Using Synthetic Data to Improve the Reproducibility of Statistical Results in Psychological Research

Simon Grund [1,2], Oliver Lüdtke [1,2], and Alexander Robitzsch [1,2]

[1] Leibniz Institute for Science and Mathematics Education

[2] Centre for International Student Assessment

Author Note

Simon Grund 0000-0002-1290-8986

Oliver Lüdtke 0000-0001-9744-3059

Alexander Robitzsch 0000-0002-8226-3132

All additional files concerning this article, including scripts, data, and the supplemental materials are available at: https://osf.io/3a5uq/

Correspondence concerning this article should be addressed to Simon Grund, Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany; (+49)-431-880-5653; grund@leibniz-ipn.de

Abstract

In recent years, psychological research has faced a credibility crisis, and open data are often regarded as an important step toward a more reproducible psychological science. However, privacy concerns are among the main reasons that prevent data sharing. Synthetic data procedures, which are based on the multiple imputation (MI) approach to missing data, can be used to replace sensitive data with simulated values, which can be analyzed in place of the original data. One crucial requirement of this approach is that the synthesis model is correctly specified. In this article, we investigated the statistical properties of synthetic data with a particular emphasis on the reproducibility of statistical results. To this end, we compared conventional approaches to synthetic data based on MI with a data-augmented approach (DA-MI) that attempts to combine the advantages of masking methods and synthetic data, thus making the procedure more robust to misspecification. In multiple simulation studies, we found that the good properties of the MI approach strongly depend on the correct specification of the synthesis model, whereas the DA-MI approach can provide useful results even under various types of misspecification. This suggests that the DA-MI approach to synthetic data can provide an important tool that can be used to facilitate data sharing and improve reproducibility in psychological research. In a worked example, we also demonstrate the implementation of these approaches in widely available software, and we provide recommendations for practice.

Keywords: reproducibility, robustness, synthetic data, open science, data sharing

Using Synthetic Data to Improve the Reproducibility of Statistical Results in Psychological Research

Recent years have constituted a period of change in psychological research, and the scientific community has expended significant effort to improve the credibility of its findings and overcome a credibility crisis in psychology (Baker, 2016; Nelson et al., 2018; Nosek et al., 2015; Nosek et al., 2021). There are varying definitions of what it means to conduct credible research, but this term is often used to convey the idea that research should be reproducible, robust, replicable, and generalizable (e.g., Asendorpf et al., 2013; Bollen et al., 2015; Nosek et al., 2021). In this context, *reproducibility* is sometimes used to refer to researchers' ability to obtain the same findings as the ones obtained in an original study when using the same data, procedures, and materials. By contrast, *robustness* and *replicability* sometimes refer to the ability to obtain similar findings with different procedures and data, respectively, and *generalizability* sometimes refers to the ability to obtain similar findings when both the data and procedures differ from the ones used in the original study (Nosek et al., 2021).

Many authors have stressed the importance of sharing the original data from psychological studies to improve the credibility of their findings because this allows researchers to independently reproduce the original findings and to investigate the robustness of the results to different specifications of the statistical analyses (Artner et al., 2020; Lindsay, 2017; Martone et al., 2018; Munafò et al., 2017; Perrino et al., 2013; Wicherts & Bakker, 2012). However, research has also shown that data sharing is rare. For example, Hardwicke et al. (2021) found that only four (2.1%) articles out of a sample of 188 articles published in psychological journals between 2014 and 2017 contained data availability statements, and only three (1.6%) provided access to the actual data (for similar findings, see also Towse et al., 2020; Vanpaemel et al., 2015; Wicherts et al., 2006). Although there is significant variance in researchers' attitudes toward data sharing, privacy issues—such as the need to fulfill confidentiality agreements or protect the identity of participants—are among the main barriers that prevent data from being shared more often (e.g., Houtkoop et al., 2018; Tenopir et al., 2011; Zuiderwijk et al., 2020). Public data can sometimes place the participants of a study at risk of being identified from the data, for example, if the participant possesses a unique combination of attributes (e.g., gender, ethnicity, occupation) or if a malicious third party attempts to deduce their identity from the data and other available information (e.g., location, time period, sample characteristics) about the study (e.g., Erb et al., 2021; Fleming et al., 2021; Gilmore et al., 2018; Joel et al., 2018; Meyer, 2018).

In the statistical literature, masking procedures and the generation of synthetic data have been recommended as ways to share data while protecting the identities of the participants (Little, 1993; Raghunathan et al., 2003; Rubin, 1993). Masking refers to techniques that obfuscate the original data, for example, by merging categories or adding noise to individual observations. One limitation of masking procedures is that they typically do not preserve the relationships between the observed variables, especially if strong masking is needed to protect participants' identities. Synthetic data aim to replace sensitive information through simulation procedures that are based on the multiple imputation (MI) approach (Rubin, 1987) for the treatment of missing data (for an overview, see Drechsler, 2011; for an introduction, see also Quintana, 2020). The main challenge in the application of synthetic data is that it relies on the specification of a synthesis model that must fit the intended analysis and correctly reflect the relations between the observed variables. Otherwise, the results obtained from synthetic data can be misleading. Originally developed for survey research, synthetic data are becoming more and more common in psychology and related disciplines with applications in education (Bonnéry et al., 2019; Schauer et al., 2019), business and employment research (Drechsler & Reiter, 2009; Kinney et al., 2011), medicine (Goncalves et al., 2020), and neuroscience (Yarkoni et al., 2011).

In the present article, we investigated the statistical properties of synthetic data as a tool for fostering data sharing and nurturing reproducibility and robustness in psychological research. In this context, we also compared different approaches to synthetic data, including both conventional methods that are based on MI and a novel data-augmented approach that attempts to combine the relative strengths of masking methods and the MI approach to synthetic data (Jiang et al., 2021). Our article is organized as follows. First, we present a classification of the different levels of reproducibility in psychological research. Next, we provide an overview of existing methods for confidentiality protection, including masking procedures as well as the conventional and newer approaches to synthetic data. Then, we present the results of multiple simulation studies in which we evaluated the performance of the synthetic data approaches. Finally, we demonstrate the application of these approaches in a worked example with real data, provide recommendations for practice, and close with a discussion of our findings.

Reproducibility and Robustness in Psychological Research

Reproducibility and robustness refer to the reanalysis of existing data with the same data and the same or different methods, respectively. Goodman et al. (2016) further distinguished between the reproducibility of *methods* and *results*. Similarly, robustness can refer to either *specific* changes to a study's procedures and analyses or *general* ones with a potentially large number of variations (e.g., Simonsohn et al., 2020). In the following, and as shown in Table 1, we present a classification that identifies four stages of reproducibility and robustness in psychology. The purpose of this classification is to help guide the following discussion about different approaches to synthetic data. Distinguishing between these stages is also useful because they highlight different uses of shared data, ranging from simple to complex, where each stage imposes stronger requirements on the data. For illustration purposes, suppose that a group of researchers investigated the linear relations between students' Big Five personality traits and their educational achievements after controlling for their socioeconomic status and differences in motivation (e.g., self-concept, interests).

Stage 1. In the first and most basic stage, the shared data are primarily used as a device that allows other researchers to run the computer code and verify the analytic procedures used in the original study, whereas the reproducibility of results is not a direct concern. These data would not provide interpretable results, but they would still allow other researchers to check the analyses, rule out coding errors, and so on. Naturally, this requires that the computer code or sufficient information about the analysis strategy is shared alongside the data. In the example above, this would mean that other researchers could use the shared data set to understand the statistical

Table 1

Stages of Reproducibility and Robustness in Psychological Research

Stage	Purpose	Description	Examples
1	Reproducibility (methods)	Reproduction of the methods and procedures in the original study	Code review; verification of data processing
2	Reproducibility (results)	Reproduction of the methods, procedures, and results reported in the original study	Parameter estimation; confidence intervals; statistical tests; model comparisons
3	Robustness (specific)	Additional analyses not included in but based on the original study	Covariate adjustment; changes to functional forms; alternative models; multiverse analysis
4	New analyses (general)	Additional analyses unrelated to the original study	Research synthesis; additional outcomes

analyses (e.g., multiple regression, structural equation models) and verify the computer code that was used to run the analyses.

Stage 2. In the second stage, the shared data are intended to reproduce the results reported in the original study. Specifically, the shared data are expected to provide similar parameter estimates, standard errors, confidence intervals, and so on. However, at this stage, the reproducibility of the results is still limited to the results reported in the original study, whereas additional analyses that go beyond the original study might not necessarily be supported by the data. In the example above, this would mean that the shared data would allow other researchers to reproduce the main findings and obtain estimates of the relations between personality and achievement that were either identical or very similar to those in the published study.

Stage 3. Going beyond reproducibility, public data can also be used to investigate the extent to which the findings in the original study are robust to certain changes in the analysis strategy. In the third stage, the shared data are therefore expected to provide trustworthy results for the not just the original but also alternative analysis strategies. In the example above, this would correspond to a case where the shared data would allow researchers to investigate whether there are nonlinear effects of the personality traits or whether the effects are moderated by gender. However, at this stage, the difference between the analysis strategies is still relatively specific, and the two strategies are concerned with the same research question.

Stage 4. Finally, in the fourth and most general stage, the shared data are regarded as a functional replacement of the original data. At this stage, the data are expected to provide trustworthy results for essentially arbitrary analyses or research questions, similar to those that would have been obtained in the original data. For instance, in the scenario above, this could mean that other researchers would use the shared data to investigate research questions that are unrelated to the original study, for example, in a meta-analysis on gender differences in academic interests or other variables that are contained in the same data set.

Naturally, no data can offer both a perfect protection of confidentiality and an exact reproducibility of all possible analyses, and creating confidentiality-protected data that support higher stages of reproducibility is even more difficult (e.g., Bonnéry et al., 2019). In other words, there is a fundamental trade-off between the reproducibility of statistical analyses with shared data and the ability to protect the participants' identities (Duncan et al., 2011; see also Karr et al., 2006; Snoke et al., 2018). However, different methods for confidentiality protection may offer reproducibility to different extents. This is especially important for the reproduction of results (Stage 2) and the investigation of specific and general robustness or changes to the analysis strategy (Stages 3 and 4), where researchers who plan to share data may be able to anticipate some but not all of the analyses that other researchers will apply to the data. In the following, we provide a brief review of some common methods for confidentiality protection before we focus on novel approaches to synthetic data.

Popular Methods for Confidentiality Protection

Suppose that a researcher is interested in publishing a data set in which a subset of sensitive variables $\mathbf{x} = (x_1, \ldots, x_p)$ should not be released in their original form, whereas other variables $\mathbf{z} = (z_1, \ldots, z_q)$ can be left unaltered. Two popular approaches for preventing the release of sensitive data can be found in the literature: masking and synthetic data. Masking procedures attempt to obfuscate the original data, for example, by adding noise to individual observations, whereas the name synthetic data refers to simulation procedures that are designed to replace sensitive information with simulated values through the use of MI. In the following section, we provide a brief review of these methods.

Masking Procedures

In masking approaches, noise is added (or similar modifications are applied) to the sensitive variables, and the original variables $\mathbf{x}$ are replaced by masked copies $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)$. For example, if the $\mathbf{x}$ are continuous, a common masking approach is to add random noise to the original observations of each unit i ($i = 1, \ldots, n$):

$$\mathbf{x}_i^* = \mathbf{x}_i + \mathbf{e}_i^* \ , \quad \mathbf{e}_i^* \sim N(\mathbf{0}, \mathbf{\Lambda}) \ , \tag{1}$$

where $\mathbf{\Lambda}$ is a diagonal matrix that controls the amount of noise added to each variable. This approach is simple but somewhat naïve in the sense that it ignores the relations between the variables by independently simulating random noise for each variable. A more sophisticated approach that also takes the relations between the variables into account can be implemented as follows (Fuller, 1993; Little, 1993):

$$\mathbf{x}_i^* = \frac{1}{\sqrt{1+\lambda}}(\mathbf{x}_i - \bar{\mathbf{x}} + \mathbf{e}_i^*) + \bar{\mathbf{x}} \ , \quad \mathbf{e}_i^* \sim N(\mathbf{0}, \lambda\mathbf{S}) \ , \tag{2}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean and covariance matrix of $\mathbf{x}$, respectively, and λ is a tuning parameter that controls the amount of noise added to the data. Other popular masking procedures include rounding, merging categories, aggregating, deleting, or grouping the values of sensitive variables as well as swapping rows or columns in the data (e.g., Duncan & Pearson, 1991; Little, 1993). Although masking procedures are popular and easy to implement, their main limitation is that they typically do not preserve the relations that existed between the variables and therefore require the use of special analysis methods that correct for the error introduced in each masking step (for a detailed discussion, see Little, 1993).

Synthetic Data

The second approach is based on the popular MI methodology for handling missing data (Rubin, 1987) and attempts to generate $M \geq 1$ synthetic data sets with simulated values $\mathbf{x}_{syn}^{(m)}$

($m = 1, \ldots, M$) in place of the original values in $\mathbf{x}$ (Raghunathan et al., 2003; Rubin, 1993). The basic idea of this method is to draw M replacements of the original data $\mathbf{x}$ from the predictive distribution $\mathbf{x}_{syn}^{(m)} \sim P(\mathbf{x}|\mathbf{z})$, thereby generating synthetic data on the basis of the original data and a statistical model (the synthesis model). In practice, this method requires the specification of a joint distribution, $P(\mathbf{x}, \mathbf{z})$, for the original variables, implying a predictive distribution for the sensitive variables in $\mathbf{x}$ from which the synthetic values can be generated.

One important requirement of synthetic data is that this joint distribution $P(\mathbf{x}, \mathbf{z})$ is correctly specified and congenial with the intended analyses (Meng, 1994). Otherwise, the results obtained from the synthetic data can be misleading. Broadly speaking, the synthesis model must be at least as general as the analyses that are meant to be reproduced with the data. This is particularly important for reproducibility and the investigation of robustness. For reproducibility, this means that the synthesis model must take the original analyses into account. For the investigation of robustness, it means that the additional analyses must be identified a priori and taken into account when specifying the synthesis model. For instance, in the example above, reproducibility would reqiure that the synthesis model reflects the linear relations between personality and educational achievement, and robustness analyses (e.g., on the nonlinear effects of personality) would require that these analyses are identified a priori and taken into account in the synthesis model (Seaman et al., 2012; von Hippel, 2009).

Specifying the Synthesis Model

Three strategies for specifying $P(\mathbf{x}, \mathbf{z})$ can be distinguished in the literature (Lüdtke et al., 2020; Murray, 2018): joint modeling, sequential modeling, and fully conditional specification (FCS). In joint modeling, the joint distribution is specified directly, and the same synthesis model is used to draw synthetic data for all variables in $\mathbf{x}$ simultaneously. In sequential modeling, the joint distribution is factorized into a sequence of univariate models, one for each variable, where each model is conditioned on the variables placed earlier in the sequence. Specifically, if $\mathbf{x}_{(<j)} = (x_1, \ldots, x_{j-1})$ denotes the variables in $\mathbf{x}$ that occur before x_j in the sequence, then the m-th synthetic data set is drawn from:

$$\mathbf{x}_{syn}^{(m)} \sim P(\mathbf{x}|\mathbf{z}) = \prod_{j=1}^{p} P(x_j|\mathbf{x}_{(<j)}, \mathbf{z}) \ . \tag{3}$$

Finally, in FCS, the joint distribution is also specified as a sequence of univariate models, but each model is applied separately and by conditioning on all other variables in the data set. Specifically, if $\mathbf{x}_{(-j)} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$ denotes the variables in $\mathbf{x}$ other than x_j, then the m-th set of synthetic values for each x_j is drawn from:

$$x_{j,syn}^{(m)} \sim P(x_j|\mathbf{x}_{(-j)}, \mathbf{z}) \qquad \text{for all } j = 1, \ldots, p \ . \tag{4}$$

The three approaches sometimes differ in the flexibility that they provide for specifying the joint distribution $P(\mathbf{x}, \mathbf{z})$. For example, if all variables are normally distributed and linearly related, then the predictive distributions implied by the three strategies are equivalent (Hughes et al., 2014). However, if the relations between the variables are nonlinear or the variables comprise a mixture of discrete and continuous data, then it can be difficult to specify a synthesis model that correctly represents the joint distribution of the data, and specifying a sequence of univariate conditional models can be beneficial (e.g., Raab et al., 2018; Reiter, 2005b; see also Raghunathan et al., 2001).

Analysis of Synthetic Data

Once generated, each synthetic data set $(\mathbf{x}_{syn}^{(m)}, \mathbf{z})$ $(m = 1, \ldots, M)$ is analyzed with standard statistical methods, and the results are pooled to obtain one final set of parameter estimates and inferences (Reiter, 2003; see also Raghunathan et al., 2003). Specifically, if $\hat{Q}_m$ is the m-th point estimate of a parameter Q, and $\hat{U}_m$ is its estimated variance (i.e., squared standard error), then the pooled point estimate is:

$$\bar{Q} = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}_m \ , \tag{5}$$

The pooled variance can be calculated as:

$$T = \bar{U} + \frac{B}{M} \ , \tag{6}$$

where

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} \hat{U}_m \quad \text{and} \quad B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{Q}_m - \bar{Q} \right)^2 \tag{7}$$

are the within- and between-imputation variance. The within-imputation variance $\bar{U}$ reflects the uncertainty about Q in the original data, and the between-synthesis variance B represents the uncertainty added by the synthesis. These pooling formulas are similar to those used in missing data analysis (Rubin, 1987) but differ in the expression for the pooled variance, because the variance between the synthetic data sets is a result of only the synthesis but not incomplete data. Other pooling methods exist for different types of analyses (e.g., model comparisons; see Reiter and Raghunathan, 2007)

Data-Augmented MI of Synthetic Data (DA-MI)

Recently, Jiang et al. (2021) introduced a *data-augmented* MI approach (DA-MI) to synthetic data that combines traditional masking procedures with the MI approach. The main motivation of DA-MI is to improve the robustness of the MI approach against misspecification in the synthesis model. The DA-MI approach is illustrated in Figure 1 and consists of two steps. In the first step (masking step), this method creates masked copies of the sensitive variables in **x** by adding noise to the individual observations. The copies are generated as

$$\mathbf{x}^* \sim P(\mathbf{x}^* | \mathbf{x}; \boldsymbol{\lambda}) \ , \tag{8}$$

where $\boldsymbol{\lambda}$ is a vector of tuning parameters that specifies the amount of noise added to the original values in **x**. For example, for a continuous variable x_j, a masked copy can be generated as follows. For each unit i,

$$x^*_{ji} = x_{ji} + e_{ji} \quad \text{with} \quad e_{ji} \sim N(0, \sigma^2_e) \ , \tag{9}$$

where the variance σ^2_e controls the amount of noise that is added to the original values x_{ji}. One or multiple copies can be generated in this manner.

In the second step (synthesis step) of the DA-MI approach, synthetic data are generated

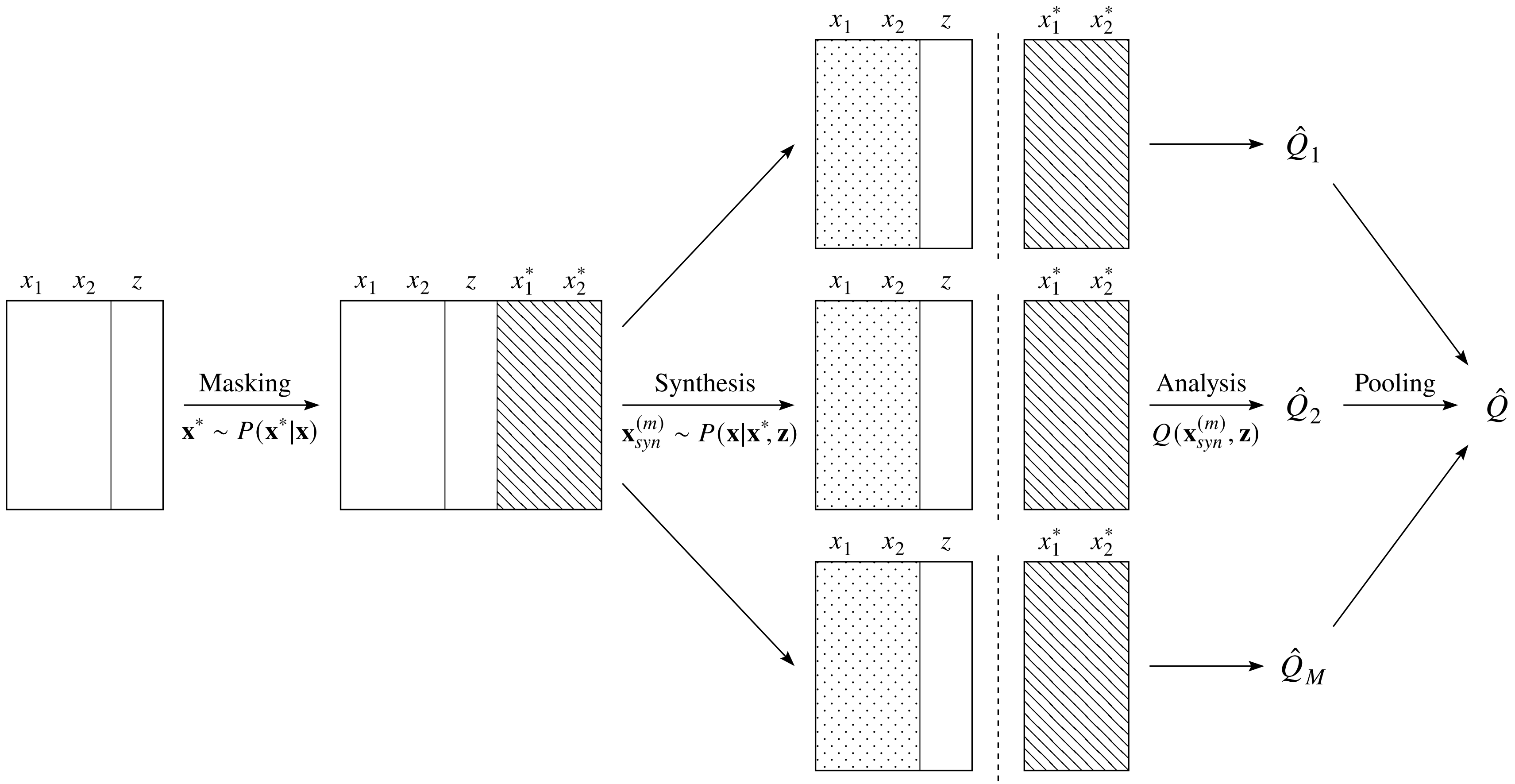

x_1
x_2
z
Masking
$\mathbf{x}^* \sim P(\mathbf{x}^*|\mathbf{x})$
x_1^*
x_2^*
Synthesis
$\mathbf{x}_{syn}^{(m)} \sim P(\mathbf{x}|\mathbf{x}^*, \mathbf{z})$
Analysis
$Q(\mathbf{x}_{syn}^{(m)}, \mathbf{z})$
$\hat{Q}_1$
$\hat{Q}_2$
$\hat{Q}_M$
Pooling
$\hat{Q}$

Figure 1

Schematic representation of the data-augmented multiple imputation (DA-MI) approach to synthetic data. The procedure involves the creation of masked copies (x_1^* and x_2^*) of the sensitive variables (x_1 and x_2, masking step), after which the synthetic data are generated on the basis of both the observed data and the masked copies (synthesis step). The masked copies are then deleted, and the synthetic data are released to the public. Finally, the synthetic data are analyzed with standard statistical methods (analysis step), and the results are pooled to obtain a final set of results (pooling step).

from the predictive distribution of $\mathbf{x}$ given both $\mathbf{z}$ and the masked copies. This is similar to the conventional MI approach to synthetic data but with the copies $\mathbf{x}^*$ included as additional variables in the synthesis model. The role of the copies in the synthesis model can be compared to the role of auxiliary variables in the imputation of missing data. Specifically, the m-th synthetic data set is simulated by drawing from the predictive distribution:

$$\mathbf{x}_{syn}^{(m)} \sim P(\mathbf{x}|\mathbf{x}^*, \mathbf{z}) \ . \tag{10}$$

The synthetic data are then analyzed separately, and the results are pooled as in the conventional MI approach to synthetic data (see Reiter, 2003; Reiter & Raghunathan, 2007). Note that the masked copies $\mathbf{x}^*$ themselves will not be synthesized or released with the data. The only purpose of including the copies in the synthesis is that they provide additional information about the original data that is independent on the relations specified in the synthesis model.

Due to the inclusion of the masked copies, the DA-MI approach can be expected to be more robust to misspecification in the synthesis model than the conventional MI approach, which relies exclusively on the relations specified in the synthesis model to generate the synthetic data. However, an important side-effects of adding the masked copies is that the synthesized values in DA-MI will resemble the original values more closely than in MI, implying a somewhat weaker protection of confidentiality. These properties depend on the amount of noise added in the masking step: The more noise is added, the more protected the original values will be at the expense of making the procedure more susceptible to misspecification in the synthesis model. As a result, the DA-MI approach provides a mechanism for tuning the utility of the data for the required level of confidentiality protection by carefully choosing the amount of noise added in the masking step (Jiang et al., 2021). Techniques such as shuffling the rows in the synthetic data sets can often be used to offset this loss of confidentiality protection. In this context, it is also worth noting that the MI approach can be regarded as a special case of DA-MI. Specifically, as the amount of noise in the masked copies increases, the contribution of the masked copies in the synthesis decreases, making DA-MI more similar to MI.

Specifying the Synthesis Model in DA-MI

The DA-MI approach again requires the specification of a joint distribution, $P(\mathbf{x}, \mathbf{x}^*, \mathbf{z})$, that implies a predictive distribution for the variables to be synthesized, and there are two broad strategies for specifying this distribution. The crucial difference between them is how the relationship between the original data $\mathbf{x}$ and the masked copies $\mathbf{x}^*$ is taken into account (for a similar discussion related to auxiliary variables, see Daniels et al., 2014). In the first strategy, the masked copies $\mathbf{x}^*$ are treated as additional *predictor* variables in the synthesis model, and the m-th synthetic data set is generated directly from:

$$\mathbf{x}_{syn}^{(m)} \sim P(\mathbf{x}|\mathbf{x}^*, \mathbf{z}) \ . \tag{10 revisited}$$

In this approach, both the masked copies and the variables in $\mathbf{z}$ are treated as fixed, and no distributional assumptions are made about them. Henceforth, we refer to this strategy as DA-MI$_\text{P}$. In the second strategy, the masked copies $\mathbf{x}^*$ are treated as additional *outcome* variables, and the m-th synthetic data set is generated from:

$$\mathbf{x}_{syn}^{(m)} \sim P(\mathbf{x}|\mathbf{x}^*, \mathbf{z}) \propto P(\mathbf{x}^*|\mathbf{x})P(\mathbf{x}, \mathbf{z}) \ , \tag{11}$$

where $P(\mathbf{x}, \mathbf{z})$ is the joint distribution of the original data, and $P(\mathbf{x}^*|\mathbf{x})$ is the masking model that was used in the masking step. This is the strategy recommended by Jiang et al. (2021), and we will call it DA-MI$_\text{O}$ for short. In this strategy, the joint distribution of the original data can itself be specified in multiple ways. For example, $P(\mathbf{x}, \mathbf{z})$ can be specified by treating all variables in $\mathbf{z}$ as predictor variables, $P(\mathbf{x}, \mathbf{z}) \propto P(\mathbf{x}|\mathbf{z})$, which again means that the variables in $\mathbf{z}$ can be treated as fixed. As an alternative, $P(\mathbf{x}, \mathbf{z})$ can be specified as a sequence of univariate conditional models. Let $\mathbf{v} = (\mathbf{x}, \mathbf{z})$ denote the combined set of variables. Then the sequence can be specified as $P(\mathbf{x}, \mathbf{z}) = P(\mathbf{v}) = \prod_{k=1}^{p+q} P(v_k|\mathbf{v}_{(<k)})$ where $\mathbf{v}_{(<k)}$ denotes all the variables placed earlier in the sequence, and where the variables in $\mathbf{x}$ and $\mathbf{z}$ can appear at any point in the sequence.

Note that if the $\mathbf{x}$ follow a multivariate normal distribution conditioned on $\mathbf{z}$, and normal masking models for continuous data are used (Equation 9), then the two strategies coincide, and the predictive distribution $P(\mathbf{x}|\mathbf{x}^*, \mathbf{z})$ is also multivariate normal (Gelman et al., 2014). Specifically, suppose that the original data $\mathbf{x}|\mathbf{z}$ are multivariate normally distributed, $\mathbf{x}|\mathbf{z} \sim N(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{z}})$, and that the masked copies also follow a multivariate normal distribution, $\mathbf{x}^*|\mathbf{x} \sim N(\mathbf{x}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is a known diagonal matrix. Then the posterior predictive distribution of $\mathbf{x}|\mathbf{x}^*, \mathbf{z}$ is also multivariate normal with covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{x}^*,\mathbf{z}} = (\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|\mathbf{z}} + \boldsymbol{\Lambda}^{-1})^{-1}$ and mean vector $\boldsymbol{\mu}_{\mathbf{x}|\mathbf{x}^*,\mathbf{z}} = (\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|\mathbf{z}} + \boldsymbol{\Lambda}^{-1})^{-1} \, (\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|\mathbf{z}}\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}} + \boldsymbol{\Lambda}^{-1}\mathbf{x})$. In the special case of normally distributed and linearly related variables, treating the copies $\mathbf{x}^*$ as outcomes or predictors should therefore yield similar results.

In more general cases, the two strategies have different advantages and disadvantages. For example, if the relations between the variables are nonlinear, then DA-MI$_\text{P}$ can add to the misspecification in the synthesis model, whereas DA-MI$_\text{O}$ can accommodate nonlinear effects in a more flexible manner. This is because DA-MI$_\text{P}$ specifies a conditional distribution for $\mathbf{x}$ given $\mathbf{x}^*$, which is difficult to specify correctly when there are nonlinear effects, whereas DA-MI$_\text{O}$ includes the (known) masking model directly in the synthesis. From a practical perspective, however, DA-MI$_\text{P}$ can be advantageous because it allows the variables in $\mathbf{z}$ to be treated as fixed and can be implemented with software for conventional MI of synthetic data (e.g., the “synthpop” package in R; Nowok et al., 2016). By contrast, DA-MI$_\text{O}$ requires the use of specialized software (see Jiang et al., 2021).

Illustrative Example

We now illustrate the two strategies for implementing DA-MI (DA-MI$_\text{P}$ and DA-MI$_\text{O}$) for a simple example with two continuous variables x_1 and x_2 and a covariate z, where the aim is to generate synthetic data for x_1 and x_2. For instance, x_1 and x_2 may represent conscientiousness and academic achievement, whereas z may represent interest in the subject. We assume that masked copies x_1^* and x_2^* of the original x_1 and x_2 have been created and added to the data set. In DA-MI$_\text{P}$, the masked copies x_1^* and x_2^* are treated as additional predictors, and the synthetic data are simulated from:

$$\mathbf{x}_{syn}^{(m)} \sim P(x_1, x_2 | x_1^*, x_2^*, z) \tag{12}$$

In cases where this model can be specified directly, conventional software for MI can be used to draw the synthetic values $\mathbf{x}_{syn}^{(m)}$ (e.g., the packages "norm" or "jomo" in R; Quartagno et al., 2019; Schafer and Olsen, 1998). Otherwise, the synthetic data can be generated from a sequential model:

$$\mathbf{x}_{syn}^{(m)} \sim P(x_2 | x_1, x_1^*, x_2^*, z) P(x_1 | x_1^*, x_2^*, z) \ , \tag{13}$$

which can be implemented in the "synthpop" package in R (Nowok et al., 2016) or similar software by adding the masked copies as additional predictor variables in the synthesis model.

In DA-MI$_\text{O}$, the masked copies x_1^* and x_2^* are treated as outcomes, and the joint distribution of the original variables, x_1, x_2, and z can be specified in multiple ways. For example, if z is treated as a (fixed) covariate, the synthetic data can be generated from:

$$\mathbf{x}_{syn}^{(m)} \sim P(x_1^* | x_1) P(x_2^* | x_2) P(x_1, x_2 | z) \ , \tag{14}$$

where $P(x_1^* | x_1)$ and $P(x_2^* | x_2)$ are the masking models for x_1 and x_2. As an alternative, a sequential modeling approach can be used, where x_1, x_2, and z can occur at any point in the sequence. For example, if the intended analyses regard z as an outcome of both x_1 and x_2, the model can be specified as:

$$\mathbf{x}_{syn}^{(m)} \sim P(x_1^* | x_1) P(x_2^* | x_2) P(z | x_1, x_2) P(x_1 | x_2) P(x_2) \ . \tag{15}$$

The key difference by which DA-MI$_\text{O}$ differs from DA-MI$_\text{P}$ is that the masking models $P(x_1^* | x_1)$ and $P(x_2^* | x_2)$ are directly included in the synthesis model in DA-MI$_\text{O}$, whereas the relations between the original variables and the masked copies are modeled implicitly in DA-MI$_\text{P}$.

Regardless of the strategy used to implement DA-MI, its key feature is that it incorporates the masked copies as additional variables in the synthesis model. The masked copies act as close proxies for the original values and provide information that is independent of the presumed relations between the variables that are specified in the synthesis model, potentially making DA-

MI more robust to misspecification (Jiang et al., 2021). For this reason, DA-MI may be an attractive alternative to the conventional MI approach to synthetic data, particularly when synthetic data are intended to support analyses that go beyond the analyses in the original study and that may be difficult to anticipate. In the sections that follow, we describe the results from a simulation study that evaluated the properties of the MI and DA-MI approaches to synthetic data with regard to reproducibility and robustness.

Study 1: Reproducibility and Robustness

In Study 1, we attempted to evaluate the statistical properties of the MI and DA-MI approaches to synthetic data in terms of reproducibility and the investigation of robustness. To this end, we investigated the behavior of these methods both with correctly specified synthesis models and under different types of misspecification, reflecting applications of synthetic data in which the analyses were either consistent with the relations that were featured in the original analysis and specified in the synthesis model (reproducibility Stages 1 and 2) or went beyond them (Stages 3 and 4). The materials used to conduct this study, including the computer code and all syntax files, are provided on the OSF (https://osf.io/3a5uq/).

Data Generation

The data generating model featured three standardized normal variables x, y, and z. To investigate different types of misspecification, we assumed that the relation between x and y was either linear or quadratic and that z represented an additional variable that could be used in additional analyses but was omitted from the synthesis model. We generated the data in two steps. First, we simulated x and z from a bivariate normal distribution. For unit i ($i = 1, \dots, n$):

$$(x_i, z_i)^T \sim N\left(0, \begin{bmatrix} 1 & \\ \rho_{xz} & 1 \end{bmatrix}\right) . \tag{16}$$

Second, we simulated y as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \ , \quad N(0, 1 - R^2_{yx}) \ , \tag{17}$$

Table 2

Synthesis Methods and Specifications of the Synthesis Model in Study 1

	Synthesis method		
	MI	DA-MI$_P$	DA-MI$_O$
Specification	$P(y\|x)P(x)$	$P(y\|x, y^*, x^*)P(x\|y^*, x^*)$	$P(y^*\|y)P(x^*\|x)P(y\|x)P(x)$
Linear	$P(y\|x)$: $y \sim x$ $P(x)$: x	$P(y\|x, y^*, x^*)$: $y \sim x + y^* + x^*$ $P(x\|y^*, x^*)$: $x \sim x^* + y^*$	$P(y\|x)$: $y \sim x$ $P(x)$: x $P(y^*\|y)$: fixed[a] $P(x^*\|x)$: fixed[a]
Linear + quadratic	$P(y\|x)$: $y \sim x + x^2$ $P(x)$: x	$P(y\|x, y^*, x^*)$: $y \sim x + x^2 + x^* + y^* + x^{*2} + y^{*2}$ $P(x\|y^*, x^*)$: $x \sim x^* + y^* + x^{*2} + y^{*2}$	$P(y\|x)$: $y \sim x + x^2$ $P(x)$ x $P(y^*\|y)$: fixed[a] $P(y^*\|y)$: fixed[a]

[a] The parameters in the masking models were fixed to the values that were used to generated the masked copies in the masking step.

where β_0, β_1, and β_2 were chosen in such a way that y would have a mean of zero, and the quadratic effect of x would contribute to the total R^2_{xy} with a given weight w (e.g., $w = .50$ for half of the total R^2_{yx}).

Synthesis

To generate synthetic data, we used three different methods with two different specifications of the synthesis model. The three methods included (1) the MI approach, (2) the DA-MI approach with the masked copies treated as predictors (DA-MI$_P$), and (3) the DA-MI approach with the masked copies treated as outcomes (DA-MI$_O$). The two specifications for each method are summarized in Table 2 and included (1) a specification with only linear effects and (2) a model with both linear and quadratic effects of x on y. The specification with only linear effects is misspecified when the true relation between x and y is quadratic, whereas the specification with linear *and* quadratic effects is specified correctly (or overspecified).

For DA-MI, we generated one masked copy each for x and y in accordance with Equation 9, where the noise variance σ^2_e was chosen in such a way that the masked copy would represent the original variable with a reliability of

$$\lambda = \frac{1}{1 + \sigma_e^2} \ . \tag{18}$$

To implement MI and DA-MI$_\mathrm{P}$, we used the sequential modeling approach in "synthpop" (Nowok et al., 2020). To implement DA-MI$_\mathrm{O}$, we used the Stan modeling language[1] and the R package "rstan" (Stan Development Team, 2020, 2021).

Simulated Conditions

In the data generating model, we fixed the sample size (n) to 250, the correlation between x and z (ρ_{xz}) to .50, and the total R^2_{yx} to .25. The constant values were chosen on the basis of preliminary simulations in which we had found that these factors had little impact on the results. We varied the contribution of the quadratic effect to the total R^2_{yx} (w) to reflect conditions with ($w = .50$) and without ($w = 0$) a quadratic effect of x on y. For the synthesis methods, we fixed the number of syntheses to $M = 20$. For DA-MI$_\mathrm{P}$ and DA-MI$_\mathrm{O}$, we varied the reliability (λ) of the masked copies (λ = .01, .20, .40, .60, .80, .90, .95, .99). This allowed us to investigate the effects of different amounts of noise added in the masking step. Each condition was replicated 10,000 times.

These simulated conditions, in combination with the different synthesis methods, allowed us to investigate the effects of misspecification in the synthesis model with regard to both the functional relation between x and y and the omitted variable z. Specifically, in conditions with quadratic effects, a linear synthesis model reflects applications in which the researcher generating the synthetic data does not take the quadratic effects into account, for example, because the original analysis did not include them. By contrast, a linear and quadratic synthesis model reflects applications in which the researcher takes the quadratic effects into account, either because such effects were included in the original analysis or because the researcher anticipates that other researchers might conduct robustness analyses that include quadratic effects. In conditions with

[1] The implementation of DA-MI$_\mathrm{O}$ involved two steps. First, we obtained maximum likelihood (ML) estimates of the parameters of the synthesis model. Second, we sampled synthetic values for the variables on the basis of the estimated parameters, the masked data, and the known parameters of the masking models. In this step, we used 1,000 burn-in iterations and generated the 20 synthetic data sets spaced 100 iterations apart. This reflects an "improper" approach to synthetic data, similar to the one used in "synthpop" (Raab et al., 2018), which relies on estimated parameters rather than samples from a Bayesian posterior distribution (Reiter & Kinney, 2012).

only linear effects, all synthesis methods correctly specify this relation. The omitted variable z reflects conditions in which an analyst requests or adds data for additional variables that were not included in the synthesis.

Analysis

In the analysis of the synthetic data, we calculated the means and standard deviations of x and y. In addition, we fit the quadratic regression of y on x and the linear regression of z on x. Each data set was analyzed separately with standard methods, and the results were pooled in accordance with Reiter (2003) to obtain the final point estimates, standard errors, and 95% confidence intervals (CIs).

Evaluation criteria

To evaluate the performance of the synthesis methods, we considered two types of properties. First, to evaluate the accuracy of the parameter estimates and inferences with respect to the true (i.e., population) values, we calculated the bias and the coverage rates of the 95% CIs. Second, to investigate how well the synthetic data reproduced the results in the original data in each replication of the simulation, we calculated the standardized bias

$$\text{Std. Bias} = E\left(\frac{\left|\hat{\theta}_i^{syn} - \hat{\theta}_i^{orig}\right|}{\widehat{SE}_i^{orig}}\right) , \tag{19}$$

which provides a measure of absolute bias in "standard error units," and the relative overlap between the 95% CIs (Karr et al., 2006)

$$\text{CI Overlap} = E\left(\max\left\{0\,,\ \frac{1}{2}\left[\frac{U_i^{over} - L_i^{over}}{U_i^{orig} - L_i^{orig}} + \frac{U_i^{over} - L_i^{over}}{U_i^{syn} - L_i^{syn}}\right]\right\}\right) , \tag{20}$$

where L and U refer to the lower and upper bounds of the confidence intervals, and where $L_i^{over} = \max(L_i^{orig}, L_i^{syn})$ and $U_i^{over} = \min(U_i^{orig}, U_i^{syn})$ are the bounds of the portion that overlaps between the CIs.

Results

The results are summarized in Table 3. Due to the large number of results, we present only the main results and focus on the estimates of the means and the regression coefficients. The remaining results are provided on the OSF and in Supplement C in the online supplemental materials (https://osf.io/3a5uq/).

The bias in the parameter estimates obtained with MI was highly dependent on the specification of the synthesis model. Specifically, the MI approach led to unbiased estimates of the quadratic effect of x on y only when that effect was either zero or when it was included in the synthesis model. Likewise, MI led to biased estimates of the effect of x on z throughout. By contrast, DA-MI$_\text{P}$ and DA-MI$_\text{O}$ provided estimates with little or no bias even when the synthesis model was misspecified, provided that the reliability of the masked copies was sufficiently high ($\lambda \geq .95$). This was true for both the quadratic effect of x on y, when it was omitted from the synthesis model, and the effect of x on z. The coverage rates for the 95% CIs followed roughly the same pattern and were usually close to the nominal value of 95% when the estimates were unbiased. However, in the MI approach, the coverage rates for the quadratic effect of x on y were sometimes above or below 95% even when the estimates were unbiased.

The effect of the reliability of the masked copies in DA-MI on the bias is shown in more detail in Figure 2. Two findings are worth noting. First, DA-MI$_\text{P}$ and DA-MI$_\text{O}$ provided nearly the same results as MI when the reliability was near zero ($\lambda = .01$) and nearly unbiased results when the reliability was near perfect ($\lambda = .99$). Second, when the synthesis model included both linear and quadratic effects, only DA-MI$_\text{O}$ provided unbiased estimates of the quadratic effect of x on y, whereas DA-MI$_\text{P}$ led to biased estimates of the quadratic effect of x when the reliability was at an intermediate level ($.20 \leq \lambda \leq .80$), even when the quadratic effect was included in the synthesis model.

The results for the standardized bias and the CI overlap suggested that MI and DA-MI allowed us to reproduce the results in the original data to similar extents when the synthesis model was correctly specified. However, when the synthesis model was misspecified, then DA-MI

Table 3

Results for the Bias and Standardized Bias of the Parameter Estimates and for the Relative Overlap and Coverage Rates of the 95% Confidence Intervals for Selected Parameters in Conditions with a High Reliability of the Masked Copies in DA-MI (λ = .95) in Study 1

			Bias			95%-CI Coverage (%)			Std. Bias			95%-CI Overlap (%)		
w	Par.	True	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$
Synthesis Model with Linear Effects														
0	$\hat{\mu}_x$	0.000	−0.001	−0.001	−0.001	95.1	95.1	94.6	0.178	0.039	0.175	95.4	99.0	95.5
	$\hat{\mu}_y$	0.000	−0.001	0.000	0.000	95.0	95.1	94.4	0.181	0.040	0.176	95.3	99.0	95.5
	$\hat{\beta}_{y\sim x}$	0.500	0.000	0.000	0.000	94.9	94.7	93.2	0.212	0.081	0.270	94.5	97.9	93.1
	$\hat{\beta}_{y\sim x^2}$	0.000	0.000	−0.001	−0.001	100.0	96.6	96.7	0.823	0.336	0.337	79.3	91.4	91.3
	$\hat{\beta}_{z\sim x}$	0.500	−0.500	−0.024	−0.024	0.0	93.6	93.3	9.137	0.451	0.450	0.0	88.6	88.6
0.5	$\hat{\mu}_x$	0.000	0.000	0.000	0.000	95.4	95.5	94.9	0.179	0.040	0.173	95.4	99.0	95.6
	$\hat{\mu}_y$	0.000	0.000	0.000	0.000	95.0	94.8	94.3	0.178	0.040	0.177	95.4	99.0	95.5
	$\hat{\beta}_{y\sim x}$	0.354	0.000	−0.005	−0.005	89.3	95.0	94.0	0.736	0.244	0.346	81.9	93.8	91.2
	$\hat{\beta}_{y\sim x^2}$	0.250	−0.250	−0.038	−0.038	0.0	86.0	85.4	6.368	0.967	0.973	0.3	75.7	75.4
	$\hat{\beta}_{z\sim x}$	0.500	−0.500	−0.024	−0.024	0.0	93.5	93.5	9.139	0.456	0.455	0.0	88.5	88.5
Synthesis Model with Linear and Quadratic Effects														
0	$\hat{\mu}_x$	0.000	−0.001	−0.001	−0.001	95.0	95.1	94.5	0.181	0.040	0.175	95.3	99.0	95.5
	$\hat{\mu}_y$	0.000	0.000	0.000	0.000	95.3	95.1	94.4	0.181	0.039	0.177	95.3	99.0	95.5
	$\hat{\beta}_{y\sim x}$	0.500	0.000	0.000	0.000	94.9	94.7	93.2	0.178	0.070	0.267	95.3	98.2	93.2
	$\hat{\beta}_{y\sim x^2}$	0.000	−0.001	−0.001	−0.001	95.0	95.3	92.9	0.181	0.203	0.307	94.5	94.7	92.0
	$\hat{\beta}_{z\sim x}$	0.500	−0.500	−0.024	−0.024	0.0	93.4	93.4	9.143	0.453	0.450	0.0	88.6	88.6
0.5	$\hat{\mu}_x$	0.000	0.000	0.000	0.000	95.4	95.6	94.9	0.179	0.040	0.172	95.4	99.0	95.6
	$\hat{\mu}_y$	0.000	0.003	0.000	0.000	95.0	94.8	94.2	0.181	0.040	0.177	95.2	99.0	95.5
	$\hat{\beta}_{y\sim x}$	0.354	0.000	0.000	0.001	95.2	95.5	93.6	0.178	0.104	0.293	95.3	97.3	92.5
	$\hat{\beta}_{y\sim x^2}$	0.250	0.000	−0.019	0.001	95.0	93.0	92.8	0.182	0.514	0.330	94.5	86.9	91.5
	$\hat{\beta}_{z\sim x}$	0.500	−0.500	−0.024	−0.024	0.0	93.5	93.6	9.138	0.454	0.448	0.0	88.5	88.7

Note. w = contribution of x^2 to the total R^2_{yx}; μ_x, μ_y = means of x and y; $\beta_{y\sim x}$, $\beta_{y\sim x^2}$ = coefficients in the quadratic regression of y on x; $\beta_{z\sim x}$ = coefficient in the linear regression of z on x; MI = synthesis with MI; DA-MI$_\text{P}$ = synthesis with DA-MI and masked copies as predictors; DA-MI$_\text{O}$ = synthesis with DA-MI and masked copies as outcomes.

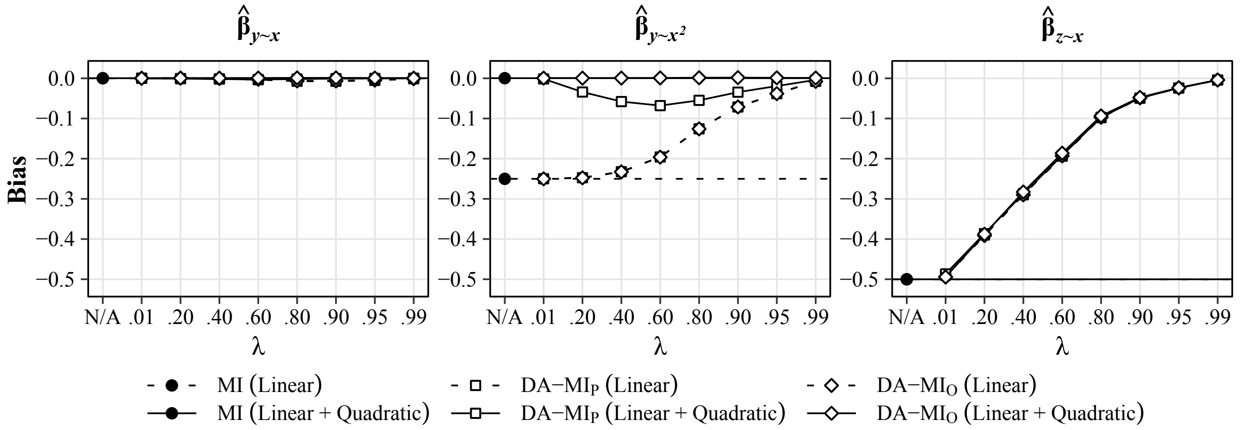


Figure 2

Bias in the estimated coefficients for the regression of y on x ($\beta_{y\sim x}$ and $\beta_{y\sim x^2}$) and the regression of z on x ($\beta_{z\sim x}$) depending on the synthesis method and the reliability (λ) of the masked copies in DA-MI in Study 1. MI = synthesis with MI; DA-MI_P = synthesis with DA-MI and masked copies as predictors; DA-MI_O = synthesis with DA-MI and masked copies as outcomes.

tended to outperfom MI. For example, in the condition with a quadratic effect of *x* on *y* ($w = 0.5$), the average difference between the estimated quadratic effect in MI versus the original data was more than nine times the size of the standard error with a CI overlap of 0.3%. In DA-MI ($\lambda = .95$), the average difference was reduced to less than one standard error unit with a CI overlap of about 75%. The differences between DA-MI_P and DA-MI_O tended to be small, but DA-MI_P provided estimates of the means of *x* and *y* and the linear effect of *x* on *y* that were more similar to those in the original data than the ones in DA-MI_O.

Summary

Overall, the results showed that the DA-MI approach has the potential to be more robust to misspecification in the synthesis model and provide results closer to the original data than the conventional MI approach, depending on the reliability of the masked copies. This indicates that DA-MI may be an attractive alternative to MI for generating synthetic data that allow for a close reproduction of the original findings and support a wider variety of robustness analyses that may be difficult to anticipate a priori when specifying the synthesis model. The results also showed that the two strategies for DA-MI (DA-MI_P and DA-MI_O) led to similar albeit not identical results.

Specifically, when a linear synthesis model was used, both DA-MI$_\mathrm{P}$ and DA-MI$_\mathrm{O}$ provided estimates with similar amounts of bias, but DA-MI$_\mathrm{P}$ tended to reproduce the original results more closely. By contrast, only DA-MI$_\mathrm{O}$ was able to reproduce nonlinear effects without bias. This suggests that treating the masked copies as outcomes (DA-MI$_\mathrm{O}$) is the more flexible approach overall, whereas treating the masked copies as predictors (DA-MI$_\mathrm{P}$) can improve reproducibility in simple applications with only linear effects. In the following, we further explore the properties of these methods with particular regard to the number of masked copies and the number of syntheses.

Study 2a: Number of Masked Copies

In Study 2a, we aimed to explore the effects of the number of masked copies created in the masking step of DA-MI. Specifically, although Jiang et al. (2021) recommended using multiple copies per variable (10 to 20), we argue that DA-MI can be implemented even with a single copy per variable as long as the (total) reliability of the masked copies remains the same (for a more formal argument, see the Appendix). This would simplify the application of DA-MI and provide a clearer mechanism for balancing the utility of the synthetic data with the required level of confidentiality protection.

To this end, we used the same simulation procedures that we used in Study 1 but now generated k masked copies per variable in DA-MI. For each copy, we chose the noise variance in such a way that the sum of the copies represented the original variable with a (total) reliability of

$$\lambda_k = \frac{1}{1 + \sigma_e^2/k} = \frac{k\left(\frac{1}{1+\sigma_e^2}\right)}{1 + (k-1)\left(\frac{1}{1+\sigma_e^2}\right)} \, , \tag{21}$$

which is just the Spearman-Brown prophecy formula, with the reliability of an individual copy as defined in Equation 18. To evaluate the impact of the number and reliability of the masked copies, we varied the number of masked copies (k = 1, 5, 10, 15, 20) and the total reliability (λ_k = .80, .95) with which they reflected the original values.

Results

The results for the bias in selected parameters are summarized in Figure 3. The remaining results are not shown in detail but are provided in Supplement C. The bias in the parameter estimates obtained from DA-MI was virtually unaffected by the number of masked copies. Specifically, as in Study 1, the bias in DA-MI was lower when the masked copies represented the original variables with high reliability ($\lambda_k = .95$), and the bias was essentially the same regardless of whether a single copy or multiple copies were used (k). The results for the CI coverage rates, the standardized bias, and the CI overlap followed a similar pattern and were also unaffected by the number of masked copies. These results show that the performance DA-MI depends on the (total) reliability of the masked copies rather than their number. This suggests that DA-MI can be implemented even with a single copy, which simplifies the specification of the masking and synthesis models in practice. However, recall that the masked copies are only included to provide additional information about the original data in the synthesis and are not released with the synthetic data.

Study 2b: Number of Syntheses

In the statistical literature, there are diverging recommendations for the number of syntheses that should be used with synthetic data (e.g., Raab et al., 2018; Reiter, 2003). For this reason, in Study 2b, we investigated the effects of the number of syntheses on the statistical properties of MI and DA-MI in terms of reproducibility and robustness. To this end, we again used the same procedures that we used in Study 1. However, we held the reliability of the masked copies constant ($\lambda = .95$) and varied the number of syntheses ($M = 5, 10, 20, 50, 100$).

Results

The results for the bias, the standardized bias, and the relative overlap of the 95% CIs are summarized in Table 4 for selected parameters. The remaining results are provided in Supplement C. For simplicity, we focus on the results for only the specification of the synthesis methods with

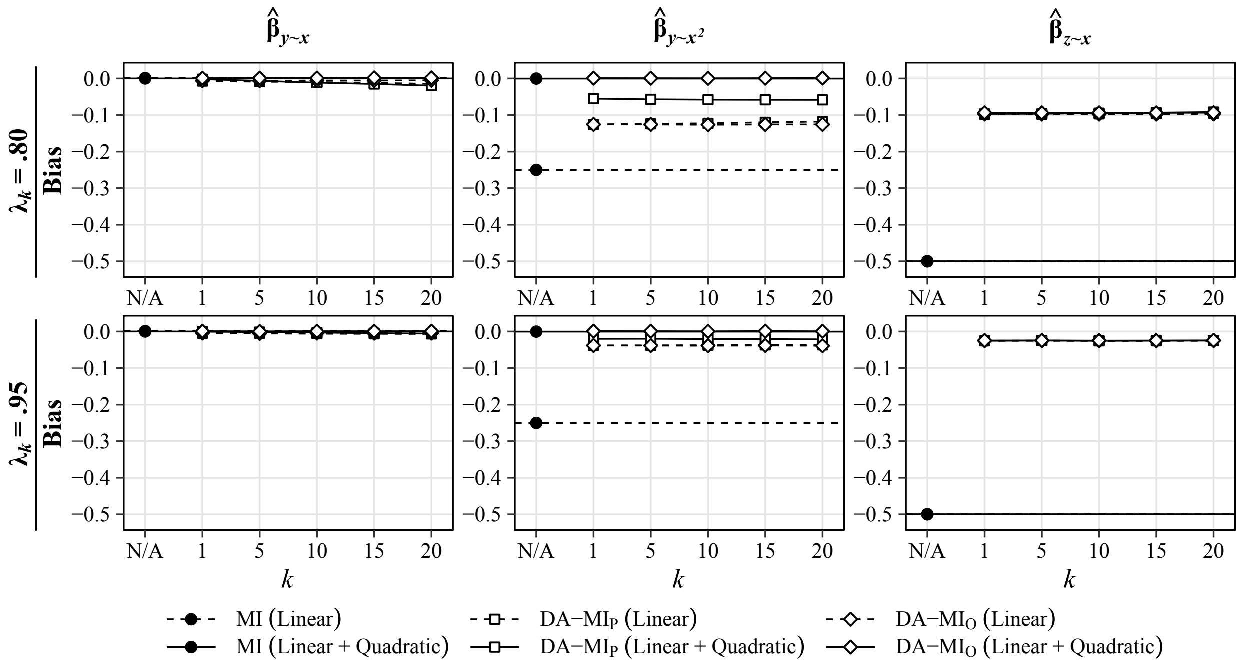


Figure 3

Bias in the estimated coefficients for the regression of y on x ($\beta_{y\sim x}$ and $\beta_{y\sim x^2}$) and the regression of z on x ($\beta_{z\sim x}$) depending on the synthesis method and the total reliability (λ_k) and the number (k) of the masked copies in DA-MI in Study 2a. MI = synthesis with MI; DA-MI$_P$ = synthesis with DA-MI and masked copies as predictors; DA-MI$_O$ = synthesis with DA-MI and masked copies as outcomes.

both linear and quadratic effects. The results indicated that the bias was virtually unaffected by the number of syntheses. However, the standardized bias and the relative overlap of the 95% CIs tended to decrease as the number of syntheses increased, indicating that the parameter estimates were closer and the CIs were more similar to those in the original data. These effects tended to be weaker for DA-MI$_P$ and DA-MI$_O$ and stronger for MI, especially when the parameters of interest were correctly specified in the synthesis model. By contrast, the number of syntheses did not improve the quality of the parameter estimates that were represented incorrectly in the synthesis model. The relative benefits of increasing the number of syntheses were largest when the initial number of syntheses was small (e.g., 5 vs. 20 syntheses), but even a large number of syntheses (e.g., 50 or 100) sometimes still improved the reproducibility of the results. Overall, these results demonstrate that, although even a small number of syntheses can provide parameter estimates with good statistical properties (e.g., low bias), choosing a larger number of syntheses tends to

Table 4

Results for the Standardized Bias of the Parameter Estimates and the Relative Overlap of the 95% Confidence Intervals for Selected Parameters Depending on the Number of Syntheses in Study 2b

			Bias			Std. Bias			95%-CI Overlap (%)		
Par.	M	True	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$	MI	DA-MI$_\text{P}$	DA-MI$_\text{O}$
			Synthesis Model with Linear and Quadratic Effects								
$\hat{\beta}_{y\sim x}$	5	0.354	−0.001	−0.001	0.000	0.359	0.163	0.320	90.4	95.7	91.8
	10	0.354	−0.001	−0.001	0.000	0.255	0.127	0.303	93.3	96.7	92.3
	20	0.354	−0.001	−0.001	0.000	0.180	0.103	0.294	95.3	97.3	92.5
	50	0.354	−0.001	−0.001	0.000	0.114	0.087	0.288	97.0	97.7	92.6
	100	0.354	−0.001	−0.001	0.000	0.080	0.081	0.286	97.8	97.9	92.7
$\hat{\beta}_{y\sim x^2}$	5	0.250	0.000	−0.020	0.001	0.359	0.531	0.358	90.1	86.5	90.8
	10	0.250	0.000	−0.020	0.001	0.256	0.521	0.340	92.7	86.7	91.3
	20	0.250	0.000	−0.020	0.001	0.182	0.516	0.330	94.5	86.8	91.5
	50	0.250	0.000	−0.020	0.001	0.115	0.514	0.324	95.9	86.9	91.6
	100	0.250	0.000	−0.020	0.001	0.081	0.514	0.322	96.5	86.9	91.7
$\hat{\beta}_{z\sim x}$	5	0.500	−0.500	−0.025	−0.024	9.147	0.462	0.454	0.0	88.4	88.5
	10	0.500	−0.500	−0.025	−0.024	9.142	0.459	0.453	0.0	88.4	88.6
	20	0.500	−0.500	−0.025	−0.024	9.139	0.458	0.452	0.0	88.4	88.6
	50	0.500	−0.500	−0.025	−0.024	9.138	0.458	0.450	0.0	88.4	88.6
	100	0.500	−0.500	−0.025	−0.024	9.137	0.458	0.451	0.0	88.4	88.6

Note. M = number of syntheses; μ_x, μ_y = means of x and y; $\beta_{y\sim x}$, $\beta_{y\sim x^2}$ = coefficients in the quadratic regression of y on x; $\beta_{z\sim x}$ = coefficient in the linear regression of z on x; MI = synthesis with MI; DA-MI$_\text{P}$ = synthesis with DA-MI and masked copies as predictors; DA-MI$_\text{O}$ = synthesis with DA-MI and masked copies as outcomes.

improve the ability of the synthetic data to reproduce the results in the original study.

Example Analysis

To illustrate the application of the MI and DA-MI approaches to synthetic data, we used the data on sociosexuality and self-rated attractiveness from Jones and DeBruine (2019; see also Quintana, 2020). The data comprise a sample of N = 9,627 participants, who participated either online or as part of a lab study. They include information about the participants' sex (male, female), age, sexual orientation (men, women, either), self-rated attractiveness, and their sociosexual behavior, attitudes, and desire, which were assessed with the revised version of the Sociosexual Orientation Inventory (SOI-R; Penke and Asendorpf, 2008). For simplicity, we used only the lab sample, which comprised N = 1,346 participants (80.6% female; median age: 20

years) after the removal of five cases with missing data. The data and computer code needed to run this example are provided on the OSF (https://osf.io/3a5uq/).

In this example, we assumed the following hypothetical scenario: Researcher A, who collected the data, is interested in the relation between age and sociosexuality when controlling for age-related differences in self-rated attractiveness. To this end, he conducts linear regression analyses to estimate the effects of age and attractiveness on sociosexual behavior, attitudes, and desire. He finds that age is associated with more liberal sociosexual behavior but not with different attitudes or desire. He publishes his findings along with synthetic data that protect the confidentiality of the data and allow other researchers to reproduce the results (reproducibility Stages 1 and 2). Researcher B is interested in the same topic. However, she believes that the strict focus on linear effects might be misleading and that there could be nonlinear associations that Researcher A did not consider. For this reason, she plans to conduct polynomial regression analyses that also include quadratic and interaction effects (Stage 3).

To generate the synthetic data, we used MI and DA-MI as before. For DA-MI, we created one masked copy per variable with different masking models for the continuous and categorical data. For continuous and Likert-type data (age, attractiveness, sociosexuality), we used the same method as before (Equation 9). For binary and categorical data (sex, sexual orientation), we did the following. For each variable x_j with C_j categories and each unit i, we simulated the masked values conditional on the original values with probabilities:

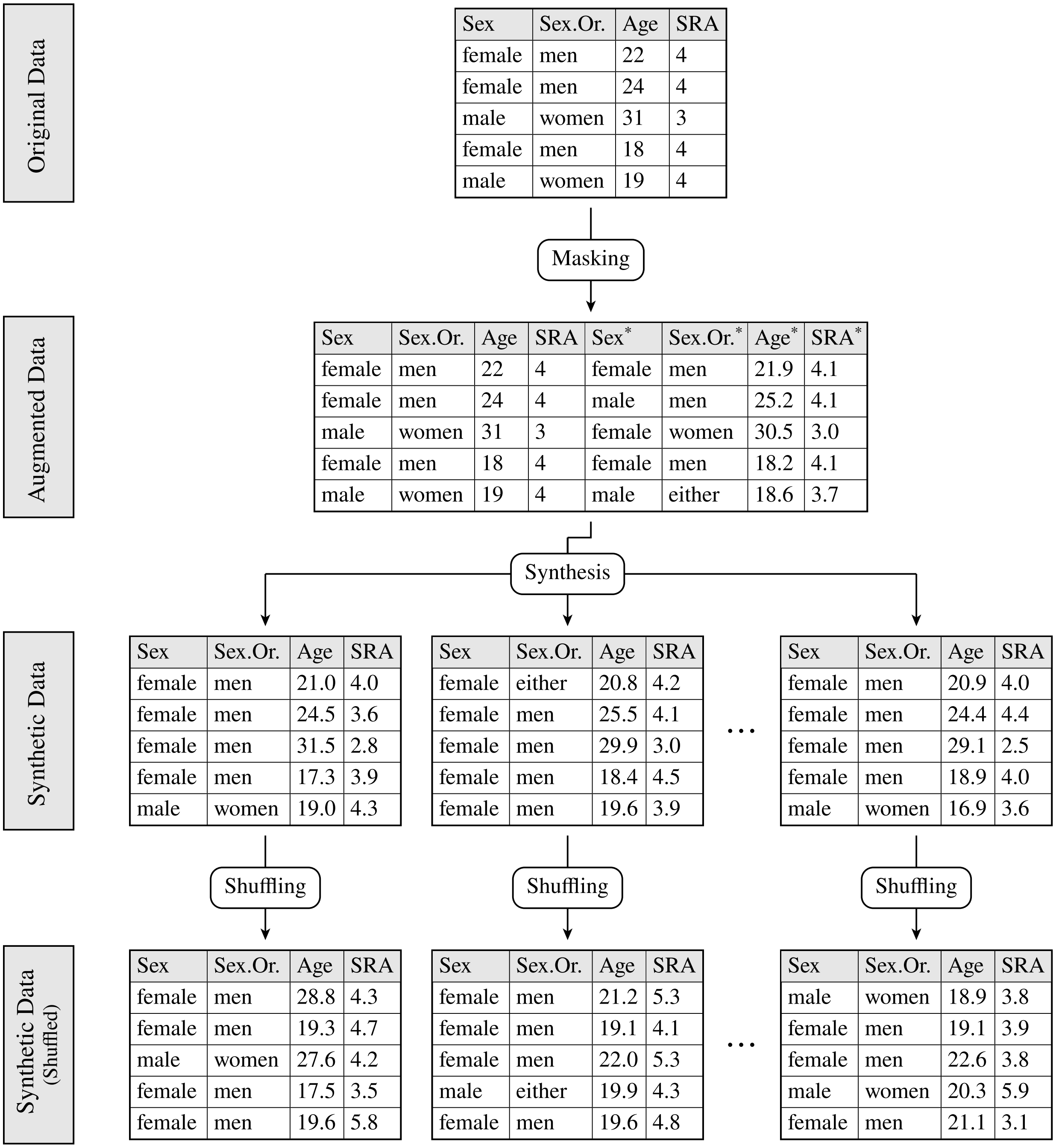
$$P(x^*_{ji} = v | x_{ji} = u) = \pi_{uv} \quad \text{with} \quad \mathbf{\Lambda} = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1C_j} \\ \vdots & \ddots & \vdots \\ \pi_{C_j1} & \cdots & \pi_{C_jC_j} \end{bmatrix} , \tag{22}$$

where $\mathbf{\Lambda}$ is a known transition matrix with rows that sum up to one (Küchenhoff et al., 2006). The diagonal entries of $\mathbf{\Lambda}$ reflect the probability that an observation i has the same value on both x_j and x^*_j, whereas the off-diagonal entries reflect the probability of category transitions. For age, we set the reliability of the masked copy to .90; for the SOI-R items, we set the reliability to .95. For sex and sexual orientation, we set the diagonal entries of the transition matrices to .80 and .60, respectively, and the off-diagonal entries to .20. These values were chosen so that sensitive information such as the participants' sex, sexual orientation, and age is obfuscated more strongly, whereas less sensitive information such as the responses to the items on the SOI-R is masked with only a little noise.

We used MI and DA-MI to generate $M = 20$ synthetic data sets. To implement MI, we used the sequential modeling approach in "synthpop" (Nowok et al., 2020). For DA-MI$_\text{P}$, we also used "synthpop" and included the masked copies as additional predictors. For DA-MI$_\text{O}$, we implemented a sequential modeling approach in a provisional R package (available at https://github.com/simongrund1/robosynth).[2] All approaches used parametric synthesis methods (i.e., linear regression for continuous variables, logistic regression for binary variables, multinomial regression for categorical variables). For both MI and DA-MI, we standardized the variables beforehand to simplify the synthesis and specification of the masking models. Once the synthetic data had been generated, we standardized the variables, computed scale scores for sociosexual behavior, attitudes, and desire, and used them to conduct (a) multiple regression analyses with linear effects of age and attractiveness (Researcher A) and (b) polynomial regression analyses with linear, quadratic, and interaction effects (Researcher B). Furthermore, for DA-MI, we shuffled the rows in each of the synthetic data sets to reduce the identifiability of the synthetic records. The full procedure for DA-MI is further illustrated in Figure 4 (DA-MI$_\text{P}$).

The results of the analyses for sociosexual behavior are shown in Table 5. In the interest of space, we focus on the results for sociosexual behavior and provide the remaining results in Supplement A in the online supplemental materials. In the linear regression analysis (Researcher A), both MI and DA-MI provided results that were very close to those in the original data and indicated a positive effect of age on sociosexual behavior, meaning that older participants tended to report more liberal sociosexual behavior. This similarity between the approaches was expected,

[2] As an alternative, the JAGS software (Plummer, 2017) could be used to implement DA-MI$_\text{O}$ in this example, and we provide the syntax files for such an implementation in Supplement B in the supplemental online materials. This is in contrast to Stan, which does not support simulating values for categorical variables and is therefore limited to applications with continuous data.



Sex	Sex.Or.	Age	SRA
female	men	22	4
female	men	24	4
male	women	31	3
female	men	18	4
male	women	19	4

Sex	Sex.Or.	Age	SRA	Sex*	Sex.Or.*	Age*	SRA*
female	men	22	4	female	men	21.9	4.1
female	men	24	4	male	men	25.2	4.1
male	women	31	3	female	women	30.5	3.0
female	men	18	4	female	men	18.2	4.1
male	women	19	4	male	either	18.6	3.7

Sex	Sex.Or.	Age	SRA
female	men	21.0	4.0
female	men	24.5	3.6
female	men	31.5	2.8
female	men	17.3	3.9
male	women	19.0	4.3

Sex	Sex.Or.	Age	SRA
female	either	20.8	4.2
female	men	25.5	4.1
female	men	29.9	3.0
female	men	18.4	4.5
female	men	19.6	3.9

Sex	Sex.Or.	Age	SRA
female	men	20.9	4.0
female	men	24.4	4.4
female	men	29.1	2.5
female	men	18.9	4.0
male	women	16.9	3.6

Sex	Sex.Or.	Age	SRA
female	men	28.8	4.3
female	men	19.3	4.7
male	women	27.6	4.2
female	men	17.5	3.5
female	men	19.6	5.8

Sex	Sex.Or.	Age	SRA
female	men	21.2	5.3
female	men	19.1	4.1
female	men	22.0	5.3
male	either	19.9	4.3
female	men	19.6	4.8

Sex	Sex.Or.	Age	SRA
male	women	18.9	3.8
female	men	19.1	3.9
female	men	22.6	3.8
male	women	20.3	5.9
female	men	21.1	3.1

Figure 4

Schematic illustration of the steps involved in DA-MI in the Example Analysis (DA-MI$_P$). Each table represent the first five rows and a subset of the columns (sex, sexual orientation, age, self-rated attractiveness) of the data at each step of the procedure.

Table 5

Results for the Linear and Polynomial Regression Models for Sociosexual Behavior in the Example Analysis

	Original Data		MI		DA-MI$_P$		DA-MI$_O$	
	β	(SE)	β	(SE)	β	(SE)	β	(SE)
Researcher A: Linear Regression								
Intercept	0.000	(0.027)	0.000	(0.027)	0.000	(0.027)	0.000	(0.027)
Age	0.152^{***}	(0.027)	0.153^{***}	(0.027)	0.156^{***}	(0.027)	0.154^{***}	(0.027)
SRA	0.099^{***}	(0.027)	0.104^{***}	(0.028)	0.099^{***}	(0.027)	0.093^{***}	(0.027)
Researcher B: Polynomial Regression								
Intercept	0.037	(0.034)	−0.009	(0.039)	0.030	(0.034)	0.029	(0.034)
Age	0.242^{***}	(0.038)	0.153^{***}	(0.027)	0.203^{***}	(0.034)	0.199^{***}	(0.034)
SRA	0.095^{***}	(0.028)	0.103^{***}	(0.028)	0.095^{***}	(0.028)	0.088^{**}	(0.028)
Age^2	-0.033^{**}	(0.010)	0.008	(0.020)	-0.023^{*}	(0.010)	-0.022^{*}	(0.010)
SRA^2	−0.004	(0.019)	0.001	(0.020)	−0.008	(0.019)	−0.008	(0.019)
Age × SRA	0.006	(0.029)	0.000	(0.028)	0.007	(0.029)	0.001	(0.029)

Note. SRA = self-rated attractiveness; MI = synthesis with MI; DA-MI$_P$ = synthesis with DA-MI and masked copies as predictors; DA-MI$_O$ = synthesis with DA-MI and masked copies as outcomes.

because the synthesis models were correctly specified and in line with the intended analyses. In the polynomial regression (Researcher B), for which the synthesis model was misspecified, MI and DA-MI led to different results. Specifically, DA-MI$_P$ and DA-MI$_O$ again provided results similar to the original data and indicated a positive linear and a negative quadratic effect of age, meaning that the age differences followed an inverted U-shape rather than a steady positive trend. By contrast, the results in MI suggested that the effect of age was again strictly linear. These results are in line with the findings presented above and indicate that DA-MI tends to be more robust to misspecification, which can allow both the reproduction of the original results as well as the investigation of additional analyses that were not taken into account in the specification of the synthesis model. In addition to these results, DA-MI$_P$ and DA-MI$_O$ also led to synthetic data that resembled the original data more closely than those in MI (for DA-MI$_P$, see Figure 4). This reflects the fact that DA-MI incorporates additional information in the synthesis through the masked copies, thus increasing the utility of the data but reducing the degree of confidentiality protection. This difference is offset by shuffling the rows in each of the synthetic data sets (last step in Figure 4). The effects of DA-MI on the confidentiality of the data are further explored in Supplement A.

Discussion

Reproducibility is one of the central tenets of credible research practices (Artner et al., 2020; Asendorpf et al., 2013; Bollen et al., 2015; Nelson et al., 2018; Nosek et al., 2021). In the present article, we investigated multiple approaches to synthetic data, which can be used to share surrogate data while protecting the identities of the study participants. Specifically, we aimed to investigate the utility of these methods for improving reproducibility and the investigation of robustness in psychological research. In this context, we distinguished between multiple stages of reproducibility, which included the reproducibility of methods and results (Stages 1 and 2) as well as specific and general robustness analyses that go beyond the analyses used in the original study (Stages 3 and 4). The main challenge in the application of synthetic data is that the synthesis model has to fit the analyses that will be applied to the data (Meng, 1994). In this study, we therefore considered both conventional approaches to synthetic data that are based on MI (Raghunathan et al., 2003; Rubin, 1993) and the novel DA-MI approach, which attempts to make the procedure more robust to model misspecification by including masked copies of the original variables in the synthesis model (Jiang et al., 2021).

Our results showed that the conventional MI approach can perform well but also that its performance depends on the correct specification of the synthesis model. By contrast, the DA-MI approach, which also includes masked copies of the original variables in the synthesis, can mitigate this problem and provide synthetic data that are more robust to different types of misspecification in the synthesis model. As a result, synthetic data generated with DA-MI can be more useful because they allow researchers to analyze the data more liberally, even when the synthesis model was not constructed with these analyses in mind. We also investigated different strategies for implementing DA-MI and found that treating the masked copies as predictors (DA-MI_{P}) often provided results that came closest to mimicking the results in the original analyses but also that this method can produce biased results when nonlinear effects are included in the synthesis model. By contrast, treating the masked copies as outcomes (DA-MI_{P}) was the most flexible approach and produced the least amount of bias with only a slight loss of efficiency.

This suggests that DA-MI could be an attractive alternative to conventional MI approach to synthetic data. The main barrier for using DA-MI in practice is the relative novelty of this approach and the scarcity of statistical software that implements it. In the present article, we demonstrated how DA-MI$_O$ can be implemented in general purpose software for Bayesian analysis (e.g., Stan, JAGS), and we presented an implementation of DA-MI$_O$ for the statistical software R (https://github.com/simongrund1/robosynth). In addition, we showed that DA-MI$_P$ can be implemented in conventional software for synthetic data (e.g., "synthpop"). At least when the synthesis model includes only linear effects, this method can therefore provide a simple option to implement DA-MI, offering the same advantages as DA-MI$_O$ and an improved efficiency in some cases.

The present study has multiple limitations. First, it is important to emphasize that DA-MI attempts to balance the utility of the data with the required level of confidentiality protection. Specifically, although the inclusion of the masked copies in DA-MI can strongly improve the robustness of the procedure, it also tends to produce data that resemble the original data more closely than in MI. In fact, one of the main advantages of the conventional MI approach is that the synthetic data are essentially unrelated to the original data, apart from the relations specified in the synthesis model. For DA-MI, this raises the question of how much noise is needed to protect the confidentiality of the data while maintaining as much of the increased robustness as possible. For example, data that pose little risk to the identification of participants (e.g., items from a personality inventory) may require only a little noise (e.g., reliabilities of .95 or even .99), whereas demographic data (e.g., gender, marital status) may require more noise to prevent identification of participant who have rare combinations of attributes. Researchers who want to apply DA-MI in practice should therefore carefully consider the amount of noise added in the masking step and incorporate additional masking techniques (e.g., shuffling the rows in the synthetic data sets) to minimize the risk of participants being identified from the data. In some cases, it may also be advisable to drop particularly extreme or unusual cases from the data (e.g., the only older participant in a sample of university students) to avoid identity disclosure by DA-MI. Future research should investigate both the overall effects of DA-MI on privacy risk (Karr et al., 2006; Snoke et al., 2018) and the effects of different masking models in more detail (e.g., heteroscedastic models that add more noise to particularly vulnerable units).

Second, our investigation was focused on relatively simple analyses and specific types of misspecification. In principle, DA-MI could also support more general uses of confidentiality-protected data, for example, for secondary analyses of existing data with different research questions (Stage 4) or research syntheses with individual participant data (Riley et al., 2010). However, more research is needed to evaluate the utility of this approach in more general applications. Similarly, we focused on parametric methods for generating synthetic data, but many authors have suggested nonparametric methods that could also offer a flexible representation of the observed variables and the relations between them (Drechsler & Reiter, 2010, 2011; Manrique-Vallier & Hu, 2018; Manrique-Vallier & Reiter, 2014; Reiter, 2005a). Future research should compare these approaches with DA-MI.

There are many other aspects of synthetic data, both theoretical and practical, that we did not consider in detail and that have yet to receive more attention in the methodological literature. For example, empirical data are often incomplete, and there are different approaches to handling missing values in the generation of synthetic data. Several authors have recommended that MI be used to treat missing data, after which the imputed data can be synthesized (e.g., Drechsler & Reiter, 2011; Jiang et al., 2021; Reiter, 2004). By contrast, Raab et al. (2018) recommended the use of missingness indicators that represent missing data and to synthesize missing values alongside the observed data. However, little is known about which method performs best.

Finally, synthetic data are not the only way to facilitate data sharing, nor can synthetic data provide the same value for reproducibility as the primary data from a study. This raises the question of what role synthetic data should play in a larger framework to promote credible research practices. Arguably the best case scenario would be that synthetic data would not be needed and that data could be shared openly. However, if this is not possible, for example, due to confidentiality concerns, then synthetic data can fulfill an important role and allow reproducibility at least to a certain extent. Especially with open data still being an exception rather than the norm (Hardwicke et al., 2021), these methods can improve the transparency and credibility of psychological research, even if they allow reproducing only the analytic procedures (Stage 1) or the main results (Stage 2) of an original study (see also Artner et al., 2020). These methods can be employed by both individual researchers who want to enable other researchers to check the validity and robustness of their results and public agencies that offer services to the scientific community, for example, by providing liberal access to synthetic data when access to the original data must be restricted. We hope that this article will promote a wider application of these methods and encourage further research on the potential utility of synthetic data in psychological research.

References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000365

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. https://doi.org/10.1002/per.1919

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*, 452. https://doi.org/10.1038/533452a

Bollen, K. A., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. A. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science*. https://www.nsf.gov

Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Lachowicz, M., Rose, B. A., Shaw, T., Stapleton, L. M., Woolley, M. E., & Zheng, Y. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*, *12*, 616–647. https://doi.org/10.1080/19345747.2019.1631421

Daniels, M., Wang, C., & Marcus, B. (2014). Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics*, *70*, 62–72. https://doi.org/10.1111/biom.12121

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control*. Springer New York. https://doi.org/10.1007/978-1-4614-0326-5

Drechsler, J., & Reiter, J. P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB establishment survey. *Journal of Official Statistics*, *25*, 589–603. https://www.scb.se/

Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, *105*, 1347–1357. https://doi.org/10.1198/jasa.2010.ap09480

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, *55*, 3232–3243. https://doi.org/10.1016/j.csda.2011.06.006

Duncan, G. T., Elliot, M., & Salazar, G. J. J. (2011). *Statistical confidentiality: Principles and practice*. Springer-Verlag. https://doi.org/10.1007/978-1-4419-7802-8

Duncan, G. T., & Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, *6*, 219–232. https://doi.org/10.1214/ss/1177011681

Erb, B., Bösch, C., Herbert, C., Kargl, F., & Montag, C. (2021). Emerging Privacy Issues in Times of Open Science. https://doi.org/10.31234/osf.io/u236e

Fleming, J. I., Wilson, S. E., Hart, S. A., Therrien, W. J., & Cook, B. G. (2021). Open accessibility in education research: Enhancing the credibility, equity, impact, and efficiency of research. *Educational Psychologist*, *56*, 110–121. https://doi.org/10.1080/00461520.2021.1897593

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, *9*, 383–406. https://www.scb.se/

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd). CRC press.

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2018). Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science*, *1*, 121–130. https://doi.org/10.1177/2515245917746500

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(108), 1–40. https://doi.org/10.1186/s12874-020-00977-1

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341-12), 1–6. https://doi.org/10.1126/scitranslmed.aaf5027

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2021). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/1745691620979806

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, *1*, 70–85. https://doi.org/10.1177/2515245917751886

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, *14*(28), 1–10. https://doi.org/10.1186/1471-2288-14-28

Jiang, B., Raftery, A. E., Steele, R. J., & Wang, N. (2021). Balancing inferential integrity and disclosure risk via model targeted masking and multiple imputation. *Journal of the American Statistical Association*. Advance online publication. https://doi.org/10.1080/01621459.2021.1909597

Joel, S., Eastwick, P. W., & Finkel, E. J. (2018). Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions. *Advances in Methods and Practices in Psychological Science*, *1*, 86–94. https://doi.org/10.1177/2515245917744281

Jones, B. C., & DeBruine, L. (2019). Sociosexuality and self-rated attractiveness. https://doi.org/10.17605/OSF.IO/6BK3W

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, *60*, 224–232. https://doi.org/10.1198/000313006X124640

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, *79*, 362–384. https://doi.org/10.1111/j.1751-5823.2011.00153.x

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, *62*, 85–96. https://doi.org/10.1111/j.1541-0420.2005.00396.x

Lindsay, D. S. (2017). Sharing data and materials in psychological science. *Psychological Science*, *28*(6), 699–702. https://doi.org/10.1177/0956797617704015

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*, 407–426. https://www.scb.se/

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods*, *25*, 157–181. https://doi.org/10.1037/met0000233

Manrique-Vallier, D., & Hu, J. (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*, 635–647. https://doi.org/10.1111/rssa.12352

Manrique-Vallier, D., & Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, *23*, 1061–1079. https://doi.org/10.1080/10618600.2013.844700

Martone, M. E., Garcia-Castro, A., & VandenBos, G. R. (2018). Data sharing in psychology. *American Psychologist*, *73*(2), 111–125. https://doi.org/10.1037/amp0000242

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*, 538–558. https://doi.org/10.1214/ss/1177010269

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*, 131–144. https://doi.org/10.1177/2515245917747656

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

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

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Almenberg, A. D., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L., Schönbrodt, F., & Vazire, S. (2021). Replicability, robustness, and reproducibility in psychological science. https://doi.org/10.31234/osf.io/ksfvq

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(1), 1–26. https://doi.org/10.18637/jss.v074.i11

Nowok, B., Raab, G. M., Dibben, C., Snoke, J., & van Lissa, C. (2020). *Synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control* (Version 1.6-0). https://CRAN.R-project.org/package=synthpop

Penke, L., & Asendorpf, J. B. (2008). Beyond global sociosexual orientations: A more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of Personality and Social Psychology*, *95*, 1113–1135. https://doi.org/10.1037/0022-3514.95.5.1113

Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., Shern, D., Pantin, H., Kaupert, S., Cano, N., Cruden, G., Bandiera, F., & Brown, C. H. (2013). Advancing science through collaborative data sharing and synthesis. *Perspectives on Psychological Science*, *8*(4), 433–444. https://doi.org/10.1177/1745691613491579

Plummer, M. (2017). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling* (Version 4.3.0). http://sourceforge.net/projects/mcmc-jags/

Quartagno, M., Grund, S., & Carpenter, J. (2019). Jomo: A flexible package for two-level joint modelling multiple imputation. *R Journal*, *11*(2), 205–228. https://doi.org/10.32614/RJ-2019-028

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, *9*(e53275), 1–12. https://doi.org/10.7554/eLife.53275

Raab, G. M., Nowok, B., & Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, *7*, 67–97. https://doi.org/10.29012/jpc.v7i3.407

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*, 85–96. http://www.statcan.gc.ca/

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

Reiter, J. P. (2005a). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*, 441–462. https://www.scb.se/

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, *29*, 181–188. http://www.statcan.gc.ca/

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, *30*, 235–242. http://www.statcan.gc.ca/

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

Reiter, J. P., & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, *28*, 583–590. https://www.scb.se/

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

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*(c221), 1–7. https://doi.org/10.1136/bmj.c221

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*, 461–468. https://www.scb.se/

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571. https://doi.org/10.1207/s15327906mbr3304_5

Schauer, J. M., Kuyper, A. M., Hedberg, E. C., Feinis, F., & Hedges, L. V. (2019). *Synthetic data disclosure control: Promise and feasibility for SLDS* [Unpublished Manuscript]. https://www.jmschauer.com

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC medical research methodology*, *12*(46), 1–13. https://doi.org/10.1186/1471-2288-12-46

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(3), 663–688. https://doi.org/10.1111/rssa.12358 _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12358

Stan Development Team. (2020). *RStan: The R interface to Stan* (Version 2.21.2).

https://CRAN.R-project.org/package=rstan

Stan Development Team. (2021). *Stan modeling language user's guide and reference manual (Version 2.26.0)*. http://mc-stan.org

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Towse, J. N., Ellis, D. A., & Towse, A. S. (2020). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01486-1

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(3). https://doi.org/10.1525/collabra.13

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40*, 73–76. https://doi.org/10.1016/j.intell.2012.01.004

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Zuiderwijk, A., Shinde, R., & Jeng, W. (2020). What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. *PLOS ONE*, *15*(9), e0239283. https://doi.org/10.1371/journal.pone.0239283

Appendix

Equivalence of DA-MI with a Single versus Multiple Copies

To show the equivalence of DA-MI with a single versus multiple copies, it needs to be shown that the two approaches imply the same predictive distributions for the variables in the synthesis model. For simplicity, we assume that the data include a continuous variable x and a covariate z, but the same holds for multivariate **x** and **z**. To generate synthetic values for x with DA-MI, the masked copies x^*_{li} $(l = 1, \ldots, k)$ are generated from:

$$x^*_{li} \sim N(x_i, \sigma^2_e) \ , \quad l = 1, \ldots, k \tag{A1}$$

with the same variance σ^2_e for all x^*_l. Suppose that DA-MI$_\text{P}$ is used to generate the synthetic data and that the synthesis model includes only linear effects. Then the synthetic values for x are simulated from the predictive distribution:

$$x^{(m)}_{syn,i} \sim N(\beta_0 + \sum_{l=1}^{k} \beta_{*l} x^*_{li} + \beta_z x_i, \sigma^2_r) \ . \tag{A2}$$

Because all x^*_{li} were generated with the same variance (σ^2_e), this is asymptotically equivalent to:

$$x^{(m)}_{syn,i} \sim N(\beta_0 + k\beta_* \bar{x}^*_i + \beta_z x_i, \sigma^2_r) \tag{A3}$$

where $\bar{x}^*_i = \sum_{l=1}^{k} x^*_{li}$ denotes the average values of the masked copies, and the same β_* applies to all x^*_l. Because the masked copies are normally distributed around x_i with variance σ^2_e, the average values $\bar{x}^*_i$ are also normally distributed around x_i with variance σ^2_e/k. Therefore, using DA-MI with k masked copies and (equal) noise variances of σ^2_e is asymptotically equivalent with DA-MI with a single and a noise variance of σ^2_e/k. For DA-MI$_\text{O}$, the same can be shown by observing that the masked copies x^*_{li} enter the likelihood—and thus the posterior predictive distribution—about x_i through the density implied by Equation A2 (the details are omitted here).