

3. Materials and Methods

We demonstrate the suitability of using `mice` for synthesization using a simulation study on the `mice::boys` data set. This data set consists of the values of 748 Dutch boys on the following 9 variables:

column	description
age	age in years
hgt	height (cm)
wgt	weight (kg)
bmi	body mass index
hc	head circumference (cm)
gen	genital Tanner stage G1-G5
phb	pubic hair Tanner P1-P6
tv	testicular volume (ml)
reg	region

Unfortunately, this data set does not differ from the vast majority of collected data sets, in the sense that it suffers from missing data. For simplicity, the data is completed using the default `mice` imputation model for all predictors except `bmi`, which is passively imputed using its deterministic relation with weight and height.

```
# create a single imputed, completely observed `boys` data set
set.seed(123)

meth <- make.method(boys)
meth["bmi"] <- "~ I(wgt / (hgt / 100)^2)"
pred <- make.predictorMatrix(boys)
```

```
pred[c("hgt", "wgt"), "bmi"] <- 0

imp <- mice(boys,
            m = 1,
            maxit = 10,
            method = meth,
            predictorMatrix = pred)

data <- complete(imp)
```

3.1. Simulation methods

To induce sampling variance, 1000 bootstrap samples of the `boys` data have been synthesized with $m = 5$ imputations for every data cell. Synthetic values are generated using the `CART` imputation method for all columns, except for `bmi`. The deterministic relation `bmi` which will be synthesized passively based on the synthetic values for `hgt` and `wgt` to preserve the relation in the synthetic data. Additional parameters that come with the use of `mice.impute.cart()` are the complexity parameter `cp` and the minimum number of observations in any terminal node `minbucket`, that both constrain the flexibility of the imputation model. The values of the parameters `cp` and `minbucket` ought to adhere to the call for imputation models that are as flexible as possible. Appropriate values for these parameters, as well as the input for the `predictorMatrix`, depend on the data at hand. In the current example, the complexity parameter is specified at `cp = 1e-08` rather than the default value `1e-04`, and the minimum number of observations in each terminal node is set at `minbucket` $= 3$ rather than the default value 5. By allowing for more complexity in the imputation model, bias in the estimates from the synthetic data set is reduced. Additionally, since the missingness pattern is monotone, the number of iterations can be set to `maxit = 1`.

To assess the performance of `mice` for synthesizing data, we compare the bootstrapped samples with the synthetic versions of these bootstrapped samples. Specifically, univariate descriptive statistics, the correlation matrix, and two linear regression models as well as one ordered logistic regression model will be considered. Subsequently, the bias in the parameters and the 95% confidence interval coverage of the synthetic data will be examined. Similarly to multiple imputation of missing data, correct inferences from synthetic data requires correct pooling over the multiply imputed data sets.

Obtaining a final point estimate of the parameter of interest Q after imputation is fairly easy and no different from pooling in the case of missing data [10]. One can calculate the average of the m point estimates $q^{(l)}$

$$\bar{q}_m = \sum_{l=1}^{m} \frac{q^{(l)}}{m}.$$

with $l = 1, \dots, m$.

Similarly to the missing data case, variances, and subsequently confidence intervals, should incorporate the increase in variance that is due to imputation [7,24]. Yet, the increase in variance due to imputation differs according to whether missing values are imputed or observed data is overimputed with missing values. Whereas the variance estimate after imputation of missing data needs to account for the fact that a certain amount of information in the data is missing, variance estimation from synthetic data does not suffer from this issue. The adjusted variance estimate that follows from using multiple synthetic data sets only suffers from the fact that a finite number of m synthetic data sets are used to resemble the observed data. Hence, the according variance estimate for synthetic data as developed by Reiter [24] yields

$$T = \bar{u}_m + \frac{b_m}{m},$$

with between-imputation variance

$$b_m = \sum_{l=1}^{m} \frac{(q^{(l)} - \bar{q}_m)^2}{(m-1)},$$

and sampling variance

$$\bar{u}_m = \sum_{l=1} \frac{u^{(l)}}{m},$$

where $u^{(l)}$ denotes the variance estimate in the lth synthetic data set.

4. Results

```
plan(multisession) # increase speed through futures

true_model_age <- lm(age ~ wgt + hgt, data) # model 1
true_model_wgt <- lm(wgt ~ age + hgt, data) # model 2
true_model_gen <- MASS::polr(gen ~ age + hc + reg, data, Hess = TRUE) # model 3

coefs_age <- broom::tidy(true_model_age)$estimate # extract coefficients of model 1
coefs_wgt <- broom::tidy(true_model_wgt)$estimate # extract coefficients of model 2
coefs_gen <- broom::tidy(true_model_gen)$estimate # extract coefficients of model 3

nsim <- 1000 # use 1000 iterations

bootstrap_samples <-
  modelr::bootstrap(data = data, n = nsim) %$%
  strap %>%
  map(as_tibble)

post <- make.post(data)
post["bmi"] <- "imp[[j]][, i] <- imp[['wgt']][, i] / (imp[['hgt']][, i] / 100)^2"

synthetic_samples <-
  bootstrap_samples %>%
  future_map(function(x) {
    x %>% mice(m = 5,
               maxit = 1,
               method = "cart",
               minbucket = 3,
               cp = 1e-08,
               predictorMatrix = pred,
               post = post,
               where = matrix(TRUE, nrow(data), ncol(data)),
               print = F)
}, .options = furrr_options(seed = as.integer(123)))
```

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

4.1. Subsection Heading Here

Subsection text here.

4.1.1. Subsubsection Heading Here

Bulleted lists look like this:

- First bullet
- Second bullet
- Third bullet

Numbered lists can be added as follows:

1. First item
2. Second item
3. Third item

The text continues here.

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

Figure 1. This is a figure, Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Table 2. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	**Title 2**	**Title 3**
entry 1	data	data
entry 2	data	data

This is an example of an equation:

$$\mathbb{S} \tag{1}$$

Example of a theorem:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:

Example of a proof:

Proof of Theorem 1. Text of the proof. Note that the phrase 'of Theorem 1' is optional if it is clear which theorem is being referred to. □

The text continues here.

5. Discussion

Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

6. Conclusion

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

7. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Acknowledgments: All sources of funding of the study should be disclosed. Please clearly indicate grants that you have received in support of your research work. Clearly state if you received funds for covering the costs to publish in open access.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “X.X. and Y.Y. conceive and designed the experiments; X.X. performed the experiments; X.X. and Y.Y. analyzed the data; W.W. contributed reagents/materials/analysis tools; Y.Y. wrote the paper.’’ Authorship must be limited to those who have contributed substantially to the work reported.

Conflicts of Interest: ‘The authors declare no conflict of interest.’

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text. For example, explanations of experimental details that would disrupt the flow of the main text, but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with ‘A’, e.g., Figure A1, Figure A2, etc.

References

1. Gewin, V. Data sharing: An open mind on open data. *Nature* **2016**, *529*, 117–119. Number: 7584 Publisher: Nature Publishing Group, doi:10.1038/nj7584-117a.
2. Molloy, J.C. The open knowledge foundation: open data means better science. *PLoS Biol* **2011**, *9*, e1001195.
3. Walport, M.; Brest, P. Sharing research data to improve public health. *The Lancet* **2011**, *377*, 537–539.

4. Lazer, D.; Pentland, A.S.; Adamic, L.; Aral, S.; Barabasi, A.L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; Van Alstyne, M. Life in the network: the coming age of computational social science. *Science* **2009**, *323*, 721.
5. Ohm, P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* **2009**, *57*, 1701–1778.
6. Council, N.R. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*; The National Academies Press: Washington, DC, 2007. doi:10.17226/11865.
7. Drechsler, J. *Synthetic datasets for statistical disclosure control: theory and implementation*; Vol. 201, Springer Science & Business Media, 2011.
8. Rubin, D.B. Statistical disclosure limitation. *Journal of official Statistics* **1993**, *9*, 461–468.
9. Little, R.J. Statistical analysis of masked data. *Journal of Official statistics* **1993**, *9*, 407–426.
10. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley, 1987.
11. Van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **2011**, *45*, 1–67.
12. Neyman, J. 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* **1934**, *97*, 123–150.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.
14. Van Buuren, S.; Brand, J.P.L.; Groothuis-Oudshoorn, C.G.M.; Rubin, D.B. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation* **2006**, *76*, 1049–1064.
15. Fredriks, A.M.; Van Buuren, S.; Burgmeijer, R.J.; Meulmeester, J.F.; Beuker, R.J.; Brugman, E.; Roede, M.J.; Verloove-Vanhorick, S.P.; Wit, J.M. Continuing positive secular growth change in The Netherlands 1955–1997. *Pediatric research* **2000**, *47*, 316–323.
16. Reiter, J.P.; Raghunathan, T.E. The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association* **2007**, *102*, 1462–1471.
17. Murray, J.S. Multiple imputation: a review of practical and theoretical findings. *Statistical Science* **2018**, *33*, 142–159.
18. Rubin, D.B. Multiple imputation after 18+ years. *Journal of the American statistical Association* **1996**, *91*, 473–489.
19. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and regression trees*; CRC press, 1984.
20. Reiter, J.P. Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics* **2005**, *21*, 441.
21. Burgette, L.F.; Reiter, J.P. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology* **2010**, *172*, 1070–1076.
22. Doove, L.L.; Van Buuren, S.; Dusseldorp, E. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational statistics & data analysis* **2014**, *72*, 92–104.
23. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Vol. 112, Springer, 2013.
24. Reiter, J.P. Inference for partially synthetic, public use microdata sets. *Survey Methodology* **2003**, *29*, 181–188.

Sample Availability: Samples of the compounds ...... are available from the authors.