On Testability and Goodness of Fit Tests in Missing Data Models (Supplementary Material)

Razieh Nabi[1] **Rohit Bhattacharya**[2]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA
[2]Department of Computer Science, Williams College, Williamstown, Massachusetts, USA

In Appendix A, we cover additional preliminaries: (i) we present the odds-ratio parameterization of a missing data process and demonstrate the estimation of an odds ratio through a straightforward example, (ii) we elaborate more on parameter counting in discrete models to assess whether the assumptions in a full law impose restrictions on observed data law, and (iii) we provide additional details on substantive edge distinctions between $\{V_i^*, R_j\}$ vs $\{V_i, R_j\}$ in the permutation model. Appendix B contains additional discussions on the goodness-of-fit tests in the sequential MNAR model using likelihood approaches. It also includes an automated algorithm for performing a sequential goodness-of-fit tests based on weighted likelihood-ratios. Appendix C contains additional discussions on the use of odds-ratio parameterization in the sequential MAR and sequential MNAR models, as well as a formalization of the goodness-of-fit tests in block-parallel MNAR models based on odds ratio calculations. Appendix D contains the proofs. Appendix E contains simulation details and additional empirical analyses.

A PRELIMINARIES

A.1 ODDS-RATIO PARAMETERIZATION

The odds-ratio parameterization of joint distributions $p(R|X)$ was introduced in Chen [2007]. Assuming we have K missingness indicators, $p(R \mid X)$ can be expressed as follows:

$$p(R \mid X) = \frac{1}{Z} \times \prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, X) \times \prod_{k=2}^{K} \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X), \tag{1}$$

where $R_{-k} = R \setminus R_k, R_{\prec k} = \{R_1, \ldots, R_{k-1}\}, R_{\succ k} = \{R_{k+1}, \ldots, R_K\}$, and

$$\text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X) = \frac{p(R_k \mid R_{\succ k} = 1, R_{\prec k}, X)}{p(R_k = 1 \mid R_{\succ k} = 1, R_{\prec k}, X)} \times \frac{p(R_k = 1 \mid R_{-k} = 1, X)}{p(R_k \mid R_{-k} = 1, X)}.$$

Z in Eq. (1) is the normalizing term and is equal to $\sum_r \Big\{ \prod_{k=1}^{K} p(r_k \mid R_{-k} = 1, X) \times \prod_{k=2}^{K} \text{OR}(r_k, r_{\prec k} \mid R_{\succ k} = 1, X) \Big\}$.

Estimating equations for computing odds ratios.

Consider the no self-censoring model with two variables, shown in Fig. 4(b). Let $\theta(r_1, r_2) = \text{OR}(R_1 = r_1, R_2 = r_2 \mid X_1, X_2)$. We can estimate $\theta(r_1 = 0, r_2 = 0)$ with the following unbiased estimating equation where an odds-ratio parameterization of $p(R|X)$ is used in place. We have:

$$p(R_1 = r_1, R_2 = r_2 \mid X) = \frac{1}{Z} \times p(R_1 = r_1 | R_2 = 1, X_2) \times p(R_2 = r_2 | R_1 = 1, X_1) \times \theta(r_1, r_2).$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Therefore,

$$\begin{aligned}
&\mathbb{P}_n\Big[R_1R_2 \times \frac{p(R_1=0,R_2=0 \mid X)}{p(R_1=1,R_2=1 \mid X)} - (1-R_1)(1-R_2)\Big] \\
&\quad = \mathbb{P}_n\Big[R_1R_2 \times \frac{p(R_1=0|R_2=1,X_2) \times p(R_2=0|R_1=1,X_1) \times \theta(R_1=0,R_2=0)}{p(R_1=1|R_2=1,X_2) \times p(R_2=1|R_1=1,X_1) \times \theta(R_1=1,R_2=1)} - (1-R_1)(1-R_2)\Big] \\
&\quad = \mathbb{P}_n\Big[R_1R_2 \times \frac{p(R_1=0|R_2=1,X_2) \times p(R_2=0|R_1=1,X_1)}{p(R_1=1|R_2=1,X_2) \times p(R_2=1|R_1=1,X_1)} \times \theta(R_1=0,R_2=0) - (1-R_1)(1-R_2)\Big] \\
&\quad = 0.
\end{aligned}$$

The first equality holds by definition, the second equality holds because $\text{OR}(R_1 = 1, R_2 = 1) = 1$, and the third equality can be simply proved with tower laws of expectations. Given the above, we can find a closed form estimator for $\theta(R_1 = 0, R_2 = 0)$:

$$\theta(R_1=0,R_2=0) = \frac{\mathbb{P}_n\Big[(1-R_1) \times (1-R_2)\Big]}{\mathbb{P}_n\Big[R_1 \times R_2 \times \frac{p(R_1=0|R_2=1,X_2) \times p(R_2=0|R_1=1,X_1)}{p(R_1=1|R_2=1,X_2) \times p(R_2=1|R_1=1,X_1)}\Big]}.$$

For $K > 2$, we need to compute odds ratio terms of the form $\theta(R_k = 0, R_j = 0) \coloneqq \text{OR}(R_k = 0, R_j = 0|R_{-kj} = 1, X)$. The following unbiased estimating equation that incorporates R_{-kj} can be used to estimate $\theta(R_k = 0, R_j = 0)$:

$$\mathbb{P}_n\Big[\prod_{i=1}^{K} R_i \times \frac{p(R_k=0|R_{-k}=1,X_{-k}) \times p(R_j=0|R_{-j}=1,X_{-j})}{p(R_k=1|R_{-k}=1,X_{-k}) \times p(R_j=1|R_{-j}=1,X_{-j})} \times \theta(R_k=0,R_j=0) - \prod_{i \neq \{j,k\}} R_i(1-R_k)(1-R_j)\Big] = 0.$$

Using the tower laws of expectations, it is easy to show why the above estimating equation holds.

A.2 PARAMETER COUNTING ARGUMENT

How does one know that a missing data DAG imposes restrictions that are testable from the observed data distribution? When all substantive variables take on values in a finite discrete state space, one simple check is to compare the number of parameters in the full law using the DAG factorization in (1) and the saturated observed data law using the *pattern-mixture* factorization [Rubin, 1976]. The pattern-mixture factorization is given by the marginal distribution of R and the conditional distribution of X^* given R. If a missing data DAG with an identified full law can be described with fewer parameters than the saturated pattern-mixture model, we may conclude that the restrictions on full law impose constraints on the observed data distribution. Shpitser [2016] has used parameter counting to give an intuition for why the no self-censoring model is identified. Nabi et al. [2020] also have relied on a parameter counting argument to prove the completeness of their results for full law identification in missing data DAG models.

As an example, consider a missing data model with two substantive binary variables X_1 and X_2. Assume the full law satisfies the assumptions of the permutation model in (2), which are $R_1 \perp\!\!\!\perp X_1|X_2$ and $R_2 \perp\!\!\!\perp X_1, X_2 \mid R_1, X_1^*$. The full law then factorizes as $p(X_1, X_2) \times p(R_1|X_2) \times p(R_2|R_1, X_1^*)$. We need 3 parameters for parameterizing $p(X_1, X_2)$, 2 parameters for $p(R_1 \mid X_2)$, and 3 parameters for $p(R_2|R_1, X_1^*)$; thus a total of 8 parameters. (We excluded the deterministic terms $p(X_1^*|R_1, X_1)$ and $p(X_2^*|R_2, X_2)$ as they do not add any parameters.) On the other hand, the pattern-mixture factorization of the observed data law $p(R, X^*)$ can be written as $p(R_1, R_2) \times p(X_1^*, X_2^*|R_1, R_2)$. Since R_1 and R_2 are binary, it requires at most 3 parameters to parameterize $p(R_1, R_2)$. Using chain rule factorization, we have $p(X^*|R) = p(X_1^*|R_1, R_2) \times p(X_2^*|R_1, R_2, X_1^*)$. Due to the deterministic relations, if $R_1 = 0$ then $X_1^* =$ "?", thus we need at most 2 parameters to parameterize $p(X_1^*|R_1, R_2)$. Similarly, we need at most 3 parameters to parameterize $p(X_2^*|R_1, R_2, X_1^*)$. In total, 8 parameters are required to encode a saturated observed data law. As expected, the number of parameters in the full law of the permutation model (which is proven to be identified as a function of observed data) and the saturated observed data law are the same, reaffirming the fact that permutation model is saturated and places no restrictions on the observed data distribution.

As another example of a saturated model, consider the no self-censoring model in Fig. 4(b). The odds-ratio parameterization

of the missingness mechanism $p(R|X)$ is as follows:

$$
\begin{aligned}
&p(R_1 = r_1, R_2 = r_2 \mid X_1, X_2) \qquad (2)\\
&\quad = \frac{1}{Z} \times p(R_1 = r_1 \mid R_2 = 1, X_1, X_2) \times p(R_2 = r_2 \mid R_1 = 1, X_1, X_2) \times \text{OR}(R_1 = r_1, R_2 = r_2 \mid X_1, X_2)\\
&\quad = \frac{1}{Z} \times p(R_1 = r_1 \mid R_2 = 1, X_2) \times p(R_2 = r_2 \mid R_1 = 1, X_1) \times f(R_1 = r_1, R_2 = r_2),
\end{aligned}
$$

where $Z = \sum_{r_1, r_2} p(R_1 = r_1 \mid R_2 = 1, X_2) \times p(R_2 = r_2 \mid R_1 = 1, X_1,) \times \text{OR}(R_1 = r_1, R_2 = r_2 | X_1, X_2)$. The second equality in (2) holds because $R_1 \perp\!\!\!\perp X_1 | R_2, X_2$ and $R_2 \perp\!\!\!\perp X_2 | R_1, X_1$. Further, $\text{OR}(R_1 = r_1, R_2 = r_2 \mid X_1, X_2)$ is just a function of R_1 and R_2 because:

$$
\begin{aligned}
\text{OR}(R_1 = r_1, R_2 = r_2 \mid X_1, X_2) &= \frac{p(R_1 = r_1 \mid R_2 = r_2, X_2)}{p(R_1 = 1 \mid R_2 = r_2, X_2)} \times \frac{p(R_1 = 1 \mid R_2 = 1, X_2)}{p(R_1 = r_1 \mid R_2 = 1, X_2)}\\
&= \frac{p(R_2 = r_2 \mid R_1 = r_1, X_1)}{p(R_2 = 1 \mid R_1 = r_1, X_1)} \times \frac{p(R_2 = 1 \mid R_1 = 1, X_1)}{p(R_2 = r_2 \mid R_1 = 1, X_1)}\\
&= f(R_1, R_2).
\end{aligned}
$$

The first equality holds because $R_1 \perp\!\!\!\perp X_1 \mid R_2, X_2$, the second equality holds because $R_2 \perp\!\!\!\perp X_2 \mid R_1, X_1$, and together they imply the last equality which means $\text{OR}(R_1, R_2 \mid X_1, X_2)$ is a function of R_1, R_2 (all observed data). In the above argument, we have used the fact that odds ratios is symmetric (i.e., $\text{OR}(A, B|Z) = \text{OR}(B, A|Z)$). Assuming X_1 and X_2 are binary, the full law in a no self-censoring model would have 8 parameters (same number as in a saturated observed data law). Those parameters are as follows: 3 parameters for $p(X_1, X_2)$, 1 parameter for $\text{OR}(R_1 = 0, R_2 = 0 | X_1, X_2) = f(R_1, R_2)$ (since the OR evaluated at other levels of R_1 and R_2, i.e., the reference values, is always one), 2 parameters for $p(R_1 = 1 | R_2 = 1, X_2)$, and 2 parameters for $p(R_2 = 1 | R_1 = 1, X_1.)$

Examples of the three class of missing data models that we are interested in are provided in Fig. 1(a), 4(a), and 4(d), where $X = \{X_1, X_2\}$. Here, we compare the full law parameterization of each example against the pattern-mixture parameterization as an illustrative step to show that the conditional independence restrictions on the full law impose restrictions on the observed data law. Given the MAR model in Fig. 1(a), the full law factorizes as $p(X_1, X_2) \times p(R_1) \times p(R_2 \mid R_1, X_1^*)$. Given the MNAR model in Fig. 4(a) (without the dashed edge), the full law factorizes as $p(X_1, X_2) \times p(R_1 \mid X_2) \times p(R_2 \mid R_1)$. Given the MNAR model in Fig 4(d), the full law factorizes as $p(X_1, X_2) \times p(R_1 \mid X_2) \times p(R_2 \mid X_1)$. In all the three examples, the full law requires 7 parameters to encode the independencies (less than the number of parameters in the saturated observed data law). The above implies that there must be a testable implication, at least in the binary case, on the observed data laws of the three classes of missing data models that we consider. The parameter counting argument can be simply generalized to discrete data. Results in the main draft confirm that this generalizes to situations where no distributional assumptions are made.

A.3 ON EDGES FROM PROXY VARIABLES TO MISSINGNESS INDICATORS

The convention in previous work on missing data DAGs (e.g., Mohan et al. [2013] and Mohan and Pearl [2021]) has often been to avoid including edges from proxy variables to missingness indicators. However, allowing for $X_i^* \rightarrow R_j$ edges enables exploration of a broader class of missing data DAG models and MNAR mechanisms. For instance, the permutation MNAR model introduced by Robins [1997] can only be represented graphically if we permit proxy variables to point to missingness indicators. Without such edges, this model would lack a graphical characterization. A more comprehensive discussion on this topic can be found in [Nabi et al., 2022]. Models like the permutation model are particularly interesting as they represent nonparametrically saturated models with nonparametrically identified full laws. Thus, incorporating these edges allows our work to have a broader scope and naturally builds upon the foundations laid out in earlier research on testability in missing data DAGs, including the framework proposed by Mohan and Pearl [2014].

Here, we explore the substantive distinctions between models with edges $X_i^* \rightarrow R_j$ (as in the permutation model) and models with edges $X_i \rightarrow R_j$. To illustrate the dissimilarities between these two models, let us assume that X_i is a binary variable, and we consider two structures: (1) $R_i \rightarrow R_j \leftarrow X_i$ and (2) $R_i \rightarrow R_j \leftarrow X_i^*$. In the first structure, $p(R_j = 1 \mid R_i, X_i)$ has four parameters, with each parameter corresponding to a specific combination of values for X_i and R_i. On the other hand, in the second structure, $p(R_j = 1 \mid R_i, X_i^*)$ only has three parameters due to the deterministic relationship between R_i and X_i^*. These structural differences indicate qualitative differences as well. An $X_i \rightarrow R_j$ edge implies that the missing variable X_i might have an impact on R_j. Conversely, an $X_i^* \rightarrow R_j$ edge suggests that the variable affects R_j when it is observed,

but when it is missing, its absence influences future missingness rather than its actual unobserved value. These differences have implications for identification. If we change the edges in Fig. 3(a) to be $X_i \to R_j$, neither the full law nor the target law is identifiable. However, if we retain the edges as they are, the models are identifiable, as they represent the permutation model. Identifiability also plays a crucial role in determining testability, as discussed in the main manuscript.

Finally we note that testing the absence of dashed edges involving proxy variables in Fig 3(a) is not entirely equivalent to testing edges involving their counterfactual counterparts. In other words, if for instance $R_2 \perp\!\!\!\perp X_1 | R_1 = 1$ or equivalently $R_2 \perp\!\!\!\perp X_1^* | R_1 = 1$ holds in the observed data, there is no guarantee that R_2 and counterfactual X_1 are independent in the full law; because for the the independence in the full law to hold, we must show that $R_2 \perp\!\!\!\perp X_1$ even among rows where $R_1 = 0$. This may be possible under a further assumption like *faithful observability* used by Tu et al. [2019] (which is a stronger assumption than standard faithfulness) where independences in the observed data "do not lie" about independences in the full data. But in the case where the full/target law is not identified, an assumption like this could be misleading – in this case $p(R_2 | R_1 = 0, X_1)$ is not identified and there is no way to confirm the validity of the test in the full data law. However, this was not a particular issue for the method proposed in [Tu et al., 2019], as they consider a subclass of MNAR models where the full law is always identified. In future research, it would be interesting to explore the additional constraints imposed by assumptions like faithful observability, which may lead to $X_i \to R_j$ edges resembling edges from a proxy variable rather than the actual underlying counterfactual.

B MORE ON GOODNESS-OF-FIT TESTS IN THE SEQUENTIAL MNAR MODEL

B.1 GENERAL ALGORITHM FOR GOODNESS-OF-FIT TESTS USING LIKELIHOOD APPROACHES

Algorithm 1 illustrates how to perform a sequential goodness-of-fit tests based on weighted likelihood-ratios for K greater than 3 variable in sequential MNAR models.

Algorithm 1 TESTING SEQUENTIAL MNAR $(\prec, \mathcal{M}, \mathcal{D}_n)$

Let $\prec$ index variables by $k = 1, \ldots, K$.
Let $\Omega_{K+1} = 1$.
for $k \in \{K, \ldots, 2\}$ **do**
    Let $W_k(\beta_k^o) \coloneqq p(R_k | R_{\prec k}, X_{\succ k}; \beta_k^o)$ and
    $W_k(\beta_k^a) \coloneqq p(R_k | R_{\prec k}, X_{\succ k}, X^*_{\prec k}; \beta_k^a)$.
    Estimate β_k^o and β_k^a via the following:

$$\mathbb{P}_n\big[\Omega_{k+1} \times U(\beta_k^o)\big] = 0, \quad \mathbb{P}_n\big[\Omega_{k+1} \times U(\beta_k^a)\big] = 0,$$

    where $\mathbb{P}_n\big[U(\beta_k^o)\big] = 0$ and $\mathbb{P}_n\big[U(\beta_k^a)\big] = 0$ are estimating equations for β_k^o and β_k^a wrt the full law.
    Compute a weighted likelihood-ratio as follows:

$$\rho = n\mathbb{P}_n\left[\Omega_{k+1} \times \log\Big(\frac{W_k(\widehat{\beta}_k^a)}{W_k(\widehat{\beta}_k^o)}\Big)\right].$$

    Test ρ with α significance level.
    if $\mathcal{M}_o$ is rejected (i.e., $R_k \not\perp\!\!\!\perp X^*_{\prec k} | R_{\prec k}, X_{\succ k}$) **then**
        return not sequential MNAR
    else $\Omega_{k+1} = \frac{\mathbb{I}(R_{\succ k} = 1)}{\prod_{j \succ k}^K W_j(\widehat{\beta}_j^o)}$.
return sequential MNAR

B.2 ALTERNATIVE SUPERMODELS IN THE SEQUENTIAL MNAR MODEL

Consider the m-DAG in Fig. 4(a). We are interested in the absence of an edge between X_1 and R_2 which implies $R_2 \perp\!\!\!\perp X_1 | R_1$. The no self-censoring supermodel is drawn in Fig. 4(b) (with R_1, R_2 edge undirected). We can evaluate this independence by showing $p(R_2 | R_1, X_1)$ is not a function of X_1. See Appendix B.2 for details on how to set up such a test.

For this, we use the following odds-ratio factorization of $p(R|X)$ [Chen, 2007]:

$$p(R \mid X) = \frac{1}{Z(X)} \times p(R_1 \mid R_2 = 1, X_2) \times p(R_2 \mid R_1 = 1, X_1) \times \text{OR}(R_1, R_2|X), \tag{3}$$

where $Z(X)$ is a normalizing term and $\text{OR}(R_1, R_2|X)$ is the conditional odds ratio between R_1 and R_2. Since the no self-censoring model is identified, each piece above must be a function of observed data. This is trivial for the univariate conditionals, however, it can also be shown that $\text{OR}(R_1, R_2|X) = f(R_1, R_2)$, i.e., is not a function of X (see Appendix A, Eq. 2.) By definition $p(R_2|R_1, X_1) = p(R|X)/\sum_{R_2} p(R|X)$; to show $p(R_2|R_1, X_1)$ is not a function of X_1, it suffices to show $p(R|X)$ is not a function of X_1 which using (3) only requires us to show $p(R_2|R_1 = 1, X_1)$ is not a function of X_1 which is easy to evaluate. This can be generalized to $K > 2$, but it involves higher order interactions terms in the odds-ratio parameterization, which is why we prefer the permutation model as our supermodel choice; see Appendix C.1 for more details.

C MORE ON GOODNESS-OF-FIT TESTS WITH ODDS RATIOS

C.1 SEQUENTIAL MNAR MODEL AS A SUBMODEL OF NO SELF-CENSORING MODEL

As mentioned in Remark 1, the sequential MNAR model can be viewed as a submodel of the no self-censoring model. This provides a way to test independence restrictions of the form $R_k \perp\!\!\!\perp X_{\prec k} \mid R_{-k}, X_{\succ k}$. We provided an example with two variables using the m-DAG in Fig. 4(a) and showed how to use odds-ratio parameterization of the missingness mechanism to test the absence of an edge between X_1 and R_2 which implied $R_2 \perp\!\!\!\perp X_1|R_1$. Extending the idea to sequential MNAR models with $K > 2$ involves higher order interaction terms in the odds-ratio parameterization. We use the sequential MNAR model with three variables, shown in Fig. 1(a), to illustrate this point. The no self-censoring supermodel is shown in Fig. 1(b). We are interested in testing the absence of $X_1 \to R_2, X_1 \to R_3, X_2 \to R_3$ edges which implies the independence restrictions: $R_3 \perp\!\!\!\perp X_1, X_2|R_1, R_2$ and $R_2 \perp\!\!\!\perp X_1|R_1, R_3, X_3$. Let us focus on the former independence, i.e, $R_3 \perp\!\!\!\perp X_1, X_2|R_1, R_2$ which entails showing that $p(R_3|R_1, R_2, X_1, X_2)$ is not a function of X_1 and X_2. Note that $p(R_3|R_1, R_2, X_1, X_2) = p(R|X)/\sum_{R_3} p(R|X)$. The odds-ratio parameterization of $p(R|X)$ is as follows:

$$\begin{aligned} p(R \mid X) &= \frac{1}{Z} \times p(R_1|R_2 = R_3 = 1, X) \times p(R_2|R_1 = R_3 = 1, X) \times p(R_3|R_1 = R_2 = 1, X) \\ &\quad \times \text{OR}(R_2, R_1|R_3 = 1, X_1, X_2, X_3) \times \text{OR}(R_3, R_1, R_2|X) \\ &= p(R_1|R_2 = R_3 = 1, X_2, X_3) \times p(R_2|R_1 = R_3 = 1, X_1, X_2) \times p(R_3|R_1 = R_2 = 1, X_1, X_2) \\ &\quad \times f(R_2, R_1, X_3) \times \text{OR}(R_3, R_1, R_2|X). \end{aligned}$$

The equality uses assumptions in the no self-censoring supermodel: $R_k \perp\!\!\!\perp X_k|R_{-k}, X_{-k}, \forall k$ and the symmetry of the odds ratio to show $\text{OR}(R_2, R_1|R_3 = 1, X_1, X_2, X_3) = f(R_1, R_1, X_3)$. Thus, to show $p(R_3|R_1, R_2, X_1, X_2)$ is not a function of X_1 and X_2, it suffices to show that $p(R_3|R_1 = 1, R_2 = 1, X_1, X_2) \times \text{OR}(R_3, R_1, R_2|X)$ is not a function of X_1, X_2. Here, we see the higher order interaction term $\text{OR}(R_3, R_1, R_2|X)$ appearing. Even though estimating equations have been discussed in Malinsky et al. [2021] to estimate these higher order terms, they make the tests more challenging.

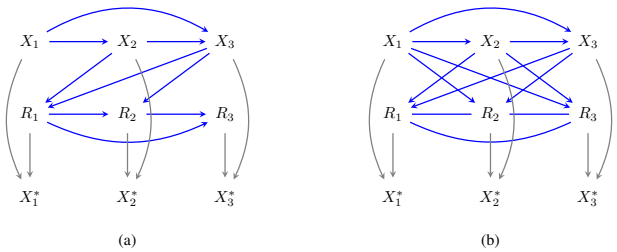


Figure 1: (a) Example of a sequential MNAR model; (b) The permutation supermodel.

The above representation becomes more complex as the number of variables increase. This makes it clear why using the saturated permutation model is relatively easier to test the sequential MNAR models.

C.2 SEQUENTIAL MAR MODEL AS A SUBMODEL OF PERMUTATION MODEL

Here, we discuss odds ratio independence test as an alternative to likelihood-ratio goodness-of-fit test in sequential MAR models (as submodels of permutation model). The independence restrictions we would like to test are: $R_k \perp\!\!\!\perp X_{\succ k} | R_{\prec k}, X^*_{\prec k}, \forall k$. We break down the independencies involving R_k into $K-k$ individual tests, i.e., we would like to test $R_k \perp\!\!\!\perp X_j | R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}, \forall X_j \in X_{\succ k}$, where $X_{\succ k, \prec j}$ denotes $\{X_{k+1}, \ldots, X_{j-1}\}$. As mentioned in the main draft, the conditional independence $A \perp\!\!\!\perp B | C$ holds if and only if $\text{OR}(A, B|C) = 1$ for all values of A, B, C. Therefore, to show the independence between R_k and X_j, we need to show that the following odds ratio is one for all levels of R_k, X_j with statistical significance-level α:

$$\begin{aligned}
&\text{OR}(R_k = r_k, X_j = x_j \mid R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}) \\
&\quad = \frac{p(R_k = r_k \mid X_j = x_j, R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}; \beta^a_k)}{p(R_k = 1 \mid X_j = x_j, R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}; \beta^a_k)} \times \frac{p(R_k = 1 \mid X_j = 1, R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}; \beta^a_k)}{p(R_k = r_k \mid X_j = 1, R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}; \beta^a_k)}.
\end{aligned}$$

To estimate the odds ratio, we need an estimate of β^a_k parameters. We use weighted estimating equations to estimate β^a_k. The intuition is as follows. Given that we have the permutation model as the supermodel, the independence restriction involving R_k and X_j is equivalent to the following Verma constraint:

$$R_k \perp\!\!\!\perp X_j \mid R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}, \text{do}(R_{\succ k, \prec j+1} = 1), \ \forall X_j \in X_{\succ k},$$

where the post intervention distribution is defined as follows:

$$p(. \mid \text{do}(R_{\succ k, \prec j+1} = 1)) = \frac{p(V)}{\prod_{i=k+1}^{j} p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))}\Bigg|_{R_{\succ k, \prec j+1}=1}.$$

Let $W_k(\beta_k) \coloneqq p(R_k | R_{\prec k}, X^*_{\prec k}, X_{\succ k, \prec j}, X_j; \beta_k)$ and let $\mathbb{P}_n\big[U(\beta_k)\big] = 0$ be an unbiased estimating equation for β_k wrt the full law (i.e., had there been no missingness). We can estimate β_k via the following weighted estimating equation:

$$\mathbb{P}_n\left[\frac{\mathbb{I}(R_{\succ k, \prec j+1} = 1)}{\prod_{i=k+1}^{j} \omega_i(\widehat{\eta}_i)} \times U(\beta_k)\right] = 0,$$

where $\omega_i(\eta_i) \coloneqq p(R_i \mid \text{pa}_{\mathcal{G}}(R_i); \eta)$, and $\widehat{\eta}_i$ denotes an estimate of η_i.

Since we have to evaluate the odds ratio for all values of X_j, the tests can become expensive in discrete cases and even more challenging in continuous cases, [Chen, 2021]. Hence, the likelihood-ratio test in Algorithm 1 might be preferred over odds ratio independence tests for larger graphs.

C.3 SEQUENTIAL MNAR MODEL AS A SUBMODEL OF PERMUTATION MODEL

The independence restrictions we would like to test are: $R_k \perp\!\!\!\perp X^*_{\prec k} | R_{\prec k}, X_{\succ k}, \forall k$. We break down the independencies involving R_k into $k-1$ individual tests, i.e., $R_k \perp\!\!\!\perp X^*_j | R_{\prec k}, X_{\succ k}, X^*_{\prec j}, \forall X^*_j \in X^*_{\prec k}$. As mentioned in the main draft, this is a context-specific independence restriction and is equivalent to $R_k \perp\!\!\!\perp X_j | R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}$. This independence holds if and only if the following odds ratio is one for all levels of X_j with statistical significance-level α :

$$\begin{aligned}
&\text{OR}(R_k = r_k, X_j = x_j \mid R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}) \\
&\quad = \frac{p(R_k = r_k \mid X_j = x_j, R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}; \beta^a_k)}{p(R_k = 1 \mid X_j = x_j, R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}; \beta^a_k)} \times \frac{p(R_k = 1 \mid X_j = 1, R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}; \beta^a_k)}{p(R_k = r_k \mid X_j = 1, R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}; \beta^a_k)}.
\end{aligned}$$

We can estimate the odds ratio by estimating the parameters β^a_k. We use weighted estimating equations to estimate the parameters and the intuition behind the choice of weights is that the restriction between R_k and X^*_j can be viewed as the following Verma constraint (under the permutation supermodel):

$$R_k \perp\!\!\!\perp X^*_j | R_{\prec k}, X_{\succ k}, X^*_{\prec j}, \text{do}(R_{\succ k} = 1), \ \forall X^*_j \in X^*_{\prec k}.$$

Let $W_k(\beta^a_k) \coloneqq p(R_k | X_j, R_{\prec k} \setminus R_j, R_j = 1, X_{\succ k}, X^*_{\prec j}; \beta^a_k)$ and let $\mathbb{P}_n\big[U(\beta^a_k)\big] = 0$ is unbiased estimating equation for β^a_k wrt the full law (had there been no missingness). We can estimate β^a_k via the following weighted estimating equation:

$$\mathbb{P}_n\left[\frac{\mathbb{I}(R_{\succ k} = 1)}{\prod_{j=k+1}^{K} p(R_j | \text{pa}_{\mathcal{G}}(R_j); \widehat{\eta}_j)} \times U(\beta^a_k)\right] = 0,$$

where $\widehat{\eta}_j$ is an estimate of η_j that parameterize the conditional density of $p(R_j|\,\mathrm{pa}_{\mathcal{G}}(R_j))$.

Similar to the sequential MAR model, the goodness-of-fit test based on odds ratio independence test can be rather challenging with continuous variables. Hence, the weighted likelihood-ratio tests might still be preferred.

C.4 BLOCK-PARALLEL MODEL AS A SUBMODEL OF NO SELF-CENSORING

Algorithm 2 TESTING BLOCK-PARALLEL $(\mathcal{M}, \mathcal{D}_n)$

for $k \in \{1, \ldots, K-1\}$ **do**
  Let $W_k(\beta_k) \coloneqq p(R_k = 1 \mid R_{-k} = 1, X_{-k}; \beta_k)$.
  Estimate β_k (denoted by $\widehat{\beta}_k$).
for each pair $k, j \in \{1, \ldots, K\}$ s.t. $k \neq j$ **do**
  Let $\theta(r_k, r_j) = \mathrm{OR}(R_k = R_j = 0 \mid R_{-kj} = 1, X)$
  Compute $\theta(R_k = 0, R_j = 0)$ via the following:

$$\frac{\mathbb{P}_n\Big[\prod_{i\neq\{k,j\}} R_i \times (1-R_k) \times (1-R_j)\Big]}{\mathbb{P}_n\Big[\prod_{i=1}^{K} R_i \times \frac{(1-W_k(\widehat{\beta}_k)) \times (1-W_j(\widehat{\beta}_j))}{W_k(\widehat{\beta}_k) \times W_j(\widehat{\beta}_j)}\Big]}$$

  Test $\theta(R_k{=}0, R_j{=}0){=}1$ at significance level α
  if test fails (i.e., $R_k \not\perp\!\!\!\perp R_j | X$) **then**
    return not block-parallel MNAR
return block-parallel MNAR

D PROOFS

Theorem 1. The intervention distribution $p(X, R \setminus R_{\succ k}, X^*|\mathrm{do}(R_{\succ k} = 1))$ factorizes wrt a CDAG $\mathcal{G}^*$ where edges into $R_{\succ k}$ have been removed from the sequential MAR graph $\mathcal{G}$. Factorization of this intervention distribution wrt a CDAG preserves the global Markov property, i.e., d-separation can be used to read dormant independencies in the intervention distribution. In $\mathcal{G}^*$ we have $R_k \perp\!\!\!\perp X_{\succ k} | R_{\prec k}, X^*_{\prec k}$ by d-separation implying the same independence holds in the intervention distribution. Finally, testability of this dormant independence from observed data follows from the fact that the propensity scores $p(R_j|\,\mathrm{pa}_{\mathcal{G}}(R_j))$ for each $R_j \in R_{\succ k}$ is identified under the restrictions implied by the graph $\mathcal{G}$ (identification is trivial since the sequential MAR model is a submodel of a permutation model that is fully identified), and upon intervention to $R_{\succ k} = 1$, each previously partially observed variable $X_j \in X_{\succ k}$ is now observed via a consistency argument $X_j = X^*_j$.

Theorem 2. The proof is very similar to the proof of Theorem 1. Interventions on $R_{\succ k}$ preserve the global Markov property and propensity scores of $R_{\succ k}$ are all identified as functions of observed data (since sequential MNAR is a submodel of fully identified permutation model). The m-CDAG we obtain after intervening on $R_{\succ k}$ and setting them to 1 is a graph where all incoming edges into $R_{\succ k}$ are removed and all $X_{\succ k}$ are observed random variables. Thus the dormant independence are direct functions of observed data.

Theorem 4. Given the restrictions of a block-parallel model, we note that including R_{-kj} in the conditioning set of independence $R_k \perp\!\!\!\perp R_j | X$ does not spoil the independence. Hence, we can equivalently look at $R_k \perp\!\!\!\perp R_j | X, R_{-kj} = 1$. Further, we know this independence holds if and only if $\mathrm{OR}(R_k, R_j | X, R_{-kj} = 1) = 1$. All we need to show now is that $\mathrm{OR}(R_k, R_j | X, R_{-kj} = 1) = \mathrm{OR}(R_k, R_j | X_{-kj}, R_{-kj} = 1)$. Using an odds-ratio parameterization of

$p(R_k, R_j|X, R_{-kj} = 1)$ we have:

$$\begin{aligned}\text{OR}(R_k = r_k, R_j = r_j \mid X, R_{-kj} = 1) &= \frac{p(R_k = r_k \mid R_j = r_j, X, R_{-kj} = 1)}{p(R_k = 1 \mid R_j = r_j, X, R_{-kj} = 1)} \times \frac{p(R_k = 1 \mid R_j = 1, X, R_{-kj} = 1)}{p(R_k = r_k \mid R_j = 1, X, R_{-kj} = 1)} \\ &= \frac{p(R_k = r_k \mid R_j = r_j, X_{-k}, R_{-kj} = 1)}{p(R_k = 1 \mid R_j = r_j, X_{-k}, R_{-kj} = 1)} \times \frac{p(R_k = 1 \mid R_j = 1, X_{-k}, R_{-kj} = 1)}{p(R_k = r_k \mid R_j = 1, X_{-k}, R_{-kj} = 1)} \\ &= f_1(R_k, R_j, X_{-k}, R_{-kj} = 1).\end{aligned}$$

The second equality holds because $R_k \perp\!\!\!\perp X_k | R_{-k}, X_{-k}$, and

$$\begin{aligned}\text{OR}(R_j = r_j, R_k = r_k \mid X, R_{-kj} = 1) &= \frac{p(R_j = r_j \mid R_k = r_k, X, R_{-kj} = 1)}{p(R_j = 1 \mid R_k = r_k, X, R_{-kj} = 1)} \times \frac{p(R_j = 1 \mid R_k = 1, X, R_{-kj} = 1)}{p(R_j = r_j \mid R_k = 1, X, R_{-kj} = 1)} \\ &= \frac{p(R_j = r_j \mid R_k = r_k, X_{-j}, R_{-kj} = 1)}{p(R_j = 1 \mid R_k = r_k, X_{-j}, R_{-kj} = 1)} \times \frac{p(R_j = 1 \mid R_k = 1, X_{-j}, R_{-kj} = 1)}{p(R_j = r_j \mid R_k = 1, X_{-j}, R_{-kj} = 1)} \\ &= f_2(R_k, R_j, X_{-j}, R_{-kj} = 1).\end{aligned}$$

The second equality holds because $R_j \perp\!\!\!\perp X_j | R_{-j}, X_{-j}$. Due to symmetry of odds ratio, $f_1(R_k, R_j, X_{-k}, R_{-kj} = 1)$ and $f_2(R_k, R_j, X_{-j}, R_{-kj} = 1)$ must be equal. This implies $\text{OR}(R_k, R_j|X, R_{-kj} = 1) = \text{OR}(R_k, R_j|X_{-kj}, R_{-kj} = 1)$ (all a function of observed data).

Even though the odds ratio is a function of observed data, estimation of odds ratio is not straightforward. We rely on the estimating equations discussed in this Appendix and Malinsky et al. [2021] to estimate the odds ratios.

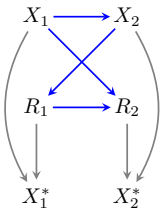
Theorem 3. To prove this result, it suffices to show that the target law in the criss-cross structure on two variables (drawn on the right hand side) is not non-parametrically identified. For this purpose, we provide an example of two different full laws that factorize according to the criss-cross model, but map into the same observed data law.

X_1	$p(X_1)$
0	a
1	$1-a$

X_2	X_1	$p(X_2 \mid X_1)$
0	0	b
1	0	$1-b$
0	1	c
1	1	$1-c$

R_1	X_2	$p(R_1 \mid X_2)$
0	0	d
1	0	$1-d$
0	1	e
1	1	$1-e$

R_2	R_1	X_1	$p(R_2 \mid R_1, X_1)$
0	0	0	f
1	0	0	$1-f$
0	0	1	g
1	0	1	$1-g$
0	1	0	h
1	1	0	$1-h$
0	1	1	i
1	1	1	$1-i$



R_1	R_2	X_1	X_2	p(FULL LAW)	X_1^*	X_2^*	p(OBSERVED LAW)
0	0	0	0	$abdf$	?	?	$d\Big[abf + (1-a)cg\Big] + e\Big[a(1-b)f + (1-a)(1-c)g\Big]$
		1	0	$(1-a)cdg$			
		0	1	$a(1-b)ef$			
		1	1	$(1-a)(1-c)eg$			
0	1	0	0	$abd(1-f)$	?	0	$d\Big[ab(1-f) + (1-a)c(1-g)\Big]$
		1	0	$(1-a)cd(1-g)$			
		0	1	$a(1-b)e(1-f)$		1	$e\Big[a(1-b)(1-f) + (1-a)(1-c)(1-g)\Big]$
		1	1	$(1-a)(1-c)e(1-g)$			
1	0	0	0	$ab(1-d)h$	0	?	$ah\Big[b(1-d) + (1-b)(1-e)\Big]$
		1	0	$(1-a)c(1-d)i$			
		0	1	$a(1-b)(1-e)h$	1		$(1-a)i\Big[c(1-d) + (1-c)(1-e)\Big]$
		1	1	$(1-a)(1-c)(1-e)i$			
1	1	0	0	$ab(1-d)(1-h)$	0	0	$ab(1-d)(1-h)$
		1	0	$(1-a)c(1-d)(1-i)$	1	0	$(1-a)c(1-d)(1-i)$
		0	1	$a(1-b)(1-e)(1-h)$	0	1	$a(1-b)(1-e)(1-h)$
		1	1	$(1-a)(1-c)(1-e)(1-i)$	1	1	$(1-a)(1-c)(1-e)(1-i)$

A concrete example is as follows:

X_1	$p(X_1)$	
	M_1	M_2
0	7/15	5/11
1	8/15	6/11

X_2	X_1	$p(X_2 \mid X_1)$	
		M_1	M_2
0	0	6/7	4/5
1	0	1/7	1/5
0	1	3/4	2/3
1	1	1/4	1/3

R_1	X_2	$p(R_1 \mid X_2)$	
		M_1	M_2
0	0	19/20	189/200
1	0	1/20	11/200
0	1	85/100	89/100
1	1	15/100	11/100

R_2	R_1	X_1	$p(R_2 \mid \mid R_1, X_1)$	
			M_1	M_2
0	0	0	268/323	7636/16821
1	0	0	55/323	9185/16821
0	0	1	208/323	16216/16821
1	0	1	115/323	605/16821
0	1	0	1/2	1/2
1	1	0	1/2	1/2
0	1	1	1/2	1/2
1	1	1	1/2	1/2

R_1	R_2	X_1	X_2	$p(R, X)$		X_1^*	X_2^*	$p(R, X^*)$
				M_1	M_2			$M_1 = M_2$
0	0	0	0	134/425	3818/24475	?	?	68/100
		1	0	104/425	8108/24475			
		0	1	67/1425	1118/30439			
		1	1	104/1425	8108/51975			
0	1	0	0	11/170	167/890	?	0	2/10
		1	0	23/170	11/890			
		0	1	11/1140	167/3780		1	1/20
		1	1	23/570	11/1890			
1	0	0	0	1/100	1/100	0	?	3/200
		1	0	1/100	1/100			
		0	1	1/200	1/200	1		2/100
		1	1	1/100	1/100			
1	1	0	0	1/100	1/100	0	0	1/100
		1	0	1/100	1/100	1	0	1/100
		0	1	1/200	1/200	0	1	1/200
		1	1	1/100	1/100	1	1	1/100

From the above example, we see that none of the parameters in red are identified.

E SIMULATIONS

As mentioned in the main draft, we describe three sets of simulations to illustrate the key results and the utility of our proposed methods – each set focuses on a class of missing data models that we considered in the main draft. For each simulation, we generate four random variables from either a multivariate normal distribution or binomial distribution. We induce missing values in all four variables according to a missingness mechanism that follows restrictions of either sequential MAR, sequential MNAR, block-parallel, or supermodels of them. All code necessary to reproduce our simulations is included with this submission. The data generating mechanism is described as follows.

Generating X: For Gaussian data, we generate four random variables from multivariate normal distribution with mean zero and covariance matrix σ where the ij-th entry is $\sigma_{ij} = 1 - |i - j| \times 0.25$. For binary data, variable X_k is generated from a binomial distribution with the probability of observing $X_k = 1$ given $X_{\prec k}$ equals to expit$\big(a^0_{x_k} + \sum_{j \prec k} a^j_{x_k} \times X_j\big)$, where expit$(x) = 1/(1 + \exp(-x))$ and parameters $a^j_{x_k}$ (for all $k = 1, \ldots, K$ and $j \prec k$) are generated uniformly from the $(-1, 1)$ interval.

Generating R: In each class of missing data model, we consider generating R according to two scenarios: one where the restrictions in the missing data model we would like to test hold true (the null hypothesis should be accepted) and one where the restrictions are violated (the null hypothesis should be rejected in favor of accepting the corresponding supermodel). All missingness indicators are generated from binomial distributions. The details on missing data parameters are as follows.

$$p(R_k = 1 \mid R_{\prec k}, X^*_{\prec k}, X_{\succ k}) = \text{expit}\left(a^0_k + \sum_{j \prec k} b^j_k \times R_j + c^j_k \times R_j X^*_j + \sum_{i \succ k} d^i_k \times X_i\right),\ k = 1, \ldots, 4 \quad \text{(Simulation 1)}$$

$$p(R_k = 1 \mid R_{\prec k}, X_{\succ k}, X^*_{\prec k}) = \text{expit}\left(a^0_k + \sum_{i \succ k} d^i_k \times X_i + \sum_{j \prec k} b^j_k \times R_j + c^j_k \times R_j X^*_j\right),\ k = 1, \ldots, 4 \quad \text{(Simulation 2)}$$

$$p(R_k = 1 \mid X_{-k}) = \text{expit}\left(a^0_k + \sum_{j \neq k} b^j_k \times X_j\right),\ k = 1, \ldots, 4 \quad \text{(Simulation 3).} \tag{4}$$

Addition of the blue terms simulate scenarios where the independence assumptions we would like to test are violated. All the parameters are randomly generated from a uniform distribution. In order to control the proportion of missing values, we run the experiments with three different ranges for the uniform distribution: $(-1, 1)$, $(-0.5, 1.5)$, and $(0, 2)$.

Generating X^*: For each given sample, if $R_k = 1$ then $X^*_k = X$, otherwise $X^* = \text{NA}$.

Our objective is to test the missing data restrictions by relying only on observed data, i.e., (R, X^*) samples.

Simulation 1. In the first set of simulations, we focused on testing the sequential MAR model defined via the set of restrictions in (4). The results were provided and discussed in the main draft.

We briefly add that when true underlying missingness mechanism satisfies the assumptions of the sequential MAR model, missingness indicators are generated from (4) without the blue terms. When the restrictions are no longer valid, missingness indicators are generated from (4) with the blue terms.

Simulation 2. In the second set of simulations we focus on testing the sequential MNAR model defined via the set of restrictions in (5). We follow Algorithm 1 to test the independence restrictions, which entails running a total of $K - 1$ tests. Our test statistic is $2 \times \rho$ and we use a chi-square distribution with $k - 1$ degrees of freedom to evaluate the goodness-of-fits – the degree of freedom is chosen as the difference between number of parameters in $W_k(\beta^a_k)$ and $W_k(\beta^0_k)$, as defined in the algorithm. If the p-values are all greater than 0.05, we accept the sequential MNAR model.

For a fixed sample size, we simulate 100 different datasets and calculate the acceptance rate of a sequential MNAR model. The acceptance rate is plotted as a function of sample size in Fig. 2. The sample size ranges from $1,000$ to $15,000$ with 500 increments. In each panel, there are three plots that vary in terms of the proportion of complete cases in the dataset, i.e, $6\%, 30\%, 48\%$. The top row illustrates the results when the true underlying missingness mechanism satisfies the assumptions of the sequential MNAR model (missingness indicators are generated from (4) without the blue terms) and the bottom row illustrates results for when the restrictions are no longer valid (missingness indicators are generated from (4) with the blue terms). As it is shown, the acceptance rate is quite low when the independence restrictions of a sequential MNAR model are not valid; even when we only have 6% of complete cases the tests perform well. When the sequential MNAR model assumptions are true, the acceptance rate increases as missing rate decreases and reaches very close to 1 when we have only 48% complete cases.

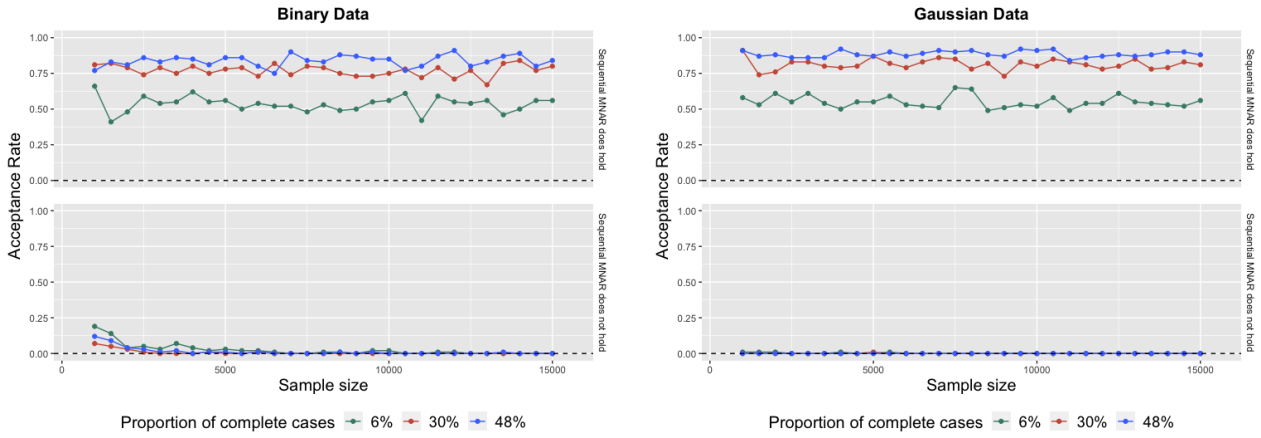


Figure 2: Results on testing **sequential MNAR** models. In the top row, the sequential MNAR model captures the true underlying missingness mechanism. The assumptions of sequential MNAR model are violated in the bottom row.

Simulation 3. In the third set of simulations we focus on testing independencies between missingness indicators in a block-parallel MNAR model defined via the set of restrictions in (6). Testing the full model requires following Algorithm 2 which entails running a total of $\binom{K}{2}$ tests (between all distinct pairs of missingness indicators.) For illustration purposes, we focus on testing only one pair of missingness indicator in two different scenarios: one where the true underlying missingness mechanism follows the restrictions of a block-parallel model – thus $R_k \in R$ is generated using (4), and one where the missingness mechanism factorizes as $\prod_{k=1}^{K} p(R_k|R_{\succ k}, X_{\prec k})$ which is still a submodel of the no-self censoring model but violates the assumptions of the block-parallel model. We focus on testing the independence $R_1 \perp\!\!\!\perp R_2|X$ by calculating the odds ratio $\theta := \text{OR}(R_1 = 0, R_2 = 0|X)$ via the following estimating equation and showing that the value is one.

$$\mathbb{P}_n\Big[R_1 \times R_2 \times R_3 \times \frac{p(R_1 = 0 \mid R_2 = 1, R_3 = 1, X_2, X_3) \times p(R_2 = 0 \mid R_1 = 1, R_3 = 1, X_1, X_3)}{p(R_1 = 1 \mid R_2 = 1, R_3 = 1, X_2, X_3) \times p(R_2 = 1 \mid R_1 = 1, R_3 = 1, X_1, X_3)} \times \theta$$
$$- R_3 \times (1 - R_1) \times (1 - R_2)\Big] = 0.$$

For a fixed sample size, we simulate 100 different datasets and calculate the odds ratio via the above estimating equation. We provide the boxplots in Fig. 3. The x-axis is sample size that ranges from $1,000$ to $10,000$ with $2,000$ increments. The left panel illustrates the boxplots for binary and Gaussian data when the true missingness mechanism follows the restrictions of the block-parallel model, and in the right panel it does not. As it is shown, the boxplots are centered around 1 in the left panel as expected, but move away from 1 when the independence does not hold. To perform a formal test, we can construct confidence intervals for each sample size via bootstrapping the data generations and odds ratio calculations.

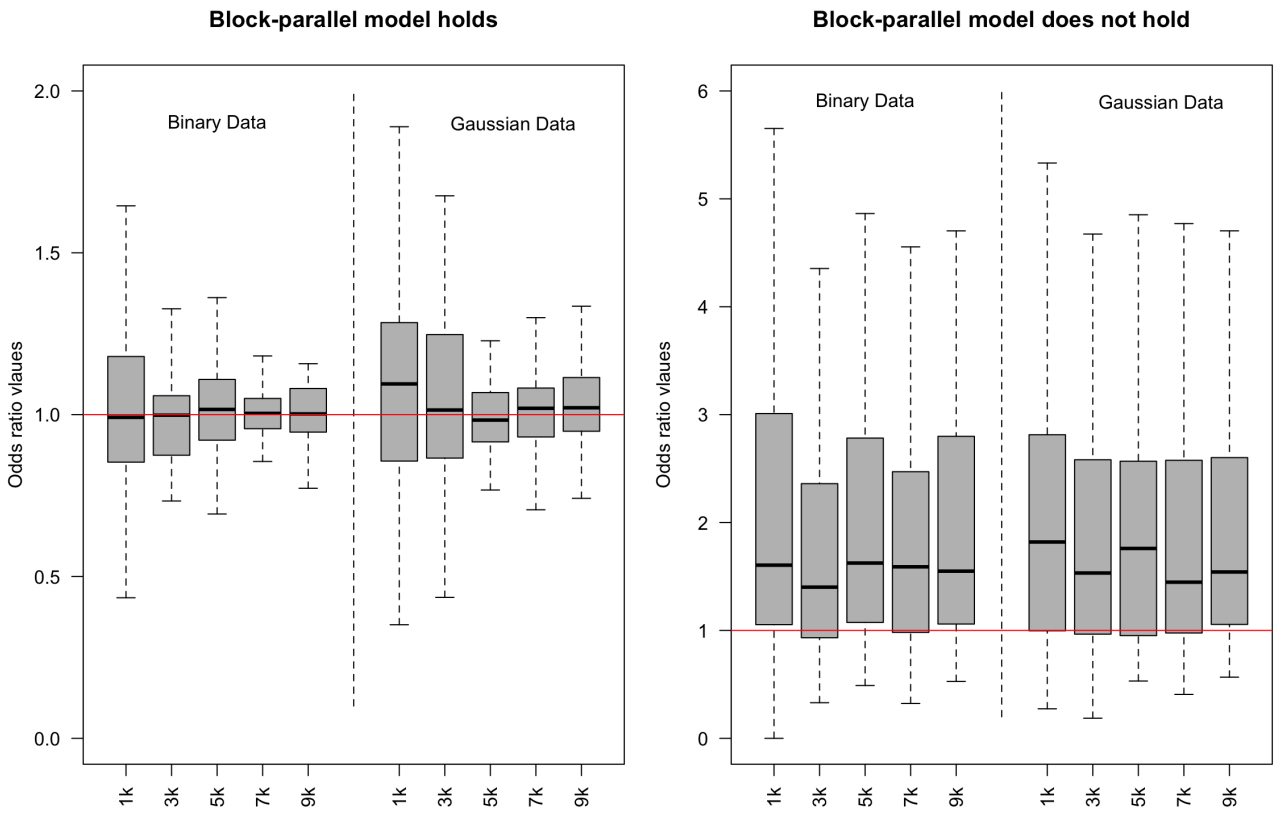


Figure 3: Results on computing (conditional) odds ratio between a pair of missingness indicators to test an independence restriction between them. On the left panel, the block-parallel MNAR model captures the true underlying missingness mechanism. The assumptions of block-parallel MNAR model are violated on the right panel.

References

Hua Yun Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421, 2007.

Hua Yun Chen. *Semiparametric Odds Ratio Model and Its Applications*. Chapman and Hall/CRC, 2021.

Daniel Malinsky, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9, 2021.

Karthika Mohan and Judea Pearl. On the testability of models with missing data. In *Artificial Intelligence and Statistics*, pages 643–650. PMLR, 2014.

Karthika Mohan, Judea Pearl, and Tian Jin. Missing data as a causal inference problem. In *Proceedings of the Neural Information Processing Systems Conference*, 2013.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the Twenty Seventh International Conference on Machine Learning (ICML-20)*, 2020.

Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, and James M Robins. Causal and counterfactual views of missing data models. *arXiv preprint arXiv:2210.05558*, 2022.

James M Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

Donald B Rubin. Causal inference and missing data (with discussion). *Biometrika*, 63:581–592, 1976.

Ilya Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. *Advances in Neural Information Processing Systems*, 29:3144–3152, 2016.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. PMLR, 2019.