










Causal Discovery with Missing Data in a Multicentric Clinical Study

Alessio Zanga^{1,2}(✉) , Alice Bernasconi^{1,3} , Peter J.F. Lucas⁴ ,
Hanny Pijnenborg⁵ , Casper Reijnen⁵ , Marco Scutari⁶ ,
and Fabio Stella¹ 

¹ DISCo, University of Milano - Bicocca, Milan, Italy

² F. Hoffmann - La Roche Ltd, Basel, Switzerland
a.zanga3@campus.unimib.it

³ Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

⁴ University of Twente, Enschede, The Netherlands

⁵ RadboudUMC, Nijmegen, The Netherlands

⁶ IDSIA, Lugano, Switzerland

Abstract. Causal inference for testing clinical hypotheses from observational data presents many difficulties because the underlying data-generating model and the associated causal graph are not usually available. Furthermore, observational data may contain missing values, which impact the recovery of the causal graph by causal discovery algorithms: a crucial issue often ignored in clinical studies. In this work, we use data from a multi-centric study on endometrial cancer to analyze the impact of different missingness mechanisms on the recovered causal graph. This is achieved by extending state-of-the-art causal discovery algorithms to exploit expert knowledge without sacrificing theoretical soundness. We validate the recovered graph with expert physicians, showing that our approach finds clinically-relevant solutions. Finally, we discuss the goodness of fit of our graph and its consistency from a clinical decision-making perspective using graphical separation to validate causal pathways.

Keywords: Causal discovery · Causal graphs · Missing data

1 Introduction

Much of the data collected in clinical research is observational, collected as part of daily clinical practice. Correctly interpreting them requires a good understanding of their characteristics and of possible sources of bias. A common one is missing values, which may arise in three different ways [1]: *data missing completely at random* (MCAR), *data missing at random* (MAR), and *data missing not at random* (MNAR) that are neither MCAR nor MAR. MNAR is common in clinical observational data and thus interesting to study, as it is often possible to unravel the reason for the missingness: for instance, a laboratory test may be skipped in favour of a more precise ones available at a later stage.

Missing values in clinical data are commonly imputed with heuristics or with single/multiple imputation. Such techniques assume that the data are MCAR or MAR; we cannot test whether these assumptions are valid without knowing the missingness mechanism but, at the same time, if these assumptions do not hold our clinical conclusions are likely to be biased [2]. A possible approach to this problem is *causal discovery*: modelling the missingness mechanism is to recover the underlying causal graph \mathcal{G}^* , given the data \mathcal{D} and the prior knowledge \mathcal{K} [3]. In our previous work on endometrial cancer [EC; 4], we proposed a new causal discovery approach based on bootstrapping for clinical data with low sample size and high missingness assuming MAR (*Bootstrap SEM*). Algorithms assuming MNAR were not available until recently when *HC-aIPW* [5] was introduced.

Our aim is to showcase how modern causal discovery techniques can model the biases in observational data, in particular for MNAR. For this purpose, we applied different causal discovery algorithms with different assumptions to data from a multicenter study on EC, highlighting the clinical implications of their biases on recovering the causal mechanisms behind the prognosis of EC.

2 Background

A causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a directed acyclic graph (DAG) where for each directed edge $(X, Y) \in \mathbf{E}$, X is a direct cause of Y and Y is a direct effect of X . The vertex set \mathbf{V} is usually split into two disjoint subsets $\mathbf{V} = \mathbf{O} \cup \mathbf{U}$, where \mathbf{O} is the set of the *fully observed* variables (with no missing values), while \mathbf{U} is the set of *fully unobserved* variables (the *latent* variables). A missingness graph $\mathcal{M} = (\mathbf{V}^*, \mathbf{E}^*)$ [6] is a causal graph where the vertices in \mathbf{V}^* are partitioned into five disjoint subsets: $\mathbf{V}^* = \mathbf{O} \cup \mathbf{U} \cup \mathbf{M} \cup \mathbf{S} \cup \mathbf{R}$, where \mathbf{M} is the set of the *partially observed* variables, that is, the variables with at least one missing value; \mathbf{S} is the set of the proxy variables, that is, the variables that are actually observed; \mathbf{R} is the set of the *missingness indicators*. Missingness graphs can be queried for independencies using *d-separation*. The set of variables \mathbf{Z} d-separates X from Y , denoted by $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, if it *blocks* every path π between X and Y . A path π is blocked by \mathbf{Z} if and only if π contains: a fork $A \leftarrow B \rightarrow C$ or a chain $A \rightarrow B \rightarrow C$ so that B is in \mathbf{Z} , or, a collider $A \rightarrow B \leftarrow C$ so that B , or any descendant of it, is not in \mathbf{Z} . MCAR, MAR and MNAR result in different independence statements [1] which are linked to the independency statements implied by the missingness graph: MCAR implies $\mathbf{O} \cup \mathbf{U} \cup \mathbf{M} \perp\!\!\!\perp \mathbf{R}$, the missingness is random and independent from the fully observed and the partially observed variables; MAR implies $\mathbf{U} \cup \mathbf{M} \perp\!\!\!\perp \mathbf{R} \mid \mathbf{O}$, missingness is random only conditionally on the fully observed variables; MNAR if neither MCAR nor MAR. Since MCAR implies MAR, any method assuming MAR can be used on MCAR.

3 Multicentric Clinical Data on Endometrial Cancer

The observational data we explore in this paper comprise 763 patients with endometrial cancer from 10 gynecological oncological clinics in Europe that are

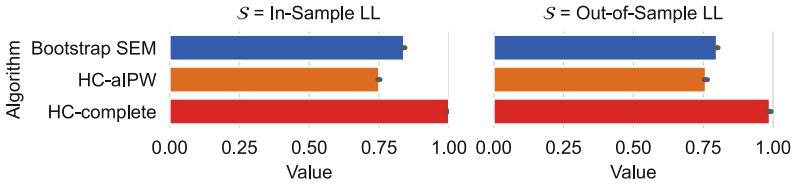


Fig. 1. In-sample and out-of-sample LL for each algorithm. Re-scaled by sample size and absolute maximum value. Lower values are better.

part of the European Network for Individualized Treatment of Endometrial Cancer (ENITEC). Clinical experts selected the variables that they considered most important for predicting the presence of lymph node metastases (LNMs) and survival [4]. The selected variables were: the cytology of the cervix uteri, the preoperative tumour grade, the postoperative tumour grade (after pathological examination of the tumour tissue obtained after surgical removal of the uterus), treatment by chemotherapy or radiotherapy, lymphovascular space invasion (that is, whether there is tumour growth into the lymph or blood vessels), the levels of estrogen and progesterone in blood, the presence of lymph node metastasis according to CT or MRI imaging, the CA125 tumour marker, L1CAM (an intracellular protein that promotes tumour cell motility), the p53 tumour suppressor gene, the number of platelets, presence of lymph node metastases, recurrence of the tumour, and lastly survival before and after 1, 3, and 5 years. The tumour markers (p53, CA125, L1CAM, estrogen and progesterone levels) are thought to offer causal prognostic information about tumour cell behaviour and thus tumour ingrowth, metastases, recurrence, and survival.

4 Experiments

We performed numerical experiments to compare the graphs recovered under MAR and MNAR by the Bootstrap SEM and HC-aIPW. For reference, we also reported the results for HC on data completed with single imputation, denoted *HC-complete*. Prior knowledge elicited from experts consists of forbidden and required edges ($\text{Survival1yr} \rightarrow \text{Survival3yr}$, $\text{Survival3yr} \rightarrow \text{Survival5yr}$).

Firstly, we evaluated the goodness of fit of the recovered graphs by computing the log-likelihood (LL) of both the data used to recover the graph (in-sample) and those held aside for validation (out-of-sample). The former allows us to see which algorithm fits a particular data set the best. The latter approximates the Kullback-Leibler divergence between the recovered causal graph and the unknown causal graph underlying the data, and allows us to see how close the two are and how well the recovered graph generalises to new data. We repeated causal discovery for 100 bootstrap replicates and computed the mean and the standard deviation of both in-sample and out-of-sample LL.

We observe (Fig. 1) that HC-aIPW, which assumes MNAR, dominates Bootstrap SEM and HC-complete, which assume MAR, for both in-sample and out-of-sample LL. In the case of in-sample LL, this may be attributed (at least in

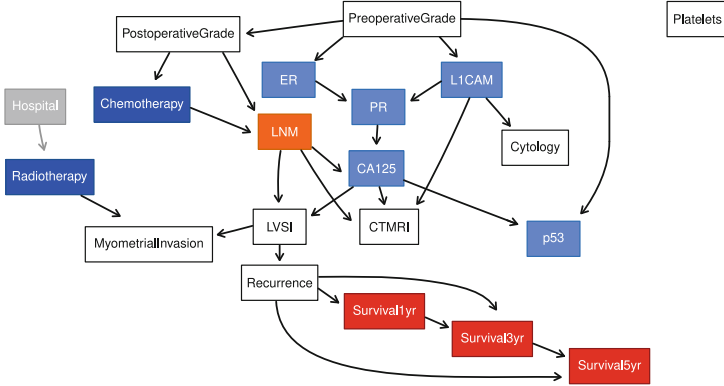


Fig. 2. Causal graph $\mathcal{M}_{\text{MNAR}}$ recovered by HC-aIPW under MNAR.

part) to making the correct missingness assumption: MCAR and MAR are too restrictive and limit how well the recovered graph fits the data; MNAR it is not strict enough when MCAR or MAR hold and let causal discovery algorithms overfit. This decreases the out-of-sample LL because overfitted models are too complex and do not generalize well. The fact that HC-aIPW, which assumes MNAR, outperforms Bootstrap SEM and HC-complete, which assume MAR, for both in-sample and out-of-sample LL suggests that the MNAR assumption is correct for the data and that it allows HC-aIPW to recover a causal graph that is close to the underlying model and that generalises better to new data.

Goodness-of-fit measures allow for a *quantitative* evaluation of the recovered graphs, but they say little about the *qualitative* information they encode in terms of *independence statements*. We denote the graph recovered by Bootstrap SEM as \mathcal{M}_{MAR} ; Fig. 2 shows that recovered by HC-aIPW, $\mathcal{M}_{\text{MNAR}}$. For readability, we colored the vertices depending on their *semantic* interpretation: treatments (Radiotherapy and Chemotherapy) are colored in blue, outcomes (Survival1yr, etc.) in red, the event of interest (LNM) in orange, relevant biomarkers (ER, PR, CA125, etc.) in lightblue and the *context* variable Hospital in grey.

Focusing on the interactions of LNM, we observe using d-separation that $\text{LNM} \perp\!\!\!\perp \{\text{CA125}, \text{p53}\} \mid \text{PostoperativeGrade}$ is true in \mathcal{M}_{MAR} , but false in $\mathcal{M}_{\text{MNAR}}$, where CA125 and p53 are effects of LNM. This makes $\mathcal{M}_{\text{MNAR}}$ close to the clinical practice where both CA125 and p53 are considered relevant biomarkers linked to LNM, providing additional information on LNM even if PostoperativeGrade is observed. Indeed, a crucial difference between \mathcal{M}_{MAR} and $\mathcal{M}_{\text{MNAR}}$ is that the biomarkers and LNM are d-separated from LNM in \mathcal{M}_{MAR} if PostoperativeGrade is observed, but are descendants of LNM in $\mathcal{M}_{\text{MNAR}}$. Hence, if our goal is to detect the presence of LNM in EC patients, measuring CA125, p53 or any of their descendants is coherent with $\mathcal{M}_{\text{MNAR}}$.

Shifting the focus to the treatment variables Chemotherapy and Radiotherapy, $\text{LNM} \perp\!\!\!\perp \text{Chemotherapy}$ does not hold in either \mathcal{M}_{MAR} or $\mathcal{M}_{\text{MNAR}}$, since Chemotherapy is a direct cause of LNM in both. Therefore, Chemotherapy is

expected to influence the likelihood of LNM, which is exactly the reason why it is prescribed by clinicians. On the other hand, LNM $\perp\!\!\!\perp$ Radiotherapy does hold in $\mathcal{M}_{\text{MNAR}}$ but not in \mathcal{M}_{MAR} , suggesting a spurious correlation induced by MAR. This is confirmed by the clinical literature: since Radiotherapy is aimed at local treatment of the tissue surrounding the uterus, and there is a clear dependence with MyometrialInvasion of the tumour (in both models), Radiotherapy effects on LNM are not expected.

5 Discussion and Conclusions

In this work we presented a systematic analysis of the impact of missingness assumptions using state-of-the-art causal discovery algorithms. We applied these methods to a real-world, observational multicentric study on EC patients, extending them to include expert prior knowledge without sacrificing theoretical soundness. Furthermore, we validated the obtained causal models with experienced physicians and clinical literature. We evaluated the goodness-of-fit of the recovered graphs with respect to the underlying data distribution, showing that stricter assumptions are associated to models that generalize poorly. Moreover, by leveraging the test for graphical separation, we explained how the missingness mechanism affects the causal pathways associated to the clinical decision-making perspective. Quantifying the bias due to missingness in other case studies and its overlap with the effects of hidden and selection variables are open problems.

Acknowledgements. Alessio Zanga is funded by F. Hoffmann-La Roche Ltd.

References

1. Rubin, D.B.: Inference and Missing Data. *Biometrika* **63**(3), 581–592 (1976). <https://doi.org/10.1093/biomet/63.3.581>
2. Stavseth, M.R., Clausen, T., Røislien, J.: How handling missing data may impact conclusions: a comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Med.* **7**, 205 (2019). <https://doi.org/10.1177/2050312118822912>
3. Zanga, A., Ozkirimli, E., Stella, F.: A survey on causal discovery: theory and practice. *Int. J. Approximate Reasoning* **151**, 101–129 (2022). <https://doi.org/10.1016/j.ijar.2022.09.004>
4. Zanga, A., Bernasconi, A., Lucas, P.J.F., et al.: Risk assessment of lymph node metastases in endometrial cancer patients: a causal approach. In: *Proceedings of the 1st Workshop on Artificial Intelligence For Healthcare* (2022). <https://ceur-ws.org/Vol-3307/>
5. Liu, Y., Constantinou, A.C.: Greedy structure learning from data that contain systematic missing values. *Mach. Learn.* **111**(10), 3867–3896 (2022). <https://doi.org/10.1007/S10994-022-06195-8>
6. Mohan, K., Pearl, J.: Graphical models for processing missing data. *J. Am. Stat. Assoc.* **116**(534), 1023–1037 (2018). <https://doi.org/10.1080/01621459.2021.1874961>