Outline paper - Multiple imputation for statistical disclosure control: Creating synthetic datasets with *mice*

Anony*mice*d shareable data: Using *mice* to create multiply imputed synthetic datasets

Thom Volker, Gerko Vink & Stef van Buuren

21-08-2020

Table of Contents

# OUTLINE

**Intro**

* Motivation/relevance
* Synthetic data (fully/partially shortly explained)
* How can this data be generated in practice (FCS - multivariate synthetic values)
* Mice adopts FCS for missing data
* It will be shown that mice is also capable of generating multivariate synthetic data
* However, multiple rules for correct variance estimates have been proposed, but the choice of one optimal variance estimate is not straightforward. Therefore, the performance of the currently available variance estimators (as outlined below) will be contrasted
* Goal: explain and show how mice can be used to create synthetic versions of the data at hand, and which variance estimator yields the best results.

**Synthetic data explained**

*Fully synthetic data*

* Explain the idea in more depth than in the intro
* Explain pooling rules by Raghunathan, Reiter & Rubin
* Adjusted variance estimate by Reiter (restricted positive)
* Simple variance estimator by synthpop authors.

*Partially synthetic data*

* Explain the idea in more depth than in the intro
* Explain pooling rules by Reiter.
* Adjusted estimate by synthpop authors.

**Mice**

* Explain algorithm (FCS) in more depth
* Observed values can be regarded as starting values for the imputation model. Since the observed data is used for creating the imputation models, no iterations are required. That is, the model is fixed, and not dependent on the imputed values.

**Methods**

* Boys dataset, considering using another dataset (two datasets with same results would prove the point even more).
* Explain synthesizing procedure (by means of cart, using fcs as implemented in mice).
* Imputation settings (CART, as proposed by Reiter).
* What will be looked at: simple means / variances of variables (univariate relationships) and estimates of regression coefficients (of multiple models (at least 2 different models)).
* Eventually: distinction between fully (everything replaced) and partially (only subset of variables replaced) synthetic data.
* Eventually: distinction between number of imputations (however, proved to yield sufficient results already in the simulations run thus far).

**Results**

* Discuss results.

**Discussion**

* Pros and cons of different methods, theoretically and practically.

# Introduction

Open science, including open data, has been marked as the future of science (Gewin, 2016), and the advantages of publicly available research data are numerous (Foster & Deardorff, 2017; Walport & Brest, 2011). HIER EEN VOORBEELD Additionally, the fact that public funds are used for data collection results in increased demands for the collected data. Nevertheless, the possibilities to share research data directly are often very limited due to restrictions with regard to privacy and confidentiality. Simply anonimizing the data is not enough to fulfil these requirements (Ohm, 2009). Over the years, several other techniques have been used to increase the confidentiality of the data, such as categorizing continuous variables, top coding values above a certain threshold or adding random noise (Drechsler, 2011). However, these methods may distort relationships between variables, reducing the data quality drastically.

An alternative has been offered by Rubin (1993), who, building on the framework of multiple imputation, proposed to release multiple synthetic datasets to the public. Conform this approach, all units that are in the population but not in the sample are treated as missing data. These values are imputed by means of conventional multiple imputation approaches, and simple random samples are drawn from the population. Ultimately, these samples are released to the public. In practice, it is not required to impute the complete population, since random samples can be drawn from the sampling frame, so that only these sampled values have to be imputed. The datasets generated under this approach are labelled as *fully synthetic datasets*. However, note that, if one draws simple random samples from the sampling frame, the possibility exists that actual observations are included.

A second procedure to creating synthetic datasets has been proposed by Little (1993), named *partially synthetic datasets*, which originated as a procedure that required synthesizing only those values that are at a high risk of being disclosive. These values can be complete variables that could be compared with publicly available datasets or registers, such as addresses, or it can be certain values that bear a high risk of being disclosive, say, income values above a certain threshold. Similarly to the fully synthetic data approach, multiple datasets containing synthetic values are released to the public, although in this instance, only the values at a high risk vary over the synthetic datasets. However, nothing keeps us from treating all values in the dataset as bearing a high disclosure risk, resulting in datasets that are completely existing of synthetic values. This approach basically implements the idea of fully synthetic data, based on the partially synthetic data procedure.

– – HIER MIST NOG EEN OVERGANG WAAR IK NOG NIET UIT BEN – –

As of today, most papers in this field address methodological issues concerning both fully and partially synthetic datasets, both with regard to analysis methods suitable for creating synthetic data, and concerning inferential procedures. The current paper aims at reducing the gap between the theory about synthetic data, and generating synthetic data in practice. Due to the conceptual similarities between imputation for missing data and imputation for synthetic data, we propose that the R-package mice is capable of bridging the gap between theory and actual applications. Originally, mice is developed as an R-package to impute values subject to non-response. However, algorithmically, the procedure of imputing values subject to non-response does not need to differ from imputing synthetic data. Both approaches can be based on fully conditional specification (FCS) as implemented in mice. Conform this approach, values are drawn iteratively from the posterior predictive distribution of the variable of interest conditional on all other variables, for each variable separately. Thus, previously imputed values are taken into account in the imputation model for variables that are imputed later on.

Since mice is currently only capable of imputing the data at hand, the approach should be labelled as *partially synthetic*. That is, although it is completely possible to create a synthetic dataset that consists of only synthetic values, mice is not (yet) capable of imputing observations that are not present in the data. Creating fully synthetic datasets by means of mice remains an area of future research.

– – NOT SURE YET: CONTRAST FULLY AND PARTIALLY SYNTHETIC DATA – –

– – SINCE PARTIALLY SYNTHETIC DATA POOLING RULES DO NOT REQUIRE ADJUSTED VARIANCE ESTIMATES – –

– – CAN WE REGARD THE OBSERVED VALUES AS THE STARTING VALUES FOR THE IMPUTATIONS ? – –

# FULLY/PARTIALLY synthetic data

## Fully synthetic data

DISCUSS APPROACH IN MORE DEPTH

Explain pooling rules (variance estimates + adjusted variance estimates + simple estimator proposed by synthpop authors)

## Partially synthetic data

DISCUSS APPROACH IN MORE DEPTH

Explain pooling rules (variance estimates)

# Mice

* Discuss the FCS algorithm in more detail.
* Discuss mice, predominantly the imputation algorithm in greater depth
* Procedure: observed values can be regarded as starting values, so that all variables can be used for the synthesis of all other variables.

# Simulations

## Methods

* Boys data - bootstrapped to obtain a population.
* explain simple synthetic data procedure
* explain imputation settings
* CART has been proposed by Reiter

## Results

## References

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: Theory and implementation* (Vol. 201). Springer Science & Business Media.

Foster, E. D., & Deardorff, A. (2017). Open science framework (osf). *Journal of the Medical Library Association: JMLA*, *105*(2), 203.

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

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

Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*(6), 1701–1778. Retrieved from <https://heinonline.org/HOL/P?h=hein.journals/uclalr57&i=1713>

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

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *The Lancet*, *377*(9765), 537–539.