Article

Sociological Methods & Research
1-16

DOI: 10.1177/0049124118799376
journals.sagepub.com/home/smr


The Dance of the Mechanisms: How Observed Information Influences the Validity of Missingness Assumptions

Rianne Margaretha Schouten[1,2] and Gerko Vink[1]

Abstract

Missing data in scientific research go hand in hand with assumptions about the nature of the missingness. When dealing with missing values, a set of beliefs has to be formulated about the extent to which the observed data may also hold for the missing parts of the data. It is vital that the validity of these missingness assumptions is verified, tested, and that assumptions are adjusted when necessary. In this article, we demonstrate how observed data structures could a priori indicate whether it is likely that our beliefs about the missingness can be trusted. To this end, we simulate complete data and generate missing values according several types of MCAR, MAR, and MNAR mechanisms. We demonstrate that in scenarios where the data correlations are either low or very substantial, strictly different mechanisms yield equivalent statistical inferences. In addition, we show that the choice of quantity of scientific interest together with the distribution of the nonresponse govern the validity of the missingness assumptions.

[1] Department of Methodology and Statistics, University of Utrecht, Utrecht, the Netherlands
[2] Samen Veilig Midden-Nederland, Information Management Team, Utrecht, the Netherlands

Corresponding Author:
Rianne Margaretha Schouten, Department of Methodology and Statistics, University of Utrecht, Sjoerd Groenman Building, Padualaan 14, 3584 CH Utrecht, the Netherlands.
Email: r.m.schouten@uu.nl

Keywords

missing data methodology, missingness assumptions, multivariate amputation

Introduction

The analysis of incomplete data forms a ubiquitous problem in scientific research, especially because any existence of missing data calls for a set of assumptions. In general, it is challenging to solve missing data problems, and although techniques like multiple imputation (MI) (Rubin 1987; Little and Rubin 2002) have proven to be very effective and intuitive, assumptions about the nature of the missingness have to be validly formulated.

Missing data assumptions describe our beliefs about the extent to which the observed data may also hold for the missing parts of the data. Specifically, we separate the class of models where the observed data alone would be sufficient for obtaining valid inference (i.e., Missing Completely At Random or Missing At Random; Rubin 1976) from those models that rely on information that has not been captured in the observed data (i.e., Missing Not At Random; Rubin 1976). Unfortunately, we can never fully distinguish between MAR and MNAR mechanisms based on the observed data alone. After all, "Every missing not at random model has a missing at random counterpart with equal fit" (Molenberghs et al. 2008:371).

Nevertheless, it is vital that the validity of missingness assumptions is verified, tested, and that assumptions are adjusted when necessary. Generally, we investigate how sensitive our inference is in the light of the assumptions we make by means of sensitivity analysis (Molenberghs et al. 2015:317-490). With δ adjustment, for instance, a fixed value δ is added to the imputed values. By repeating the imputation procedure for different values of δ and by comparing the statistical results, it is possible to *a posteriori* examine the robustness of statistical inferences under different missing data models (Van Buuren 2012:182-87).

In this article, we demonstrate how observed data structures could a priori indicate whether it is likely that our beliefs about the missingness can be trusted, could be trusted, or should be trusted. Knowing the sensitivity of an inference before actually performing the statistical analysis could enhance the way we deal with missing data. Although we cannot escape from posing beliefs about the missing data, the observed data can give us an indication whether our beliefs should be believed. The remainder of this article has the following structure. First, we will provide a more technical discussion of missing data assumptions. We will then outline our simulation study in the

third section and present our findings in the fourth and fifth section. In the sixth section, we discuss implications of these findings.

Missingness Assumptions

Consider data matrix Y with y_{ij} either observed or missing. We collectively denote the observed data by Y_{obs} and the missing data by Y_{mis}. Further, let matrix R be a response indicator with $R_{ij} = 1$ if y_{ij} is missing and $R_{ij} = 0$ if y_{ij} is observed (Van Buuren 2012:30-35). According to Schafer and Graham (2002:150, 151),

> The distribution of R is best regarded as a mathematical device to describe the rates and patterns of missing values and to capture roughly possible relationships between the missingness and the values of the missing items themselves. To avoid suggestions of causality, we therefore refer to the probability distribution for R as *the distribution of missingness* or *the probabilities of missingness*.

In this context, Rubin (1976) distinguished three types of probability distributions: MCAR, MAR, and MNAR missingnes. With MCAR missingness, the observed data form a randomly obtained subset from the population. Thus,

$$Pr(R = 1|Y_{obs}, Y_{mis}, \psi) = Pr(R = 1|\psi)$$

with ψ some fixed parameters of the probability model (Van Buuren 2012:6). In other words, the missingness is solely induced by ψ and independent from both the observed and unobserved data. As a result, the response and nonresponse are exact representations of the true data model. This is different with a MAR mechanism, where

$$Pr(R = 1|Y_{obs}, Y_{mis}, \psi) = Pr(R = 1|\psi, Y_{obs})$$

indicating that the missing data model is governed by the observed data. Although in this situation the response and nonresponse represent different parts of the population, a proper conditioning of the incomplete variables on the observed data will still result in valid statistical inferences. Therefore, Rubin (1976) called these kinds of missingness *ignorable*. With MNAR missingness, the probability of being missing also depends on the unobserved information:

$$Pr(R = 1|Y_{obs}, Y_{mis}, \psi) = Pr(R = 1|\psi, Y_{obs}, Y_{mis}).$$

As a consequence, the missing data are *nonignorable*, meaning that the observed data alone are not sufficient to infer about the population (Rubin

1976). In such cases, we have to conclude that the response and nonresponse represent not only different but also unique parts of the true data.

Even though the observed and unobserved data sets may represent different and unique parts of the true data, the degree to which they are connected may be of great influence on the legitimacy of any missing data method. It is for this reason that many authors advocate the inclusion of predictor variables in the imputation procedure (e.g., Rubin, Stern, and Vehovar 1995; Schafer 1997; Van Buuren 2012). Important work from Collins, Schafer, and Kam (2001) showed that the inclusion of a variable that correlates either with the incomplete variable or with the missingness indeed improves parameter estimation. With our research, we intent to add to existing knowledge by (1) presenting the trends in the behavior of missingness mechanisms and (2) focusing on assumptions about the missingness distributions. As a consequence, we provide a context for situations where MCAR, MAR, and MNAR mechanisms may be much more or much less comparable than one would expect based on theory.

Design of the Experiment

To demonstrate the relationship between the missingness, the observed data, and the true data model, we perform a model-based simulation in R (R Core Team 2017). The design of the simulation is summarized in Algorithm 1 and basically consists of four steps: data set sampling with changing data correlations ($j = 1{:}9$), generating missing values with varying missingness proportions ($k = 1{:}3$) and missingness mechanisms ($i = 1{:}9$), MI, and evaluation of statistical parameters. Each combination of i, j, and k is replicated 1,000 times ($l = 1{:}1{,}000$), making the simulation a Monte Carlo type of simulation with $9 \cdot 9 \cdot 3 = 243$ simulation conditions. The next paragraphs will discuss each of the four steps in more detail.

Algorithm 1. Randomization procedure of the experiment.

$$\text{mechanisms} \leftarrow \left\{ \begin{array}{l} \text{mcar, marright, marleft, marmid, martail} \\ \text{mnarright, mnarleft, mnarmid, mnartail} \end{array} \right\}$$

$\text{correlations} \leftarrow \{0.1, 0.2, \ldots, 0.9\}$
$\text{proportions} \leftarrow \{0.1, 0.5, 0.9\}$
$\text{replications} \leftarrow 1000$
for $i = 1$ **to** 9 **do**
  mech $\leftarrow$ mechanisms $[i]$
  for $j = 1$ **to** 9 **do**
    cor $\leftarrow$ correlations $[j]$

```
    for k = 1 to 3 do
      prop ← proportions [k]
      for l = 1 to replications do
        sample complete dataset
        generate missing values
        impute missing values
        calculate evaluation measures
        temporarily save evaluation measures
      end for
      average evaluation measures
    end for
  end for
  save output per mechanism
end for
```

For the data sampling, we draw $N = 1{,}000$ cases from a bivariate normal distribution with `mvrnorm` in the package `MASS` (Ripley et al. 2017). We use mean vector

$$\mu = \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2 \end{bmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We vary the correlation between X and Y with ρ in $\{0.1,\ 0.2,\ \ldots,\ 0.9\}$.

We induce missingness (i.e., amputation) in variable Y and evaluate MCAR, MAR, and MNAR missingness mechanisms. We subdivide the MAR and MNAR mechanisms into those that consider the incomplete variable's left tail (LEFT), right tail (RIGHT), both tails (TAIL) or the distributional center (MID). When we induce MAR missingness, the probability for Y to be missing is a function of X. The four distribution functions (LEFT, RIGHT, MID, and TAIL) are demonstrated in Figure 1. In contrast, when we consider MNAR missingness, the true value of Y governs the probability for Y itself to be missing. Furthermore, we generate three kinds of missingness proportions: 0.1, 0.5, and 0.9. Note that these values indicate the sampled proportion of incomplete cases in Y. We leave X as an always observed covariate. For all conditions, the missing values are generated with multivariate amputation function `ampute` (Schouten, Lugtig, and Vink 2018).

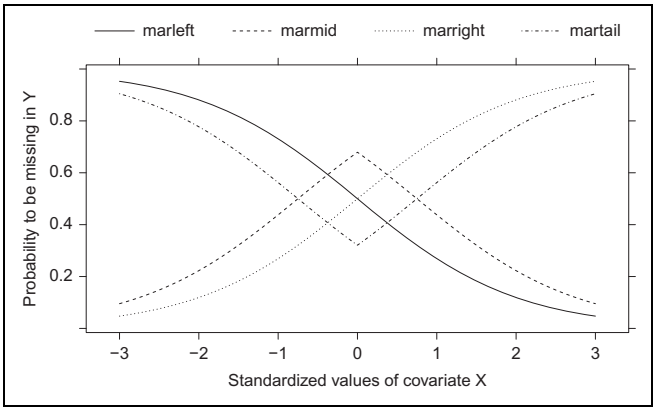

Figure 1. Four types of the logistic distribution function (Van Buuren 2012:64).

Every incomplete data set is imputed with `mice` (Van Buuren and Groothuis-Oudshoorn 2011) using Bayesian linear regression imputation as the imputation technique (`mice.impute.norm` with default settings). For every replication, we generate $m = 5$ multiply imputed data sets and combine the m completed data inferences into a single inference following Rubin's (1987) rules (pp. 76-77).

We focus our evaluation on three kinds of parameters: means, variances, and correlation. To investigate the first two, we use the bias, the coverage rate of the 95 percent confidence intervals and the width of that confidence interval of the expectation of Y (E_Y) with respect to a $N(5, 1)$ population. To evaluate the correlation coefficient ρ, we regress Y on X and use the bias, coverage rate and 95 percent confidence interval width (ciw) of regression coefficient β_X. We expect that in accordance with Neyman (1934), at least 95 percent of the confidence intervals should contain the true population value, although some room for simulation error should be taken into account to counteract the finite nature of our simulations. We compare the inferences obtained by MI to those obtained by complete case analysis (CCA).

Results

Our results display various situations where the theoretical distinctions between missingness assumptions do not appropriately describe the actual effect of a missing data problem on statistical inference making. We will first discuss the behavior of estimates of the mean, variance, and correlation under CCA and then repeat the process for MI.

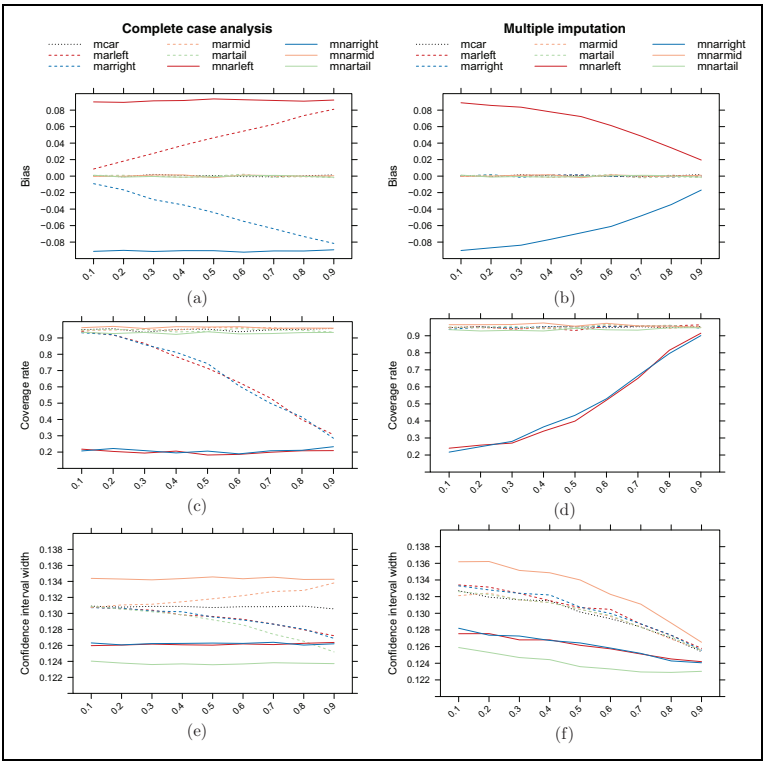

Figure 2. Coverage rate, average bias, and average confidence interval width of E_Y for complete case analysis and multiple imputation by Bayesian linear regression imputation. The *x*-axis displays the correlations between X and Y for all ρ in $\{0.1, 0.2, \ldots, 0.9\}$. Missingness proportion is 0.1.

CCA

Mean. Our simulation results show that different missingness mechanisms yield similar statistical inferences. For instance, Figure 2a shows that for all MID and TAIL types of missingness, the estimates of E_Y are unbiased. Coverage rates are about 90 percent (Figure 2c). Because these specific types of MAR and MNAR missingness preserve the symmetry of the true data distribution of Y, we find that an estimate of the mean yields equal statistical inferences under different missing data mechanisms. In other words, although the underlying relation between the response and nonresponse may take different forms, statistical inferences can still be similar.

When the missing data affect the symmetry of the true data distribution, we will obtain biased estimates of E_Y. Figure 2a and c shows that this is the situation for all RIGHT and LEFT types of missingness. With these missingness mechanisms, one side of Y's distribution obtains more missing values than the other side, thereby inducing a skewness in the observed distribution. As a result, we obtain biased estimates of E_Y (Figure 2a) and the coverage rates of the 95 percent confidence interval drop far below 95 percent (Figure 2c).

The size of the true data correlation determines whether statistical inferences are similar despite of different underlying missingness mechanisms. Figure 2a and c shows that when the correlation between variables X and Y is low, bias in the estimates of E_Y approach zero under MARRIGHT and MARLEFT missingness and coverage rates approach 95 percent (Figure 2a and c). Clearly, when the correlation between X and Y vanishes, the incomplete data form a random subset from the population even though the underlying mechanism is MAR. As a consequence, it is not necessary to assume that the complete and incomplete data describe different parts of the true data model (MAR). Rather, we can assume MCAR missingness.

Variance. The observed distribution of Y determines the precision of an estimate of E_Y. We obtain insight into the variance of the nonresponse distribution by evaluating the ciw. Figure 1 displays that especially MID and TAIL types of missingness affect the variance of the observed distribution of Y. For instance, MID missingness increases the width of the confidence interval. This finding is not surprising because MID missingness reduces the number of data points in the center of the distribution and this results in a high variance. TAIL missingness does the opposite. Remark that significance tests use these precision measures. As a result, irrespective of MAR or MNAR missingness, the assumption that the nonresponse has an MID or TAIL type of distribution will have consequences for finding significant statistical results.

Correlation. With CCA, estimates of correlation coefficient ρ are unaffected by the theoretical differences between MAR and MCAR missingness mechanisms. Figure 3a shows unbiased estimates of β_X for all MAR and MCAR missingness mechanisms. Additionally, the coverage rates are all approximately 95 percent (Figure 3c). A simplified graphical depiction of these findings can be found in Figure 4. Because the missing values are in variable Y, the data structure (i.e., the regression coefficient) is not affected

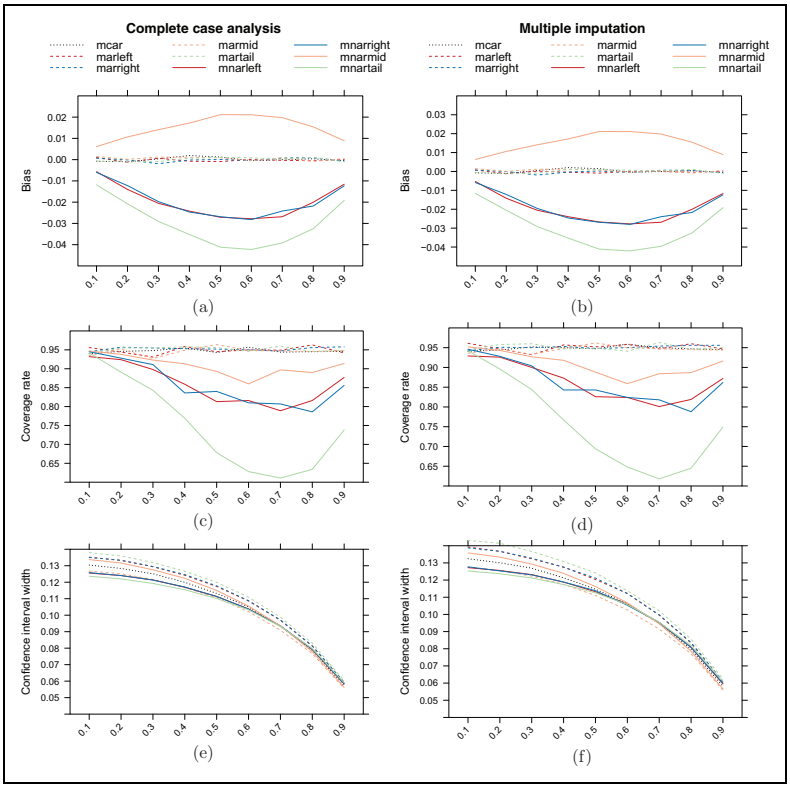

Figure 3. Coverage rate, average bias, and average confidence interval width of β_X for complete case analysis and multiple imputation by Bayesian linear regression imputation. The *x*-axis displays the correlations between *X* and *Y* for all ρ in $\{0.1, 0.2, \ldots, 0.9\}$. Missingness proportion is 0.1.

by MAR (and MCAR) missingness. Note that when X would be the incomplete variable, MNAR mechanisms would result in unbiased estimates of β_X.

With low and high true data correlations, CCA estimates of β_X under MNAR approach the values of the estimates obtained under MCAR and MAR missingness. This is visible in Figure 3a by the bias decreasing toward zero for lower and higher values of ρ. A comparable trend is visible for the coverage rates (Figure 3c). In other words, with CCA, MNAR, and MAR mechanisms yield comparable statistical estimates of β_X when data correlations become extreme.

The type of missingness affects the magnitude of possible bias and undercoverage of β_X estimates. The largest bias for MNARMID is 0.021 (when

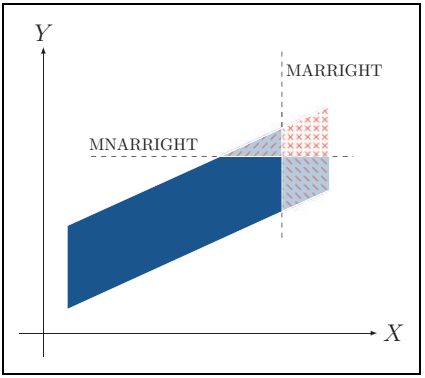

Figure 4. Schematic depiction of the relation between variables *Y* and *X* with missing values in *Y*. It is apparent that in this scenario, MARRIGHT will not influence the β coefficient. MNARRIGHT, on the other hand, has a direct influence on β_X. Note that when the missingness is in *X*, the findings will be exactly opposite.

$\rho = 0.6$), while it is -0.028 for RIGHT and LEFT missingness and -0.042 for TAIL missingness (all outcome values can be found in the Online Appendices). Clearly, the occurence of missing values at the tails of *Y*'s distribution affects the relation between *Y* and *X* more than missingness in the center of *Y*'s distribution.

MI

Mean. As expected, under MAR statistical inferences of E_Y are unbiased with MI (Figure 2b). When we use MI under the assumption of MAR missingness, we expect that the information about the missing values in variable *Y* is present in the observed data variable *X*. As such, we anticipate a relation between the response and nonresponse. MI uses this relation to obtain unbiased estimates of E_Y.

However, Figure 2b and d shows that in the situation of a theoretical MNAR mechanism, a high correlation enables MI to obtain valid statistical estimates of E_Y too. For instance, Figure 2b displays how MI reduces the bias from -0.090 to -0.017 for RIGHT-tailed missingness. In addition, MI increases the coverage rate from 0.217 to 0.901 (Figure 2d). Clearly, as data correlations increase, MI is able to use the information in *X* to create imputations in *Y*, even though the missingness generating mechanism is MNAR. In this situation, the assumption that the observed data alone are not sufficient to obtain valid statistical inferences is not conclusive. Rather, a high data correlation can allow for a legitimate use of the MAR assumption.

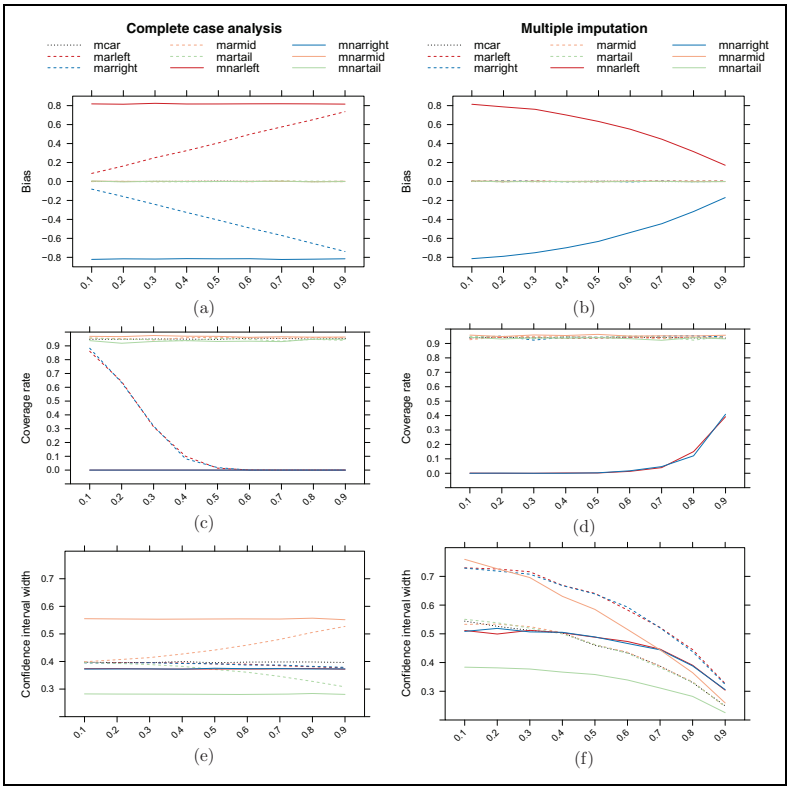

Figure 5. Coverage rate, average bias, and average confidence interval width of E_Y for complete case analysis and multiple imputation by Bayesian linear regression imputation. The *x*-axis displays the correlations between X and Y for all ρ in $\{0.1, 0.2, \ldots, 0.9\}$. Missingness proportion is 0.9.

Variance. In addition, high data correlations enable MI to thoroughly use the observed information to obtain not only valid but also efficient statistical estimates. At first, we see that with 10 percent missingness, MI increases the ciw from 0.131 (under CCA) to 0.133. This is due to between-imputation variance. However, after performing MI, a higher data correlation decreases the ciw toward 0.125 (Figure 2f). With 90 percent missingness, this effect is more pronounced with the ciw being reduced from 0.544 to 0.249 (Figure 5f).

Correlation. For estimates of β_X, using MI results in bias and coverage rates comparable with CCA. Surely, statistical inferences were already unbiased

under MAR (Figure 3a) and MI rather adds between-imputation variance. The between-imputation variance increases the ciws, especially when true data correlations are low (compare Figure 3e with Figure 3f).

When data correlations increase, MI is not able to decrease the bias or increase the coverage rates of β_X under MNAR missingness. This effect is opposite to what we saw earlier for estimates of E_Y. For estimates of β_X, on the other hand, the outcome values in Figure 3a and c are comparable to those in Figure 3b and d. Estimates of β_X remain biased for all MNAR mechanisms, even when there is a strong true data relation between X and Y. Therefore, in this situation, the assumption that the observed data alone are not sufficient to obtain valid statistical inferences is justified. In other words, the information about the missing data is truly missing from the data.

Proportion

We present the results for estimates of E_Y with a missing data proportion of 0.9 in Figure 5. Simulation results for E_Y with 50 percent missing data, and for β_X with 90 percent and 50 percent missingness are presented in the Online Appendices in Figures 7, 8, and 9, respectively. In sum, all the trends we discussed so far are present with higher missingness proportions. Thus, whether or not our missingness assumptions are valid is in essence not determined by the missingness proportion.

However, an increased missingness proportion does have an effect on the size of possible bias, coverage, and variance measures. Although the trends in Figure 5 are comparable with Figure 2, a missingness proportion of 0.9 increases the absolute values. For instance, the observed bias after CCA increases with factor 10 (Figure 5a) and the ciw's with factor 30 (Figure 5e). In addition, for RIGHT and LEFT missingness, the coverage rates drop way faster when more values are missing (Figure 5c) and even when the data correlation is 0.9, MI is no longer able to improve the coverage rates all the way to 95 percent (Figure 5d).

Key Findings

1. *Mean*: With low true data correlations, CCA estimates of E_Y are comparable between MCAR and MAR mechanisms. This is graphically depicted in Figure 6a. With increasing data correlations, MI uses the information in observed variable X to obtain unbiased

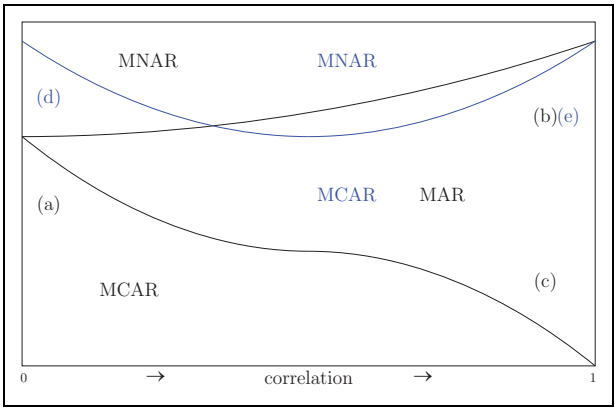

Figure 6. Graphical representation of the impact of data correlations on missingness assumptions. Black and blue lines illustrate estimates of E_Y and β_X, respectively. (a) Comparable complete case analysis estimates for MAR and MCAR missingness, (b) comparable multiple imputation estimates for MNAR and MAR missingness, (c) comparable variance with MI for MCAR and MAR missingness, (d) comparable CCA and MI estimates for MNAR and MAR missingness, and (e) same as (d).

estimates of E_Y. Even under MNAR missingness, MI uses the information in X to decrease the bias in estimates of E_Y (see Figure 6b).

2. *Variance*: Regardless of whether the underlying mechanism is MAR or MNAR, MID, and TAIL, missingness influences the variance of the observed data distribution. High data correlations enable MI to reduce the variance in estimates of E_Y, even for MCAR missingness. This is depicted in Figure 6c.
3. *Correlation*: The estimated correlation coefficient is not affected by MAR and MCAR missingness. With low and high true data correlations, CCA estimates of β_X under MNAR approach the values of the estimates obtained under MCAR and MAR (Figure 6d and e.) Despite of data correlations, MI does not affect the estimates of β_X.

Discussion

In practice, applied researchers often conveniently assume MAR missingness and proceed with incomplete data methods. We demonstrate that simply assuming a mechanism in order to warrant an analysis may be severely limiting to the statistical inference that can be obtained. It is

vital to validate the likelihood of the assumed mechanism, given the observed data structures. For example, when the variables of a data set are barely correlated, MAR and MCAR missingness may become indistinguishable. In such cases, assuming MAR to allow for the use of MI would limit statistical power. MI would merely increase the variance without the need for decreasing bias.

The reverse is also true: Assuming MNAR missingness on highly correlated data may be unnecessary as the essence of the missing data may be sufficiently covered by the observed information. The response and nonresponse are connected at the population level. Our findings confirm that with increased data relations, this link improves the performance of missing data methods. As a result, an applied researcher's hope for obtaining valid MAR-based inference on MNAR data can be justified when their variables are highly correlated. It is needless to say we can "force" such a situation by including the right variables.

A comprehensive insight into the nature of missingness mechanisms is also important for evaluations of missing data methodology. Although it may seem obvious that any simulation study evaluating the performance of a missing data method should generate a legitimate (i.e., as intended) missing data problem, our results identify situations where a theoretical appropriate mechanism can have an unexpected, not so large impact on the data. When performing simulation studies, we strongly recommend researchers to perform CCA to investigate the effects of a generated missing data problem. When CCA returns biases close to zero, coverage rates close to 95 percent or very large ciws, it is wise to revisit the simulation conditions. After all, "A missing value treatment cannot be properly evaluated apart from the modeling, estimation or testing procedure in which it is embedded" (Schafer and Graham 2002:149).

We believe that our missingness assumptions are not merely a subjective nuisance that dictates whether or not we may proceed with the inferential statistics we value so dearly. Instead, we advocate the approach that our assumptions are an elegant, interwoven part of the system that gives us our inference. We encourage researchers to obtain deeper insights into the (overlapping) behavior of MAR and MNAR mechanisms. We expect that our findings generalize to larger data structures but to investigate whether the same phenomena can be observed for data sets with more variables or higher dimensions, an extension of our simulation to other settings would be necessary. Ultimately, the goal is to obtain valid statistical inferences and the set of beliefs we formulate are the only means to bridge the gap between what is observed and what is missing.

Acknowledgments

The authors gratefully acknowledge Andrew Gelman for facilitating a variety of interesting research projects. This article is the outcome of an inspiring and pleasant visit to a beautiful city and a great team of researchers. The authors gratefully acknowledge Stef Van Buuren for reviewing a previous draft of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6:330-51.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: John Wiley.

Molenberghs, Geert, Caroline Beunckens, Cristina Sotto, and Michael G. Kenward. 2008. "Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit." *Journal of the Royal Statistical Society: Series B* 70: 371-88.

Molenberghs, Geert, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke. 2015. *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC Press.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97:557-625.

R Development Core Team. 2008. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org.

Ripley, Brian, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, and David Firth. 2017. *R-Package MASS*. Retrieved September 25, 2018 (https://Cran.r-project.org/web/packages/MASS/index.html).

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63:581-90.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, Donald B., Hal S. Stern, and Vasja Vehovar. 1995. "Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite." *Journal of the American Statistical Association* 90:822-28.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London, England: Chapman & Hall.

Schafer, Joseph L. and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7:147-77.

Schouten, Rianne Margaretha, Peter J. Lugtig, and Gerko Vink. 2018. "Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure." *Journal of Statistical Computation and Simulation* 88:2909-30.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC Press.

Van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45.

Author Biographies

Rianne Margaretha Schouten works as an external PhD candidate in the field of missing data methodology. She started her research with the development and implementation of a multivariate amputation procedure (i.e. a way to generate sophisticated missingness in complete datasets) and is now focusing on the evaluation of missing data methods. Rianne applies her statistical and programming knowledge as Developer Data & Analytics at Samen Veilig Midden-Nederland, a Dutch youth care organization.

Gerko Vink is a statistical scientist with a passion for educating people. He aims to be at the cutting edge of both teaching and research and has an interest in new developments concerning the presentation of data, results and knowledge. Gerko has a specific interest for incomplete data problems, optimisation and programming. He is based in the Netherlands and works as an assistant professor at Utrecht University.