| Statistica Sinica Preprint No: SS-2025-0164 | |
|---|---|
| **Title** | Discussion of "Causal and Counterfactual Views of Missing Data Models" by Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, James Robins |
| **Manuscript ID** | SS-2025-0164 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202025.0164 |
| **Complete List of Authors** | Zeyi Wang and<br>Mark J. van der Laan |
| **Corresponding Authors** | Zeyi Wang |
| **E-mails** | zwang107@gmail.com |

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164



# Discussion of "Causal and Counterfactual Views of Missing Data Models" by Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, James Robins

Zeyi Wang and Mark J. van der Laan

*Oklahoma State University*

*University of California, Berkeley*

## 1. Introduction

**Viewing missingness as longitudinal intervention enabled closed-form efficient estimation for challenging bivariate censoring problems.**

Opposite to the common view of causal inference as missing data problems, viewing missingness as (longitudinal) intervention has also been a fruitful strategy, leveraging identification and estimation techniques from causal inference literature. For example, in the challenging bivariate censoring problem, viewing the complex censoring mechanism as longitudinal interventions under sequential randomization assumptions (SRA; it can be shown that SRA $\subset$ MAR for multivariate right-censored data and SRA = MAR for univariate right-censored data) leads to closed-form efficient influence curves which would have not been possible with only the MAR assumptions (Rubin, 1976; van der Laan and Robins, 2003). For example, if the full data is $X_j(t) = I(T_j \leq t), j = 1, 2$ defined by survival times $T_1, T_2$ that are subject to censoring times $C_1, C_2$, then only assuming MAR is significantly weaker than assuming $(C_1, C_2) \perp (X_1, X_2)$, resulting in an efficient influence curve that does not exist in closed form. Even though such

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164

SRA based estimators are inefficient if one truly only assumes MAR, they can flexibly incorporate time-dependent covariate information and provide meaningful efficiency improvement (van der Laan and Robins, 2003). There also exist scenarios (for example, when $C_1$ and $T_2$ are dependent, such as when monitoring, defined as interval censoring, is part of the intervention; see Carone et al. (2012); van der Laan (2018)) where SRA is plausible whereas MAR is violated.

**The missing data model is a special case of multivariate/bivariate censoring.**

In Nabi et al., the missing data model defined by full data variables $L_1^{(1)}, \ldots, L_K^{(1)}$, missingness indicators $R_1, \ldots, R_K$, and observed data variables $L_1, \ldots, L_K$ is a special case of multivariate censoring. For $k = 1, \ldots, K$, let $T_k \in \mathbb{R}$, $C_k \in \{-\infty, \infty\}$, and define $\tilde{T}_k = \min\{T_k, C_k\}$ with a non-censoring indicator $\Delta_k = I(C_k > T_k)$. Then $L_k^{(1)} = T_k, R_k = \Delta_k, L_k = \tilde{T}_k$ define the same missing data model, except that when $R_k = 0$ we let $L_k = -\infty$ instead of $L_k =?$. Therefore, the examples with $K = 2$ (e.g. Figure 2-4) are special cases of bivariate censoring. In fact, when $T_k$ truly represents a survival time, the multivariate/bivariate censoring models allow additional time-dependent covariates that predict both censoring and survival.

**SRA may be limited without actual time series data structure.**

Although SRA is a practical MAR assumption in many real-world data problems, it can be limited in missing data problems without censoring and longitudinal components.

Consider the missing data model induced by the special case of bivariate censoring ($K = 2$) above. Let $T_k, k = 1, 2$ be discrete survival time variables or categorical variables with supports $0, 1, \ldots, p$. Let $L_k^{(1)}(t)$ include $I(T_k \leq t)$ as component. Then $L_k^{(1)}(0) = I(T_k = 0)$. Let $L_k(t) = -\infty$ if $C_k = -\infty$ and $L_k(t) = L_k^{(1)}(t)$ if $C_k = \infty$. Now $SRA \subset MAR$ is equivalent to: $p(\boldsymbol{R}(t)|\bar{\boldsymbol{R}}(t-1), \boldsymbol{L}^{(1)}) = p(\boldsymbol{R}(t)|\bar{\boldsymbol{R}}(t-1), \boldsymbol{L}(t))$ for $t = 1, \ldots, p$. But for such missing data models, $R_k(t) = R_k(0)$ for $t = 1, \ldots, p$. At $t = 0$, it reduces to $p(\boldsymbol{R}|\boldsymbol{L}^{(1)}) = p(\boldsymbol{R}|\boldsymbol{L}(0))$. As $\boldsymbol{L}(0) = \emptyset$, SRA here becomes trivially MCAR; yet MAR is a more complex concept allowing for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164

dependence of censoring and full data. Under a missing data structure without time ordering, a submodel of MNAR may be more plausible and/or easier to identify than MAR and SRA, highlighting the need for practical tools to identify and analyze realistic, identifiable MNAR submodels.

## 2. Contributions of Missing Data DAGs

**Viewing missingness as intervention, in combination with graphical representation, creates missing data DAGs that can construct flexible, identifiable submodels of MNAR.**

In a missing data DAG, the fully observed counterfactuals are defined and presented as vertices. By incorporating fully observed counterfactuals into causal graphs, a missing data DAG naturally leverages the idea of viewing missingness as intervention. This makes it particularly suitable for representing and identifying some MNAR (and some MAR) models

### 2.1 Representing Existing Models with Missing Data DAGs

Permutation missingness (which can be generalized to sequential CAR, a layered extension of CAR; see Robins (1997); Gill and Robins (1997)) is a plausible MNAR submodel in many real-world applications. To see how it works, consider baseline covariates $W = (W_1, W_2)$, a randomized treatment $A \in \{0, 1\}$, and potential outcomes $Y^{(0)}, Y^{(1)}$. However, coarsening indicator $\Delta = 0$ hides a key confounder $W_1$. The actual observed data is $(\Delta, \Delta W_1, W_2, A, Y^{(A)})$. Sequential CAR analyzes the following factorization:

$$p(A = a, \Delta = \delta | W, Y^{(0)}, Y^{(1)}) = p(A = a | W_1, W_2) p(\Delta = \delta | W_2, A, Y^{(A)}), \tag{2.1}$$

which is a direct violation of CAR as A depends on partially observed $W_1$. CAR would only allow $p(A = a | \Delta = \delta, \delta W_1, W_2)$ depending on observed components which is much more restrictive.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164



A two-stage sampling procedure, where the second stage measures with controlled randomization a confounder omitted in the first stage (such as $W_1$ in (2.1)), fits naturally into sequential CAR. In other scenarios, multiple permutations (orderings) of the variables may construct sequential CAR models; however, which ordering actually generated the data cannot be learned from the observed information without further knowledge on the data generating process.

Permutation missing models and sequential CAR conditions are formally defined without the graphical component. Some missing data DAGs, through d-separation, imply conditional independence conditions that are equivalent to the permutation missingness. However, each one of such missing data DAGs (or its Markov equivalent classes) will be inherently more restrictive than the non-graphical definition, due to the additional local Markov conditions implied by the graphical structure.

Block-conditional MAR (BCMAR Zhou et al. (2010)) is a natural MNAR extension of MAR under non-monotone missingness. A time or causal order still exists for blocks of variables (each block consists of the full data within-block and its missingness indicators), which makes BCMAR a subset of block-sequential models. A subset of BCMAR (referred to as "block-sequential MNAR" in this manuscript, where the vertices need to be generalized to random vectors to truly represent the blocks) is represented and identified as missing data DAGs. Interestingly, the identification result involves only the monotone-missing cases even though the missingness model is not monotone.

## 2.2 Identifying Novel MNAR Submodels

Missing data DAGs become particularly appealing as it provides a principled identification strategy, fixing missingness variables in a way similar to fixing treatment variables in causal

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164

DAGs. The resulting identification strategies are highlighted by sequentially or simultaneously fixing the missingness variables. This makes identification easily available for a large class of (graphically represented) MNAR models, some previously not identified or considered.

Graph-based rules for detecting non-identifiable scenarios are also represented by edge patterns in missing data DAGs, such as self-censoring and criss-cross structures. The necessary and sufficient conditions of identifiability of missing data DAGS remain an open question.

## 3. Realistic Examples

The plausibility of previously mentioned models can be evaluated under the following conceptual examples. We start with two variables, the observed version, $L_1, L_2$, the full version $L_1(1), L_2(1)$ and the missingness indicators $R_1, R_2$. The following ones are relevant when there exists time ordering from $L_1(1), L_1, R_1$ to $L_2(1), L_2, R_2$ and can be generalized to more time points.

1. MAR is appropriate when missingness $R_2$ only depends on the observed components $L_1, R_1$. MAR is equivalent to SRA in this scenario. For example, when the missingness is monotone (e.g. dropouts) and there exists no unmeasured confounding. This can be satisfied in well-planned clinical studies with scheduled visits, where unmeasured confounding and measurement error remain the common violations.

$$p_{l_1(1),l_2(1)}(l_1, l_2) = \frac{p(l_1, l_2, r_1 = 1, r_2 = 1)}{p(r_2 = 1 \mid r_1 = 1, l_1) \times p(r_1 = 1)}. \tag{3.2}$$

2. BCMAR is appropriate when missingness $R_2$ depends on the full version $L_1(1)$ and $R_1$. For example, when the missingness is non-monotone but a time or causal ordering still presents from $(L_1(1), R_1)$ to $(L_2(1), R_2)$. This is also common in clinical studies with repeated measurements where missingness is intermittent (allowing $(R_1, R_2) = (0, 1)$)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164

rather than drop-out.

$$p_{l_1(1),l_2(1)}(l_1,l_2) = \frac{p(l_1,l_2,r_1=1,r_2=1)}{p(r_2=1 \mid r_1=1,l_1)\times p(r_1=1)}. \tag{3.3}$$

3. Permutation missingness (or sequential CAR) is appropriate if $L_2(1)$ is a key confounder deciding the missingness $R_1$ of $L_1(1)$, and if $R_2$ represents a fully controlled additional random sampling procedure decided by fully observed $L_1, R_1$. This is a practical assumption in follow-up rounds of surveys, where later rounds sample the key missingness confounders in the prior round until it reaches MAR (for $R_2$ in this example).

$$p_{l_1(1),l_2(1)}(l_1,l_2) = \frac{p(l_1,l_2,r_1=1,r_2=1)}{p(r_2=1 \mid r_1=1,l_1)\times \frac{\sum_{l_1} p(l_2|r_2=1,r_1=1,l_1)\times p(r_1=1,l_1)}{\sum_{l_1,r_1} p(l_2|r_2=1,r_1,l_1)\times p(r_1,l_1)}}$$

There exist other missing data DAGs that are identifiable but appear to be less connected to real-world data analysis applications, even for a simple model such as block-parallel MNAR.

When more variables are included in a missing data DAG, the number of ways of identifications increases, especially for sequential identification strategies under Markov equivalence. A unified and efficient search algorithm for identification strategies is of interest. It also remains an open question which of them are most relevant in real-world data problems; research is needed showcasing their applications, either as the main model or in sensitivity analysis.

Lastly, estimation based on aforementioned identifications poses another challenge. For example, although the identification results (3.2) and (3.3) coincide, BCMAR posits unintuitive model restrictions on the observed data distribution (Hunt, 2020), complicating the construction of efficient estimators. In addition, different forms of propensity scores are impacted to different degrees by near-violation of positivity assumptions (Petersen et al., 2012), which warrants future research. Finally, the implicitly imposed restriction (e.g. by BCMAR) can challenge the validity of seemingly simple and non-parametric missingness models, requiring statistical tests of the proposed graphical models based on the statistical restriction.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164

## 4. Discussion

**Missing data hierarchy through the lens of DAGs.**

While MNAR encompasses a more complex collection of distributions, usually considered more general than MAR, missing data DAGs now allow for flexible specification for submodels of MNAR, challenging this conventional viewpoint. Because of the graphical and structural constraints, each missing data DAG (or its Markov equivalence class) can be just as restrictive or even more so, as a submodel, than MAR.

Committing to one missing data DAG, even an MNAR one, may itself be a strong model assumption, especially with a large number of variables, by enforcing all the local Markov conditions. Therefore, caution should be exercised. In practice, the time or causal ordering among variables may be partially known or indeterminable. In such cases, no DAG can accurately represent the data generating process, motivating generalizations to realistic bi-directed graphs (Richardson et al., 2023) for missing data.

Although particular missing data DAGs construct testable assumptions, the testability is rooted in strong structural restrictions that often make them less plausible. Such tests may lack sufficient power to detect violations. In contrast, untestable assumptions can be more plausible, especially when reflecting domain knowledge. Various time ordering based non-testable assumptions (e.g. MAR under dropouts, sequential CAR under follow-up surveys; or imposing subtle model restrictions, e.g. BCMAR) are more applicable to typical data generating procedures as discussed in Section 3.

It all comes down to the question: which is more relevant and appropriate in real-world applications? As numerous novel MNAR submodels can now be straightforwardly defined and identified using graphs, there is a growing need for demonstrations on real-world data. In many cases, multiple identification strategies may be equally plausible or differ only slightly

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164



in the submodel representation; a quantitative trade-off may arise. Some submodels may lead to estimators with lower variance but require slightly stronger assumptions. A slightly more realistic model may come at the cost of estimation difficulty or large variance (a trade-off similar to what occurs with weak instrumental variables). Choosing among them depends not only on identifiability but also on estimation performance and practical interpretability. An important extension discussed in Supplementary Materials is allowing hidden variables, which may construct multiple realistic models for structured and principled sensitivity analysis involving unmeasured confounding.

## References

Carone, M., M. Petersen, and M. J. van der Laan (2012). *Targeted minimum loss based estimation of a casual effect using interval censored time to event data*. Chapman & Hall/CRC, New York, NY, USA.

Gill, R. D. and J. M. Robins (1997). Sequential models for coarsening and missingness. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival analysis*, pp. 295–305. Springer.

Hunt, L. (2020). *Causal Inference and Missing Data in Longitudinal Studies*. Ph. D. thesis, Johns Hopkins University.

Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research 21*(1), 31–54.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0164

# REFERENCES

Richardson, T. S., R. J. Evans, J. M. Robins, and I. Shpitser (2023). Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics 51*(1), 334–361.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in medicine 16*(1), 21–37.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

van der Laan, M. J. (2018). HAL estimator of the efficient influence curve. *Targeted learning in data science: Causal inference for complex longitudinal studies 8*, 103–123.

van der Laan, M. J. and J. M. Robins (2003). Multivariate right-censored multivariate data. In *Unified Methods for Censored Longitudinal Data and Causality*, pp. 266–310. Springer.

van der Laan, M. J. and J. M. Robins (2003). *Unified methods for censored longitudinal data and causality*. Springer.

Zhou, Y., R. J. Little, and J. D. Kalbfleisch (2010). Block-conditional missing at random models for missing data. *Statistical Science 25*(4), 517–532.

Zeyi Wang, Department of Statistics, Oklahoma State University; School of Public Health, University of California, Berkeley.

E-mail: zeyi.wang@okstate.edu

Mark J. van der Laan, Division of Biostatistics, School of Public Health, University of California, Berkeley.

E-mail: laan@stat.berkeley.edu