Risk Assessment of Lymph Node Metastases in Endometrial Cancer Patients: A Causal Approach

Alessio Zanga[1,2], Alice Bernasconi[1,3], Peter J.F. Lucas[4], Hanny Pijnenborg[5], Casper Reijnen[5], Marco Scutari[6] and Fabio Stella[1]

[1]*Department of Informatics, Systems and Communication (DISCo), University of Milano - Bicocca, Milan, Italy*

[2]*Data Science and Advanced Analytics, F. Hoffmann - La Roche Ltd, Basel, Switzerland*

[3]*Evaluative Epidemiology Unit, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*

[4]*University of Twente, Enschede, The Netherlands*

[5]*RadboudUMC, Nijmegen, The Netherlands*

[6]*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland*

Abstract

Assessing the pre-operative risk of lymph node metastases in endometrial cancer patients is a complex and challenging task. In principle, machine learning and deep learning models are flexible and expressive enough to capture the dynamics of clinical risk assessment. However, in this setting we are limited to observational data with quality issues, missing values, small sample size and high dimensionality: we cannot reliably learn such models from limited observational data with these sources of bias. Instead, we choose to learn a causal Bayesian network to mitigate the issues above and to leverage the prior knowledge on endometrial cancer available from clinicians and physicians. We introduce a causal discovery algorithm for causal Bayesian networks based on bootstrap resampling, as opposed to the single imputation used in related works. Moreover, we include a context variable to evaluate whether selection bias results in learning spurious associations. Finally, we discuss the strengths and limitations of our findings in light of the presence of missing data that may be missing-not-at-random, which is common in real-world clinical settings.

Keywords

Causal discovery, Causal networks, Bayesian networks, Missing mechanism, Selection bias

1. Introduction

1.1. Artificial Intelligence in Medicine

State of the Art. Artificial Intelligence (AI) has found many applications in medicine [1] and, more specifically, in cancer research [2] in the form of predictive models for diagnosis [3], prognosis [4] and therapy planning [5]. As a subfield of AI, Machine Learning (ML) and in particular Deep Learning (DL) has achieved significant results, especially in image processing [6]. Nonetheless, ML and DL models have limited explainability [7] because of their black-box

design, which limits their adoption in the clinical field: clinicians and physicians are reluctant to include models that are not transparent in their decision process [8]. While recent research on Explainable AI (XAI) [9] has attacked this problem, DL models are still opaque and difficult to interpret. In contrast, in Probabilistic Graphical Models (PGMs) the interactions between different variables are encoded explicitly: the joint probability distribution P of the variables of interest factorizes according to a graph $\mathcal{G}$, hence the "graphical" connotation. Bayesian Networks (BNs) [10], which we will describe in Section 3.1, are an instance of PGMs that can be used as causal models. In turn, this makes them ideal to use as decision support systems and overcome the limitations of the predictions based on probabilistic associations produced by other ML models [11, 12].

1.2. Lymph Node Metastases in Endometrial Cancer Patients

Background. The present paper focuses on the development of a BN predictive model for endometrial cancer (EC). Endometrial cancer is cancer of the mucous lining, or endometrium, of the uterus. It is a common gynecological disease affecting hundreds of thousands of women worldwide. Although most patients with EC are diagnosed at an early stage of the disease and have a favorable prognosis, approximately 90,000 patients around the world die every year because of EC [13]. Surgery to remove the uterus (hysterectomy), possibly together with the ovaries (ovariectomy), is the typical initial treatment for EC; the choice of neo-adjuvant (pre-surgery) or adjuvant (post-surgery) treatments depends on patient outcome prognosis. The presence of pelvic and/or para-aortic lymph node metastases (LNM) is one of the most important prognostic factors for poor outcome. The identification of LNM during the primary treatment makes it possible to choose a suitable adjuvant treatment and improve survival in node-positive EC [14, 15]. However, no consensus exists on how to determine which patients will benefit from lymphadenectomy (or lymph node dissection): this procedure is usually performed after or concomitant with surgery to evaluate evidence for the spread of cancer, which helps the medical team determine the progress of and treatment options for a patient's malignancy). In clinical early-stage EC, lymphadenectomy has been observed to have a marginal impact on EC outcomes and to be associated with substantial long-term comorbidities. The diagnostic accuracy for LNM is limited: approximately 50% of LNM is found in low- or intermediate-risk patients [16, 17].

Objectives. This work uses the BN model from Reijnen et al. [18] as a starting point to improve the state of the art in two ways:

- Extending the BN model to include the hospital of treatment as an additional variable to detect, estimate and control for potential selection bias.
- Addressing the bias introduced by the missing imputation step, which could induce spurious correlations, hindering the interpretability of the discovered relationships.
- Developing a causal model that integrates domain expert knowledge with observational data to better identify patients with EC designated as low or intermediate risk to develop LNM, in order to support stakeholders for decision-making.

2. Related Work

Individualized treatment aims to minimize unnecessary exposure to therapy-related morbidity and at the same time offers proper management according to patients' risk-stratification. In the context of EC, predicting the risk of LNM before surgical treatment has received limited attention in the literature. Koskas et al. [19] evaluated the performance of BNs models within their cohort of 519 patients. Only one model achieved an AUC greater than 0.75,[1] highlighting the need for improved pre-operative risk stratification. Subsequent works [20, 21, 22] identified biomarkers such as p53 and L1CAM as potential prognostic predictors, together with patients baseline comorbidities and tumors characteristics such as histology, grading and staging. More recently, Reijnen et al. [18] developed a model for the prediction of LNM and of disease-specific survival (DSS) in EC patients. This model, called ENDORISK, is a BN built on clinical, histopathological and molecular biomarkers that can be assessed pre-operatively, allowing for patient *counseling* and shared decision-making before surgery. ENDORISK was shown to be competitive in both goodness of fit and predictive accuracy, achieving AUC values between 0.82 and 0.85 [23].

3. Methods

3.1. Causal Bayesian Networks

Firstly, we will summarize those key definitions for BNs and causal models that we will need to describe our contributions in Section 3.

Definition 1 (Graph). *A graph $\mathcal{G} =$(V, E) is a mathematical object represented by a tuple of two sets: a finite set of nodes V and a finite set of edges $E \subseteq V \times V$. In the following pages (V, E) will be omitted if not specified otherwise.*

We will focus on *directed graphs* where $(X, Y) \neq (Y, X)$, which is graphically represented as $X \to Y$. A directed graph encodes a set of ordinal relationships, i.e. in $X \to Y$ the node X is called *parent* of Y and Y is said to be the *child* of X. Therefore, the set of parents of X is $Pa(X)$, while the set of children of X is $Ch(X)$.

A directed path π is a finite ordered set of nodes $\pi = (V_0 \to \cdots \to V_n)$ such that each adjacent pair of nodes (V_i, V_{i+1}) in π is a directed edge in $\mathbf{E}$. A cycle is a path where the first and the last node are the same node. A graph is acyclic if it contains no cycle, also called a Directed Acyclic Graph (DAG).

Definition 2 (Causal Graph). *A causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ [11] is a graph that encodes the cause-effect relationships of a system.*

Causes & Effects. The set $\mathbf{V}$ contains the variables that describe the behavior of the system under study, whereas the set $\mathbf{E}$ contains the edges that make explicit the interplay of the variables. In particular, for each directed edge $(X, Y) \in \mathbf{E}$, X is said to be a direct cause of Y,

---

[1]The "Area Under the Curve", defined in page 7.

whereas Y is called direct effect of X. This definition is recursive: a variable Z that is the direct cause of X, but not of Y, is said to be an indirect cause of Y.

This mapping between a causal graph $\mathcal{G}$ and the cause-effect relationships is formalized by the causal edge assumption [24].

Definition 3 (Causal Edge Assumption). *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a causal graph. The value assigned to each variable $X \in \mathbf{V}$ is completely determined by the function f given its parents:*

$$X := f(Pa(X)) \qquad \forall X \in \mathbf{V} \tag{1}$$

The causal edge assumption allows us to interpret the edges of a causal graph in a non-ambiguous way: it enforces a recursive relationship over the structure of the graph, establishing a chain of functional dependencies. Hence, this class of graphical models is inherently explainable, even for researchers approaching them for the first time.

When the causal graph is not known a priori, it is possible to recover it from a combination of prior knowledge and data driven approaches. Such problem is called Causal Discovery [25].

Definition 4 (Causal Discovery). *Let $\mathcal{G}^*$ be the true but unknown graph in the space of possible graphs $\mathbb{G}$ from which the data set $\mathbf{D}$ has been generated. The Causal Discovery problem consists in recovering $\mathcal{G}^*$ given the data set $\mathbf{D}$ and the prior knowledge $\mathbf{K}$.*

Once the causal graph $\mathcal{G}^*$ is recovered, it is possible to build a PGM with the given structure. For example, BNs [10] are a widely known type of PGM.

Definition 5 (Bayesian Network). *Let be $\mathcal{G}$ a DAG and let $P(\mathbf{X})$ be a global probability distribution with parameters Θ. A BN $\mathcal{B} = (\mathcal{G}, \Theta)$ is a model in which each variable of $\mathbf{X}$ is a vertex of $\mathcal{G}$ and $P(\mathbf{X})$ factorizes into local probability distributions according to $\mathcal{G}$:*

$$P(\mathbf{X}) = \prod_{X \in \mathbf{X}} P(X \mid Pa(X)) \tag{2}$$

The key difference between a BN and a Causal BN (CBN) is the semantic interpretation of its edges. Indeed, in a CBN an edge represents a cause-effect relationship between two variables, whereas the same edge in a BN entails only a probabilistic dependence.

Definition 6 (Causal Bayesian Network). *A Causal BN $\mathcal{B} = (\mathcal{G}, \Theta)$ is a BN where the associated DAG $\mathcal{G}$ is a causal graph.*

3.2. Causal Discovery with Observational and Missing Data

Causal discovery algorithms are usually divided into two classes: constraint-based and score-based. The two classes have been extended to handle missing data in different ways: constraint-based algorithms rely on test-wise deletion [26] to perform conditional independence tests efficiently in order to mitigate the impact of missing observations, while score-based approaches alternate data imputation and causal discovery [27].

Causal Discovery with Missing Data. By default, causal discovery algorithms are not designed to handle incomplete data. However, we can combine them with missing value imputation approaches to complete the data and reduce the problem to a standard causal discovery. A widely-used application of this idea is the Expectation Maximization (EM) [28] algorithm. In particular, the Structural EM [27] algorithm is specifically designed to iteratively run the imputation step performed by EM and a causal discovery step performed by a score-based algorithm, alternating them until convergence.

Greedy Search: The Hill-Climbing Approach. A widely applied score-based algorithm for causal discovery is Greedy Search (GS) [29]. GS traverses the space $\mathbb{G}$ of the possible DAGs over the set of variables $\mathbf{V}$, selecting the optimal graph $\mathcal{G}^*$ by a greedy evaluation of a function $\mathcal{S}$, known as the *scoring criterion*. There are multiple strategies to implement GS, one of which is called Hill-Climbing (HC). At its core, HC repeatedly applies three fundamental operations to change the current recovered structure, moving from a graph to another, across the graphs space $\mathbb{G}$. These "moves" are the addition, deletion or reversal of an edge. If a move improves the score $\mathcal{S}$, then the graph is updated accordingly. The procedure halts when no moves improve the score and returns a DAG.

While the graphs space $\mathbb{G}$ contains every graph that could be generated given the vertices $\mathbf{V}$, only a subset of them are compatible with the probability distribution induced by the observed data. Moreover, not every graph compatible with said distribution is necessarily causal. Therefore, it is possible to shrink the search space by adding constraints in terms of structural properties, that is, by requiring or forbidding the existence of an edge in the optimal graph $\mathcal{G}^*$.

Encoding Prior Knowledge. One could restrict the set of admissible graphs by encoding prior knowledge through required or forbidden edge lists [30]. For instance, it is possible to leverage expert knowledge to identify known relationships and encode them as required edges. These lists can also encode a partial ordering when potential causes of other variables are known.

For example, suppose that clinicians want to include their prior knowledge on the interaction between biomarkers and LNM into the CBN. This inclusion would happen during the execution of the causal discovery algorithm and, therefore, requires that the experts' knowledge is encoded programmatically. Causal discovery algorithms essentially learn a set of ordinal, parent-child relationships: it is natural to encode prior knowledge in the same form. For instance, if we know that p53 is not a direct cause of LNM, then the translation of such a concept would be p53 $\notin Pa(\text{LNM})$. If, on the other hand, we know that LNM is a direct cause of L1CAM then we would have L1CAM $\in Pa(\text{LNM})$. This is a direct consequence of the Causal Edge Assumption (Definition 3). Even this simple example shows the flexibility of this approach, allowing to encode different sources of prior knowledge without any restrictions.

4. Experimental Results

Causal discovery algorithms provide a correct solution to the causal discovery problem in the limit of the number of samples [31]. However, in real-world applications the available data are finite, especially in medicine, where data samples are usually small. As a result, even small amounts of noise in the data may result in a different structure. Therefore, it is important to quantify our confidence in the presence of each edge in the causal BN, also called the "strength" of an edge.

Estimating Edge Strength: A Bootstrap Approach. The estimation of the strength of an edge was performed through a bootstrap approach [32]. Here, a custom version with Structural EM is reported in Algorithm 1, described as follows. Line 1, the procedure takes as input a data set $\mathbf{D}$, prior knowledge $\mathbf{K}$, hyperparameters $\boldsymbol{\alpha}$ for Structural EM, number of bootstraps n and number of samples to draw m. Line 2, the confidence matrix $\mathbf{C}$ is initialized. Lines 3-6, the data set $\mathbf{D}$ is re-sampled n times with replacement, drawing m observations for each bootstrap following a uniform distribution. For each sampled data set $\mathbf{D}_i \subseteq \mathbf{D}$, the causal discovery algorithm is applied to induce a corresponding graph $\mathcal{G}_i$. Finally, line 7, is responsible to compute the strength of each edge as the relative frequency of inclusion across the n bootstraps.

The causal discovery algorithm developed is described in Algorithm 2. Line 1 is based on the confidence matrix estimation computed by Algorithm 1. Line 2, the causal graph $\mathcal{G}$ is initialized to the empty graph and, line 3, the associated confidence matrix $\mathbf{C}$, i.e. the matrix containing the edges strength, is computed. Line 4 describes a generic strategy to select the edges to insert into $\mathcal{G}$ given $\mathbf{C}$. Here, we relied on a threshold λ to filter irrelevant edges to build the "average graph". Lines 5-6, the CBN parameters Θ are fitted given $\mathcal{G}$ by applying EM [28] to the data set $\mathbf{D}$ with missing data.

Definition and Selection of Variables. To conduct this analysis we used the cohort presented by Reijnen et al. An overview of the cohort and the procedures done for data collection can be found in [18]. Briefly, the retrospective multicenter cohort study included 763 patients, with a median age 65 years, surgically treated for endometrial cancer between 1995 and 2013 at one of the 10 participating European hospitals. Clinical and histopathological variables with prognostic value for the prediction of LNM were identified by a systematic review of the literature. The used variables could be divided into three major temporal tiers:

- *Pre-operative clinical, histopathological variables and biomarkers*: Estrogen Receptor (ER) expression, Progesteron Recepter (PR) expression, L1CAM (cell migration) expression, p53 (tumour suppressor gene) expression, cervical cytology, platelets counts (thrombocytosis), lymphadenopathy on MRI or CT, lymphovascular space invasion (LVSI), Ca-125 serum levels and pre-operative tumor grade,
- *Post-operative/treatment variables:* adjuvant therapy (Chemotherapy and/or Radiotherapy), post-operative tumour grade,
- *Late post-operative outcomes:* 1-,3-,5-year disease-specific survival (DSS), Lymph Nodes Metastases (LNM), Myometrial Invasion.

All the described variables are discrete variables, with cardinality ranging from 2 to 3. Two main changes were done in comparison to published works: addition of hospital of treatment (10 levels) in the model and separation of adjuvant therapy into two different dichotomous variables (chemotherapy and radiotherapy).

Training and Testing. The data set $\mathbf{D}$ was split in a train set and a test set following a 70/30 ratio. For each configuration of hyperparameters $(\boldsymbol{\alpha}, n, m, \lambda)$, we applied Algorithm 2 to the train set, with the same prior knowledge $\mathbf{K}$. The resulting BNs were evaluated on the test set by estimating the probability of LNM. The hyperparameter tuning was performed following a grid search, as suggested in [31]. While *cross validation* (CV) is generally preferred over a naïve train-test splitting, hyperparameter tuning over a learning procedure based on *Structural EM* is computationally expensive and, therefore, it would require a nonignorable amount of time when coupled with CV. Moreover, we considered the possibility to further split the train set to obtain a *validation set*, but the reduced sample size hindered the feasibility of this additional step. Finally, we computed the sensitivity, specificity, ROC and AUC[2] for each CBN model.

Definition 7 (Sensitivity & Specificity). *Given a binary classification problem, the confusion matrix is a 2×2 squared integer matrix resulting from the application of a classification algorithm. The values on the main diagonal are called true positives (TP) and true negatives (TN), while the values on the off diagonal are false positives (FP) and false negatives (FN). Then, the true positive ratio (TPR) and the true negative ratio (TNR) are defined as follows:*

$$TPR = \frac{TP}{TP + FN} \qquad (3) \qquad\qquad TPR = \frac{TN}{TN + FP} \qquad (4)$$

The TPR and TNR are also called sensitivity and specificity, respectively.

Definition 8 (ROC & AUC). *The Receiving Operating Characteristic (ROC) curve is a plot of sensitivity and (1 - specificity) measures at different thresholds. The Area Under the Curve (AUC) is the area under the ROC curve.*

Algorithm 1 Confidence matrix from missing data and prior knowledge.

1: **procedure** CONFIDENCEMATRIX($\mathbf{D}, \mathbf{K}, \boldsymbol{\alpha}, n, m$)
2:     $\mathbf{C} \leftarrow \mathbf{0}$ ▷ Initialize a $|\mathbf{V}| \times |\mathbf{V}|$ matrix, with $\mathbf{V}$ the variables in $\mathbf{D}$.
3:     **for** $i \in [1, n]$ **do**
4:         $\mathbf{D}_i \leftarrow$ SAMPLE($\mathbf{D}, m$)   ▷ Sample from $\mathbf{D}$ with replacement.
5:         $\mathcal{G}_i \leftarrow$ STRUCTURALEM($\mathbf{D}_i, \mathbf{K}, \boldsymbol{\alpha}$)   ▷ Learn $\mathcal{G}_i$ from $\mathbf{D}_i$ and $\mathbf{K}$.
6:         $\mathbf{C}[X, Y] \leftarrow \mathbf{C}[X, Y] + 1, \quad \forall (X, Y) \in \mathbf{E}_i$ ▷ Increment the edge count.
7:     $\mathbf{C} \leftarrow \mathbf{C}/n$ ▷ Normalize the confidence matrix.
8:     **return** $\mathbf{C}$

---

[2]We are aware of the issues related to the selection of a *causal* model based on its classification performances, hence, we took into account both in-sample and out-of-sample metrics.

Algorithm 2 Learn CBN from missing data and prior knowledge.

1: **procedure** CBN($\mathbf{D}, \mathbf{K}, \boldsymbol{\alpha}, n, m, \lambda$)
2:      $\mathcal{G} \leftarrow (\mathbf{V}, \varnothing)$ ▷ Initialize an empty graph over the variables $\mathbf{V}$ in $\mathbf{D}$.
3:      $\mathbf{C} \leftarrow \text{ConfidenceMatrix}(\mathbf{D}, \mathbf{K}, \boldsymbol{\alpha}, n, m)$ ▷ Compute the confidence matrix.
4:      *Insert edges into $\mathcal{G}$ following a strategy w.r.t. $\mathbf{C}$ and λ.*
5:      $\hat{\Theta} \leftarrow \text{EM}(\mathcal{G}, \mathbf{D})$ ▷ Estimate the parameters using EM.
6:      $\mathcal{B} \leftarrow (\mathcal{G}, \hat{\Theta})$ ▷ Build the CBN given $\mathcal{G}$ and $\hat{\Theta}$.
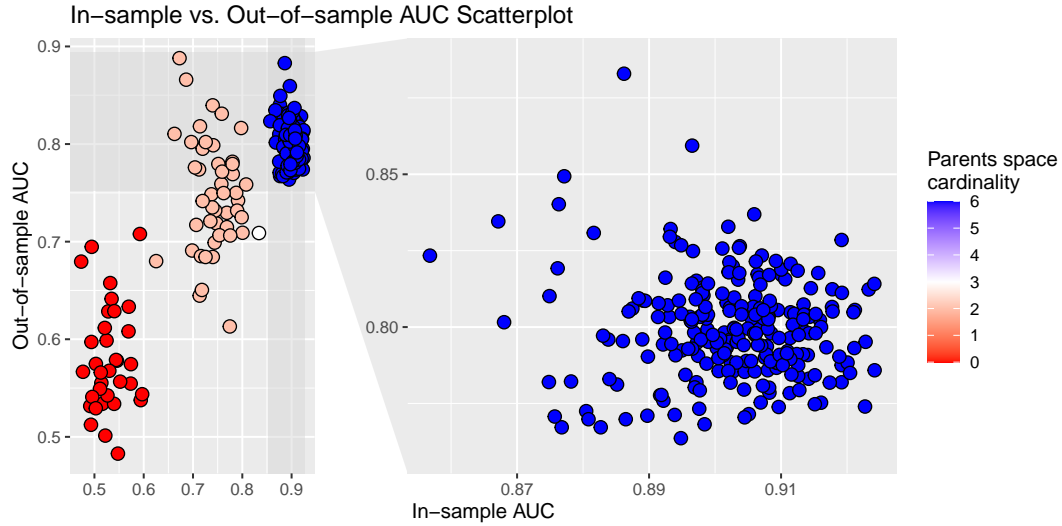7:   **return** $\mathcal{B}$

Figure 1: Scatter plot of the results of Algorithm 2. Each dot is a CBN with achieved in-sample and out-of-sample AUC on the horizontal and vertical axes, respectively. Dots color depend on the cardinality of the space of the parents of the target node LNM. To the right, a zoom of the cluster of those CBN that achieved higher values of AUC.

Figure 1 is a scatter plot of the results of the execution of Algorithm 2. The color mapping allows to clearly distinguish three well-separated clusters, grouped by the parents space cardinality. Specifically, the red cluster represents the models where *LNM* has no parents, the light-red cluster contains models where *LNM* has only *Chemotherapy* as parent, and finally the blue cluster where both *Chemotherapy* and *Histology* are parents of *LNM*.

The structure presented in Figure 2 is built by encoding prior causal knowledge elicited by clinicians and randomized controlled trials (RCTs). The encoding process is performed by adding a directed edge from the expected cause to its effect. Each edge addition is supported by biological and physiological knowledge, either obtained by querying experts or from reviewed literature, without observational data. Note, for example, that the *Therapy* node does not have incoming edges, since therapy is always assigned at random in an RCT (and only the outcome matters).

The graph presented in Figure 3 is the result of the application of Algorithm 2 on the collected data set and encoded prior knowledge based on partial temporal ordering of variables. The
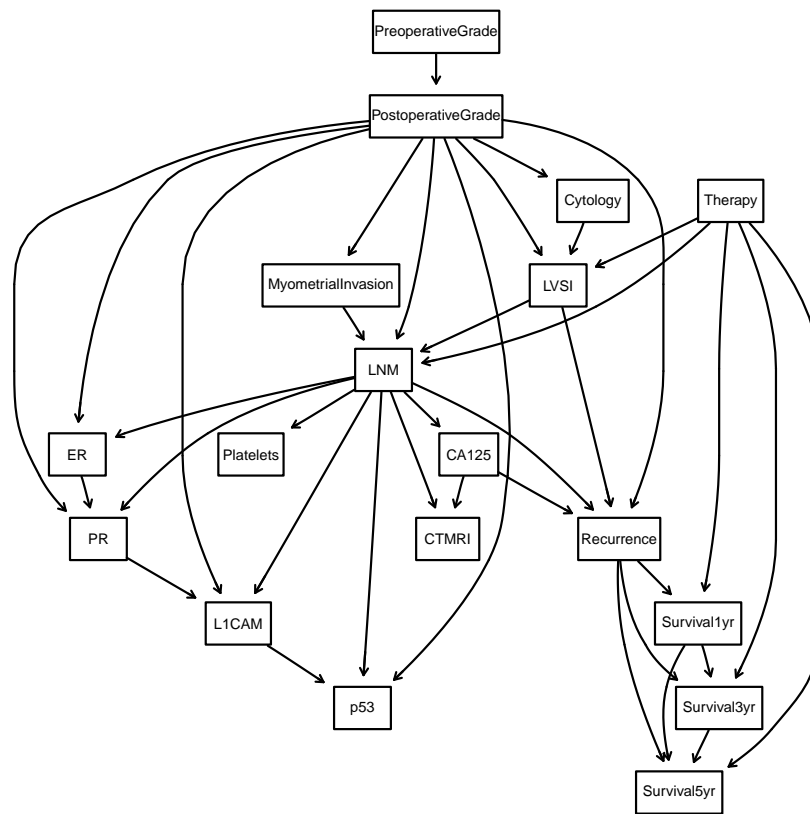
Figure 2: Reference graph obtained by experts' knowledge and randomized controlled trial.

Therapy node is split into *Radiotherapy* and *Chemotherapy* to highlight the different impact of adjuvant treatments.

The two graphs share a common subset of edges, e.g. the ones related to *Recurrence* and *Survivals*. A major difference stands in the edges related to the biomarkers cluster. Indeed, while in Figure 2 biomarkers, such as *p53*, *CA125* and *L1CAM*, are assumed to be strongly related to *LNM*, in the recovered graph the *PreoperativeGrade* is observed as common parent of the variables contained in such cluster. Moreover, no biomarker is directly connected to *LNM*, not as a parent nor as a child, calling for further analyses of the collected data.

Such similarities and differences also appear in Figure 4, where *Hospital* is introduced to explore the potential presence of latent effects and selection bias. While the graph in Figure 4 is not completely different from the one in Figure 3 in terms of observed substructures, the latter encodes different independence statements due to the presence of the newly introduced *Hospital*.

The crucial difference stands in the semantic interpretation of Hospital, which in this case is not to be intended as a direct cause of its children, but rather as a proxy for others unobserved variables or biases, i.e. a context variable. Indeed, while it could be that population heterogeneity across hospitals affects the choice of adjuvant treatments, it would be nonsensical to conclude
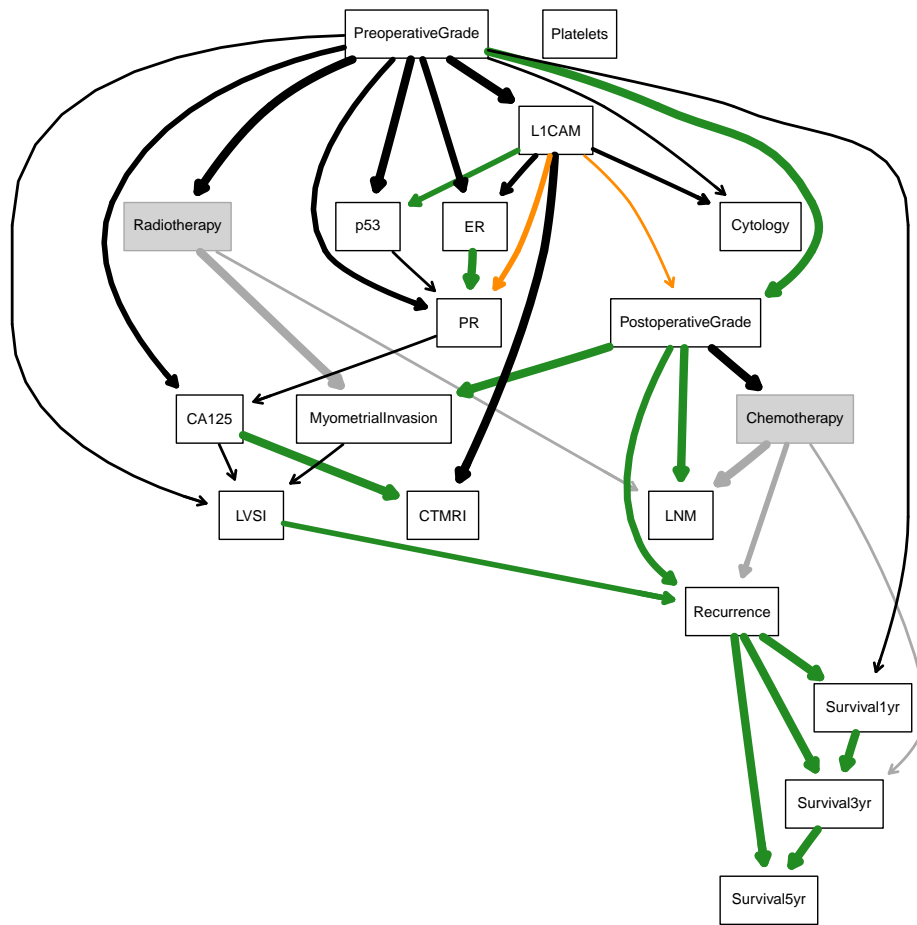
Figure 3: Strength plot of recovered CBN. The edges thickness depends on the strength of the edge itself, which is estimated by the confidence matrix **C**. The nodes and edges are colored to ease the comparison with the reference graph Figure 2. In particular, green edges are present both in reference and recovered graph with the same orientation, orange edges are present both in reference and recovered graph with reversed orientation, gray nodes and edges cannot be directly compared due to different node sets, and, finally, black edges are present only in the recovered graph.

that Hospital is a cause of Ca-125. Nonetheless, the causal discovery procedure includes a set of edges that are related to spurious associations present in the data set. For example, the directed edge that connects Hospital to p53 is an instance of such pattern, which could be caused by a missing-not-at-random (MNAR) mechanism [33]. Another example of the impact of biases is represented by the directed edge from Hospital to PostoperativeGrade. In this case, an unbalanced distribution of patients' grading across geographical regions, which Hospital is a proxy of, could act as a potential source of selection bias [34].

The ROC curve depicted in Figure 5 is obtained by predicting the probability of the LNM class on the test set, given the CBN fitted on the structure in Figure 4 and the train set. It achieves an AUC of 0.883, with associated 95% CI 0.775-0.991, which is higher than the one obtained in
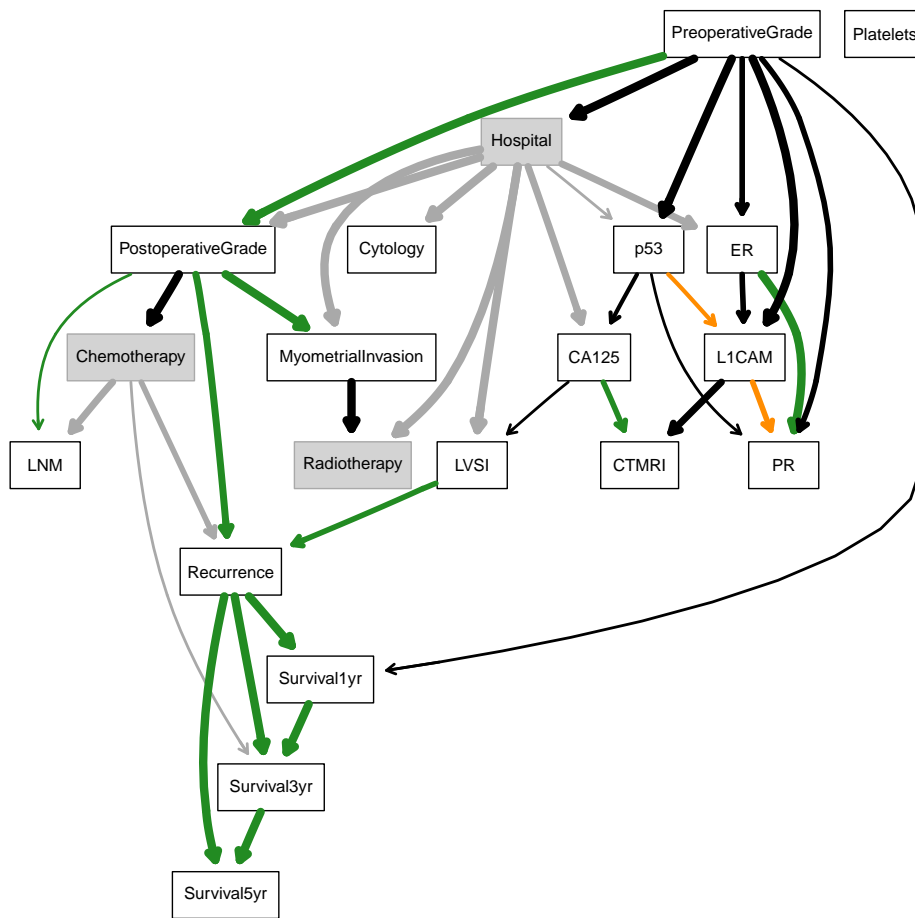
Figure 4: Strength plot of recovered CBN with the addition of the previously unobserved Hospital variable. The edges thickness and the coloring schema are the same of Figure 3.

[18], although it was not possible to compare the metrics using a significance test due to the different test sets.

5. Conclusions and Future Works

Given the known limitations of data-driven approaches when applied to observational data, causal discovery techniques are used to explore and mitigate the impact of spurious associations during the learning process. In this work we explored the task of learning a causal representation to assess the pre-operative risk of developing LNM in endometrial cancer patients. Furthermore, the recovered models were extended to include information from context variables, aiming to uncover previously unobserved effects.

The resulting procedure takes advantage of pre-existing techniques to reduce the bias introduced during the imputation step in a bootstrap approach. This enabled us to compute
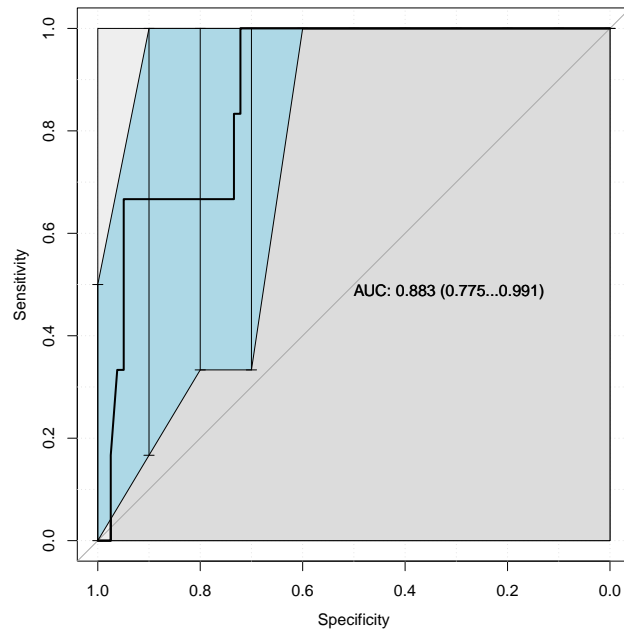
Figure 5: ROC curve and AUC value of the CBN model fitted from the graph in Figure 4 and train data.

the strength of the observed associations in the obtained models across multiple re-sampled instances, allowing a step of model averaging to recover less frequent substructures. The risk assessment is performed by predicting the probability of developing LNM using a CBN fitted on the recovered structure and given train set, showing an increased AUC over previous works.

Still, we highlighted a set of potential issues that need to be addressed in future works.

Missingness Mechanism.   With the introduction of the Hospital variable we observed a set of edges that hint to the presence of a potential missing-not-at-random pattern. If this is the case, then it would require careful consideration in order to reduce the bias introduced during the missing imputation step.

Effect of Adjuvant Therapy.   Once a causal graph is obtained, it is theoretically possible to estimate the causal effect of each adjuvant therapy, either single or combined, on the development of LNM. Before directly computing the the effect, there are assumptions that need to be carefully verified, e.g. positivity, consistency, unconfoundedness and non interference [24].

Impact of Selection Bias.   While it is clear that observing an association between Hospital and other variables it is not sufficient to conclude that, indeed, there is a selection bias, it is a strong hint that there are other unobserved variables that influence the causal mechanism. It could be interesting to assess which is the impact of the selection bias mediated by the Hospital variable alone.

Acknowledgments

References

[1] V. Kaul, S. Enslin, S. A. Gross, History of artificial intelligence in medicine, Gastrointestinal Endoscopy 92 (2020) 807–812. doi:`10.1016/J.GIE.2020.06.040`.

[2] O. Troyanskaya, Z. Trajanoski, A. Carpenter, S. Thrun, N. Razavian, N. Oliver, Artificial intelligence and cancer., Nature cancer 1 (2020) 149–152. URL: http://www.ncbi.nlm.nih.gov/pubmed/35122011. doi:`10.1038/s43018-020-0034-6`.

[3] S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges, Cancer Letters 471 (2020) 61–71. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304383519306135. doi:`10.1016/j.canlet.2019.12.007`.

[4] O. Elemento, C. Leslie, J. Lundin, G. Tourassi, Artificial intelligence in cancer research, diagnosis and therapy, Nature Reviews Cancer 21 (2021) 747–752. URL: https://www.nature.com/articles/s41568-021-00399-1. doi:`10.1038/s41568-021-00399-1`.

[5] D. Ho, Artificial intelligence in cancer therapy., Science (New York, N.Y.) 367 (2020) 982–983. URL: http://www.ncbi.nlm.nih.gov/pubmed/32108102. doi:`10.1126/science.aaz3023`.

[6] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, R. H. Mak, R. M. Tamimi, C. M. Tempany, C. Swanton, U. Hoffmann, L. H. Schwartz, R. J. Gillies, R. Y. Huang, H. J. W. L. Aerts, Artificial intelligence in cancer imaging: Clinical challenges and applications, CA: A Cancer Journal for Clinicians (2019) caac.21552. URL: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21552. doi:`10.3322/caac.21552`.

[7] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019) e1312. URL: https://onlinelibrary.wiley.com/doi/10.1002/widm.1312. doi:`10.1002/widm.1312`.

[8] L. Pumplun, M. Fecho, N. Wahl, F. Peters, P. Buxmann, Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study, Journal of Medical Internet Research 23 (2021) e29301. URL: https://www.jmir.org/2021/10/e29301. doi:`10.2196/29301`.

[9] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science Robotics 4 (2019). URL: https://www.science.org/doi/10.1126/scirobotics.aay7120. doi:`10.1126/scirobotics.aay7120`.

[10] J. Pearl, S. Russell, BAYESIAN NETWORKS, Technical Report, 2003.

[11] E. Bareinboim, J. D. Correa, D. Ibelind, T. Icard, On Pearl's Hierarchy and the Foundations

of Causal Inference, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2020, pp. 1–62. URL: https://causalai.net/r60.pdf.

[12] S. Lee, E. Bareinboim, Structural causal bandits: Where to intervene?, in: Advances in Neural Information Processing Systems, volume 2018-Decem, 2018, pp. 2568–2578.

[13] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 68 (2018) 394–424. doi:10.3322/caac.21492.

[14] S. M. de Boer, M. E. Powell, L. Mileshkin, et al., Adjuvant chemoradiotherapy versus radiotherapy alone in women with high-risk endometrial cancer (portec-3): patterns of recurrence and post-hoc survival analysis of a randomised phase 3 trial, The Lancet Oncology 20 (2019) 1273–1285. doi:10.1016/S1470-2045(19)30395-X.

[15] D. Matei, V. Filiaci, M. E. Randall, et al., Adjuvant chemotherapy plus radiation for locally advanced endometrial cancer, New England Journal of Medicine 380 (2019) 2317–2326. doi:10.1056/NEJMoa1813181.

[16] S. Bendifallah, G. Canlorbe, P. Collinet, E. Arsène, F. Huguet, C. Coutant, D. Hudry, O. Graesslin, E. Raimond, C. Touboul, E. Daraï, M. Ballester, Just how accurate are the major risk stratification systems for early-stage endometrial cancer?, British Journal of Cancer 112 (2015) 793–801. doi:10.1038/bjc.2015.35.

[17] J. Trovik, E. Wik, H. M. Werner, C. Krakstad, H. Helland, I. Vandenput, T. S. Njolstad, I. M. Stefansson, J. Marcickiewicz, S. Tingulstad, A. C. Staff, F. Amant, L. A. Akslen, H. B. Salvesen, Hormone receptor loss in endometrial carcinoma curettage predicts lymph node metastasis and poor outcome in prospective multicentre trial, European Journal of Cancer 49 (2013) 3431–3441. doi:10.1016/j.ejca.2013.06.016.

[18] C. Reijnen, E. Gogou, N. C. Visser, H. Engerud, J. Ramjith, L. J. Van Der Putten, K. Van De Vijver, M. Santacana, P. Bronsert, J. Bulten, M. Hirschfeld, E. Colas, A. Gil-Moreno, A. Reques, G. Mancebo, C. Krakstad, J. Trovik, I. S. Haldorsen, J. Huvila, M. Koskas, V. Weinberger, M. Bednarikova, J. Hausnerova, A. A. Van Der Wurff, X. Matias-Guiu, F. Amant, L. F. Massuger, M. P. Snijders, H. V. Kusters-Vandevelde, P. J. Lucas, J. M. Pijnenborg, Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study, PLoS Medicine 17 (2020). doi:10.1371/journal.pmed.1003111.

[19] M. Koskas, M. Fournier, A. Vanderstraeten, F. Walker, D. Timmerman, I. Vergote, F. Amant, Evaluation of models to predict lymph node metastasis in endometrial cancer: A multicentre study, European Journal of Cancer 61 (2016) 52–60. doi:10.1016/j.ejca.2016.03.079.

[20] G. Getz, S. B. Gabriel, K. Cibulskis, et al., Integrated genomic characterization of endometrial carcinoma, Nature 497 (2013). doi:10.1038/nature12113.

[21] F. K. Kommoss, A. N. Karnezis, F. Kommoss, A. Talhouk, F. A. Taran, A. Staebler, C. B. Gilks, D. G. Huntsman, B. Krämer, S. Y. Brucker, J. N. McAlpine, S. Kommoss, L1cam further stratifies endometrial carcinoma patients with no specific molecular risk profile, British Journal of Cancer 119 (2018). doi:10.1038/s41416-018-0187-6.

[22] L. J. V. D. Putten, N. C. Visser, K. V. D. Vijver, M. Santacana, P. Bronsert, J. Bulten, M. Hirschfeld, E. Colas, A. Gil-Moreno, A. Garcia, G. Mancebo, F. Alameda, J. Trovik,

R. K. Kopperud, J. Huvila, S. Schrauwen, M. Koskas, F. Walker, V. Weinberger, L. Minar, E. Jandakova, M. P. Snijders, S. V. D. B.-V. Erp, X. Matias-Guiu, H. B. Salvesen, F. Amant, L. F. Massuger, J. M. Pijnenborg, L1cam expression in endometrial carcinomas: An enitec collaboration study, British Journal of Cancer 115 (2016). doi:10.1038/bjc.2016.235.

[23] P. Vinklerová, P. Ovesná, J. Hausnerová, J. M. A. Pijnenborg, P. J. F. Lucas, C. Reijnen, S. Vrede, V. Weinberger, External validation study of endometrial cancer preoperative risk stratification model (endorisk), Frontiers in Oncology 12 (2022). doi:10.3389/fonc.2022.939226.

[24] J. Pearl, M. Glymour, N. P. Jewell, Causal inference in statistics: A primer, John Wiley \& Sons, 2016.

[25] A. Zanga, E. Ozkirimli, F. Stella, A Survey on Causal Discovery: Theory and Practice, International Journal of Approximate Reasoning 151 (2022) 101–129. doi:10.1016/J.IJAR.2022.09.004.

[26] E. V. Strobl, S. Visweswaran, P. L. Spirtes, Fast causal inference with non-random missingness by test-wise deletion, International Journal of Data Science and Analytics 6 (2018) 47–62. doi:10.1007/S41060-017-0094-6.

[27] N. Friedman, The Bayesian Structural EM, Proceedings of the Fourteenth Conference on Uncertainty and Artificial Intelligence (1998). URL: http://www.cs.huji.ac.il/~nir/Papers/Fr2.pdf.

[28] S. L. Lauritzen, The EM algorithm for graphical association models with missing data, Computational Statistics and Data Analysis 19 (1995). doi:10.1016/0167-9473(93)E0056-A.

[29] M. Scutari, C. Vitolo, A. Tucker, Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation, Statistics and Computing 29 (2019) 1095–1108. URL: https://doi.org/10.1007/s11222-019-09857-1. doi:10.1007/s11222-019-09857-1.

[30] C. Meek, Strong Completeness and Faithfulness in Bayesian Networks (2013). URL: http://arxiv.org/abs/1302.4973.

[31] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, Causation, prediction, and search, MIT press, 2000.

[32] N. Friedman, M. Goldszmidt, A. Wyner, Data Analysis with Bayesian Networks: A Bootstrap Approach (2013). URL: http://arxiv.org/abs/1301.6695.

[33] M. Scutari, Bayesian network models for incomplete and dynamic data (2020). doi:10.1111/stan.12197.

[34] K. M. Esterling, D. Brady, E. Schwitzgebel, The Necessity of Construct and External Validity for Generalized Causal Claims, OSF Preprints (2021) 1–41. URL: https://osf.io/2s8w5/.