Universiteit Leiden

Master Computer Science

A Clinical Prognostic Decision Support System for the Prediction of the Prognosis of Patients with Endometrial Cancer

Name: Evangelia Gogou
Student ID: s1790420

Date: 26/10/2020

1st supervisor: Prof.dr. P.J.F. Lucas
2nd supervisor: Prof.dr.ir. F.J. Verbeek

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Achieving an accurate prognosis of patients with endometrial cancer is a complicated and demanding process that involves multiple steps. It is literally about predicting the future and therefore graphical models that allow oncologists to reason about patients outcome can be extremely helpful.

Bayesian networks (BNs) are graphical probability distributions that represent (causal and other) relationships among the relevant variables and can provide more insights about a disease and its treatment to clinicians. BNs determine the joint probability distributions that are formulated in the form of networks. These networks consist of nodes that represent the participating variables. The models have the potential to assist medical experts in causal reasoning in addition to risk prediction for patients, in our case concerning outcome of endometrial cancer treatment. In this specific domain, a list of biomarkers have been found to be correlated with lymph node metastasis, but they are not used in practice to predict and reason about a patient's survival. The present thesis focuses on unveiling how Bayesian networks can incorporate such variables and be used in order to build prognostic decision support models, employing a combination of clinical background knowledge and structural machine learning algorithms. In particular, it is investigated whether it is possible to develop a reliable prognostic Bayesian network, that gynecological oncologists can use preoperatively to assess patients with endometrial cancer. In addition, a comparison is made with a Cox proportional hazards model (CPH), which is among the most well-known regression techniques used for survival analysis.

The present study is based on a dataset of 763 patients of median age of 65 years, who underwent surgery for their treatment between February 1995 and August 2013. The BN implementation process integrates expert knowledge with score-based structure learning algorithm results. In detail, the network includes different survival variables and multiple types of preoperative, histopathological, molecular biomarkers. The ENDORISK (preoperative risk stratification in endometrial cancer) BN model has developed in close collaboration with clinicians of the Department of Obstetrics and Gynaecology of Radboud University Medical Center and acts as the basis of the research paper that has been published in the PLOS Medicine Journal (May 2020) [51]. It consists of preoperative tumor grade, immunohistochemical expression of estrogen receptor (ER), progesterone receptor (PR), p53, L1 cell adhesion molecule (L1CAM), cancer antigen 125 serum lever (Ca-125), thrombocyte count, cervival cytology and imaging results variables. Bootstrapping is also used to assess the strength of the connections among the participating nodes. The comparison with a Cox regression model reveals that the Bayesian model has many advantages over the CPH model and it can be actively used by the oncologists, facilitating the decision-making process and catering for each individual with endometrial cancer. These intuitive models are easy to grasp and can be updated effortlessly based on new information. They have the ability to transform to decision models and allow their users to reason about the patient's condition based on updated evidence.

Keywords: artificial intelligence, Bayesian networks, decision models, prognosis, machine learning, Cox proportional hazards models, survival analysis, prediction, structure learning algorithms, score-based algorithms, endometrial cancer, preoperative assessment.

Acknowledgement

First of all I would like to express my sincere thank you to my supervisor, professor dr. P.J.F. Lucas, for his extensive support and great patience during all this time. I deeply appreciate you as a person, and value you as a professor. By always demonstrating motivating passion for your work, you instilled in me the persistence to continue working on my thesis despite the difficulties I encountered in a personal level. I am grateful for always taking the time and energy to react to my questions and showing that you really care. Your constructive feedback pushed me to strive for a better result and I am thankful for the opportunity to work with you on such an interesting topic, giving me this way the chance to collaborate with medical doctors and contribute to their research. Without your consistent advice and guidance I would not have had the honor and pleasure to be part of this published research paper, before finalizing my studies.

I would also like to thank Casper Reijnen and Johanna M. A. Pijnenborg for the insightful meetings and for explaining thoroughly to me the relevant medical concepts. You were extremely helpful not only at the beginning of my thesis but also later on.

Additionally, I would like to express my sincere gratitude to my family. I deeply thank my parents, Dimitrios and Agapi, who helped me move to the Netherlands and pursue my master degree. Together with my sweet sister and friend Anna, they supported me all the way. I considered myself utterly lucky and blessed for having you in my life, always being by my side and encouraging me to try my best with patience.

I would also like to deeply thank my boyfriend, Dimitrios, for the positive influence he has had on me and his great support during my studies. I have been always inspired by your tremendous persistence when striving for your goals. Thank you so much for urging me to focus on what I need to achieve for my growth.

Last but not least, I would like to thank my dearest friend Mariela. I always feel more confident with you next to me. Thank you deeply for your kindness, thoughtfulness and encouragement all this time.

Contents

1 Introduction

1.1 The role of prognostic models in medicine

Prognostic models in medicine play an important role in estimating the clinical outcome of treatment for a particular disease well in advance. In developing such models, knowledge of the functioning of the human body is crucial and much of this knowledge is causal in nature. The process of developing a prognostic model is focused on unveiling the variables that jointly impact the outcome of a disease in a particular direction. Medical doctors have always been fascinated by how information from several patients can be incorporated into a single prognostic model and effectively support them in predicting a patient's health situation in the future.

A variety of factors, such as a person's medical history, age, gender, results of a physical examination, and laboratory test results are used as variables in these models. It is highly significant that these models consist of variables that are easy to be measured. This way, it can be ensured that physicians can use them in practice. As a process, the prognosis is based on the diagnosis, the provided treatment, and the doctor's skills.

Prognostic models are divided in *population-level models* and *patient-level models*. The former category reflects common trends based on specified attributes for groups of people, while the latter category consists of models focusing on generating the necessary information to come up with proper therapy and patient-centric advice [2]. This way, the evaluation of the future outcome is improved, and doctors are led to better decision-making.

An essential attribute of prognostic models is the fact that they have a wide range of use. More specifically, they define healthcare policies worldwide by setting up universal predictive scenarios. They also enable doctors to come up with a subset of patients, for whom specific innovative therapies are suitable. These models focus on facilitating patient clinical management and support doctors in everyday decisions (e.g., treatment choice, treatment adjustment, medical test selection) [2]

The modeling process uses data from clinical studies to obtain simple prognostic scoring rules. Such rules can be formulas in which significant variables are multiplied by a positive coefficient (the larger, the more significance), whereas insignificant variables are left out (get a coefficient of 0). The result of this procedure yields prognostic knowledge based on statistics [69]. An example is the International Prognostic Index (IPI) and the revised one (R-IPI). While the former predicts two risk groups, the latter comes up with three prognostic groups for patients with non-Hodgkin lymphoma. A total score is calculated for patients older than 60, according to specific variables (e.g., increased level of blood lactate dehydrogenase). For example, an R-IPI score equal to 2 points corresponds to a substantial degree of survival (four-year progression-free survival) [59]. In addition, prognostication can be used in terms of classification of events of patients, providing information about the individual's survival according to a fixed point in time, which is used as a threshold [2].

A sophisticated alternative to scoring models for prognostics is logistic regression and Cox regression [12]. In the present thesis, the Cox regression analysis is used. In these methods, the modeling process aims at capturing the probability of survival and how several variables (deterministically) jointly influence a specific outcome probability.

Bayesian networks can also assist physicians in predicting survival while providing flexibility to their users. The goal is to acquire a good idea of how multiple factors connect in terms of causality, leading this way to a pre-specified outcome. The data sources used are either expert knowledge, medical data, literature, or all of the above in order to create this graphical representation [73].

Bayesian networks are promising and flexible tools, dealing effectively with missing information. Given that they are the focus area of the present thesis, their advantages are analyzed in detail later on.

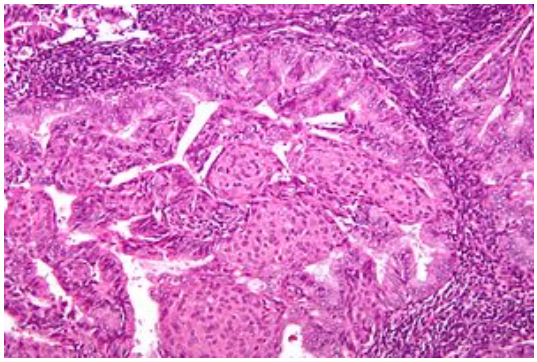


Figure 1: Histology of tissue, as shown under the microscope, obtained by a biopsy revealing endometrial adenocarcinoma

1.2 Research focus

The main point of interest of the current thesis is the ability to provide preoperatively a more precise prognosis for patients with endometrial cancer. The fact is that before surgery, limited information is available, and thus, incorporating all *pre*operative information in a predictive model as well as possible to predict *post*operative outcome becomes a matter of great importance.

The critical point of concern, triggering the current research, is the fact that in the Netherlands, the number of treatment provided to patients is not consistent with the outcome. Physicians reach a diagnosis in endometrial cancer by carefully histologically examining an endometrial biopsy, as shown in Figure 1. The aforementioned is a standard practice followed by medical doctors around the world. Doctors proceed with the surgical staging of the disease, only if they have sufficient information collected. As a consequence, the grade of the tumor is not determined adequately in a considerable number of cases. More specifically, research shows a discrepancy of 40% between the preoperative diagnosis and the final pathology. This is a crucial factor giving rise to either more or less treatment than actually required. Around 10% of patients get lymph node metastasis, yielding a very poor prognosis, the risk of which is hard to estimate. Poor outcome varies considerably among patients.

In response to this problem, the current thesis investigates how a prognostic model can be developed from data using Bayesian networks. These networks use available theoretical knowledge, cohort endometrial cancer data, and experts' guidance. They are probabilistic *graphical* models, with predictive power, that do not have as a prerequisite that all values for all available predictors in the data are needed to use them. This is a big different between the previously mentioned prognostic scoring rules that require all included variables to be known to use them.

An attempt is made to reveal the predictive strength of Bayesian networks and their potential for a more patient-oriented risk computation preoperatively. In detail, the prediction model utilizes data such as patient information, tumor attributes, and biomarkers. More specifically, selected serum-markers and bio-markers are incorporated, as research has shown that they have high predictive value. Thus we aspire to provide new insights to physicians and equip them with a useful tool to improve the prognosis of endometrial cancer.

Additionally, a qualitative comparison between the Bayesian model and the commonly used Cox proportional hazards regression model for survival data is carried out. Cox regression is the standard technique for survival analysis in clinical medicine [11]. The goal is to uncover points of convergence between the two methods and understand in which way they differ. Following this, a quantitative comparison between the established prognostic Bayesian model and a Cox regression model takes place. The goal is to acquire a better understanding of how the Bayesian model behaves as a prognostic model and to evaluate this behavior from a survival analysis point of view.

1.3 The structure of the thesis

Following the present introduction, chapter 2 provides an overview of the fundamental concepts of probability theory, Bayesian networks, and survival analysis. Chapter 3 provides the reader with all the needed information about the problem domain (endometrial cancer), the known risk factors of the disease and explains the current situation regarding prognosis and survival. Subsequently, chapter 4 focuses on the dataset used in the thesis, categorizes the variables into preoperative and postoperative ones and specifies the selected variables. Chapter 6 offers a review of useful software for the present research, describes fully all the steps taken during implementation of the models, the data processing, model fitting and extension. In addition, details are provided about how one can reason about survival, the implementation of a simple CPH model. Finally a comparison is made with BNs, concluding in which ways Cox models differ and why Bayesian networks are a really intuitive and more flexible approach for risk prediction of patients with endometrial cancer. The last chapter, chapter 7, concludes the thesis with a discussion and some directions for future research.

1.4 Contributions

The goal in the present research is to verify whether an efficient model for the prognosis of endometrial cancer can be implemented by using machine learning and specifically Bayesian networks. Medical research has brought to light new information about multimodal biomarkers that can be conveniently extracted and are closely associated with this disease. None of them are actively used in clinical practice. It was therefore necessary to assess if combining such variables with Bayesian networks can enable gynecological oncologists to achieve better prognosis preoperatively. The final Bayesian network is capable of satisfying this need and is flexible enough to be used for more than one event of interest. The medical doctors of Radboud University Medical Center have used the ENDORISK model in their research as decision support tool aiming at revealing both patients survival and also lymph node metastasis. After the developed model had been finalized, the researchers validated it externally. The continuation of their research has been already funded after the publication of the research paper. The next step is planning an implementation study for doctors to actually start using the model in their practice. The medical doctors continue with performing molecular analysis on the cohort to expand the model with new variables and the plan is to validate it later on in a big cohort in Norway (10,000 patients). Additionally the current thesis focuses on comparing this model with the regression-based Cox model and specifies in detail in which ways BNs prevail over CPHs.

2 Preliminaries

2.1 Notation

To provide a clear understanding of the technical terms employed in this thesis, we first start with reviewing the notations used. We use upper-case letters or strings, e.g., X, Y, to denote random variables and bold upper-case letters e.g., $\mathbf{X}$ and $\mathbf{Y}$, to denote a set of random variables. For a binary variable with values *true* and *false*, its values are also denoted by lower-case letters x, and $\bar{x}$, short for $X = \mathit{true}$ and $X = \mathit{false}$, respectively. Alternatively, we use $X = T$ and $X = F$, and $X = 1$ and $X = 0$, respectively. If the value of a variable is known, this is often referred to as an *observation*, an *instantiation*, or *evidence*.

2.2 Probability theory

We continue with a brief review of some key concepts from probability theory, i.e., we consider events, joint probability distributions, conditional probability distributions, the chain rule, marginalization, and conditional independence.

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a set of random variables, where $\mathrm{Val}(X)$ indicates the *domain* of $X \in \mathbf{X}$ and $\mathrm{Val}(\mathbf{X})$ the domain of $\mathbf{X}$, respectively. An (elementary) *event* $E \equiv X = x$ is any random variable X with a value x from its domain. The set of all possible Boolean combinations of events, or *Boolean algebra* denoted as $\mathcal{B}(\mathbf{X})$, is defined by using the operators: conjunction ($X = x \cap X' = x'$) (also called intersection), disjunction ($X = x \cup X' = x'$) (also called union), and negation ($\overline{X = x}$) (also called complementation). This Boolean algebra contains events such as ($X_1 = x_1 \cup X_2 = x_2$), ($X_3 = x_3 \cap X_4 = x_4$), and $\overline{X_2 = x_2}$. Events are partially ordered by $\subseteq$, with the universal lowerbound $\varnothing \in \mathcal{B}(\mathbf{X})$ and universal upperbound $\Omega \in \mathcal{B}(\mathbf{X})$, i.e., we have for each $E \in \mathcal{B}(\mathbf{X})$ that $\varnothing \subseteq E$ and $E \subseteq \Omega$. Usually ($X = x \cap X' = x'$) is represented in set notation as $\{X = x, X' = x'\}$.

A probability distribution is a function or mapping that assigns probabilities, i.e., values from the closed real interval $[0, 1]$, to any event involving variables in $\mathbf{X}$.

Definition 1 (Probability Distribution)**.** A probability distribution for a set of random variables $\mathbf{X}$ with *domain* $\mathrm{Val}(\mathbf{X})$ is defined as a function $P : \mathcal{B}(\mathbf{X}) \to [0, 1]$, such that the following axioms hold:

(1) $P(E)$ is a non-negative real value for all $E \in \mathcal{B}(\mathbf{X})$;

(2) $P(\Omega) = 1$;

(3) for any set of disjoint events $E_1, \ldots, E_n \in \mathcal{B}(\mathbf{X})$, with $(E_i \cap E_j) = \varnothing$, $1 \leq i, j \leq n$, $i \neq j$, we have that:

$$P\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} P(E_k).$$

It is a fundamental property of probability theory that it is sufficient to specify a probability distribution in terms of joint events $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$, i.e., in terms of a *joint probability distribution* $P(X_1, X_2, \ldots, X_n)$ for all values of the domain $\mathrm{Val}(\mathbf{X})$ (possibly with the exception of one element from $\mathrm{Val}(\mathbf{X})$, where its probability can be derived from the other probabilities of elements of $\mathrm{Val}(\mathbf{X})$ according to axioms (2) and (3)).

When the actual value of a random variable in an elementary event does not matter in a given context, we often also write $P(X)$ rather than $P(X = x)$ for the probability of variable X taking the value x.

The *marginal probability distribution* for a set of variables $\mathbf{Y}$ given the probability distribution for the random variables $\mathbf{X}$, with $\mathbf{Y} \subseteq \mathbf{X}$ and $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$, where $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint, is obtained

by summing out the other variables (i.e. $\mathbf{Z}$) from the joint probability distribution $P(\mathbf{X})$, and is defined as:

$$P(\mathbf{Y}) = \sum_{\mathbf{z} \in \text{Val}(\mathbf{X} \setminus \mathbf{Y})} P(\mathbf{Y}, \mathbf{Z} = \mathbf{z})$$

Let $P(\mathbf{X}, \mathbf{Y})$ be a joint probability distribution over a set of random variables $\mathbf{X}$ and $\mathbf{Y}$. A conditional probability distribution $P(\mathbf{X} \mid \mathbf{Y})$ is defined as:

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{X}, \mathbf{Y})}{P(\mathbf{Y})} \tag{1}$$

with $P(\mathbf{Y}) > 0$.

It is good to realize that $P(\mathbf{X} \mid \mathbf{Y})$ is actually a family of probability distributions, one for every value $\mathbf{y}$ of $\mathbf{Y}$. The conditional probability $P(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y})$ is the probability of the event $\mathbf{X} = \mathbf{x}$ given knowledge about the event $\mathbf{Y} = \mathbf{y}$.

The concept of conditional probability is one of the most fundamental and most important concepts in probability theory. In addition, the conditional probability plays an essential role in a wide range of domains, including classification, decision making, prediction and other similar situations, where the results of interest are based on available knowledge.

By moving the denominator on the right of Equation 1 to the left, Equation 1 can also be written as:

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X} \mid \mathbf{Y})P(\mathbf{Y}) = P(\mathbf{Y} \mid \mathbf{X})P(\mathbf{X}) \tag{2}$$

By applying Equation 2 to a set of random variables $\{X_1, X_2, \ldots, X_n\}$, this creates a chain of conditional probabilities, more formally:

Proposition 1 (Chain Rule)**.** *Let P be a joint probability distribution over a set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$. Then it holds that:*

$$P(X_1, X_2, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \cdots P(X_2 \mid X_1)P(X_1)$$

The chain rule allows us to compute the joint distribution of a set of any random variables by only making use of conditional probabilities. This rule is particularly useful in Bayesian networks, which we will introduce later in this chapter. Combined with the network structures, the use of the chain rule can facilitate the representation for a joint distribution.

Another immediate result of Equation 2 by rearranging terms is *Bayes' rule*:

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{X})P(\mathbf{Y} \mid \mathbf{X})}{P(\mathbf{Y})} \tag{3}$$

Bayes' rule tells us how we can calculate a conditional probability given its inverse conditional probability. For example, using Bayes' rule makes it possible for us to derive the conditional probability $P(\mathbf{X} \mid \mathbf{Y})$ from its inverse conditional probability $P(\mathbf{Y} \mid \mathbf{X})$, if we also have information about the prior probability $P(\mathbf{X})$, $P(\mathbf{Y})$ of events $\mathbf{X}$ and $\mathbf{Y}$ respectively. $P(\mathbf{Y})$ also behaves as a normalizing constant.

A more general conditional version of Bayes' rule, where all probabilities are conditional on the same set of variables $\mathbf{Z}$, also holds:

$$P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z})}{P(\mathbf{Y} \mid \mathbf{Z})}$$

with $P(\mathbf{Y} \mid \mathbf{Z}) > 0$.

Another fundamental concept in probability theory is *conditional independence*. Two sets of variables $\mathbf{X}$, $\mathbf{Y}$ are said to be conditionally independent given a set of variable $\mathbf{Z}$, denoted $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$, if

$$P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) \quad \text{or} \quad P(\mathbf{Y}, \mathbf{Z}) = 0 \tag{4}$$

Equation 4 asserts that given knowledge of a set of variables $\mathbf{Z}$, knowledge of whether $\mathbf{Y}$ occurs provides no extra information on the probability of whether $\mathbf{X}$ occurs.

2.3 Bayesian networks

Bayesian networks are a compact and natural graphical representation of probability distributions. A Bayesian network, abbreviated as BN, is a probabilistic graphical model that represents a set of random variables and their conditional independences via a directed acyclic graph (DAG).

As Bayesian networks are a graphical formalism, we will use a lot of notions of graph theory and a small fraction of it is summarized next. Let the pair $G = (\mathbf{V}(G), \mathbf{E}(G))$ be a graph, often abbreviated to $G = (\mathbf{V}, \mathbf{E})$, then the set $\mathbf{V}$ is called its set of *nodes*, and the elements in the set $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ are called *edges*. For Bayesian networks, we restrict ourselves to *directed* edges or *arcs*, i.e., if $(v, v') \in \mathbf{E}$, then we assume that it is different from (v', v) and $(v', v) \notin \mathbf{E}$. An arc $(v, v') \in \mathbf{E}$ is often denoted $v \to v'$. When a graph G only contains arcs, it is called a *directed graph*. Furthermore, the concept of *children* of a node $v \in \mathbf{V}$ is defined as $\gamma(v) = \{v' \mid v \to v' \in \mathbf{E}\}$ and the set of *parents* of a node $v \in \mathbf{V}$ is defined as $\pi(v) = \{v' \mid v' \to v \in \mathbf{E}\}$. Finally, when we follow the arcs of a graph G between two nodes v and u we have a *directed path*; when there are *no* paths in the graph G of the form $v \to w \to \cdots \to u \to v$ (first and last node of the directed path are equal) it is called *acyclic*. In the following we will no longer make a distinction between nodes $v \in \mathbf{V}$ and the associated random variable X_v, and simply indicate the node by X.

A formal definition for Bayesian networks is given in the following.

Definition 2 (Bayesian Network)**.** A Bayesian network $\mathcal{B}$ is defined as a pair $\mathcal{B} = (G, P)$, where G is an acyclic directed graph and P a probability distribution. The graph $G = (\mathbf{V}, \mathbf{E})$, consists of a set of *nodes* $\mathbf{V}$, representing random variables, and a set of directed edges or arcs $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. Let $X \in \mathbf{V}$ be a variable and $\pi(X)$ be the parents of X in graph G. The distribution P is defined as a joint distribution over variables $\mathbf{V}$, specified by multiplying conditional probability distributions for each variable $X \in \mathbf{V}$ in the form of $P(X \mid \pi(X))$, formally:

$$P(\mathbf{V}) = \prod_{X \in \mathbf{V}} P(X \mid \pi(X)) \tag{5}$$

As mentioned earlier, a directed acyclic graph G represents a set of conditional independence assumptions over a set of variables X. The aforementioned relations are defined via a core criterion of graphical causal models, which is known as *d-separation*.

An existing path between two nodes X, Y in a DAG G is composed of a sequence of edges, independently of their direction. An *ancestor node* X of node Y is a node with a sequence of arcs starting from X and going to Y, as follows: $X \to \cdots \to Y$; node Y is called the *descendant* of X. When one node Y has only converging arcs (head-to-head) i.e. $\to Y \leftarrow$, it is called a *collider*. The rest of the non-end-points are known as non-colliders and they are connecting to other nodes as follows:

- tail-to-tail arcs: $\leftarrow Y \rightarrow$
- tail-to-head arcs: $\leftarrow Y \leftarrow$
- head-to-tail arcs: $\rightarrow Y \rightarrow$

Next, the formal definition of d-separation is provided.

Definition 3 (D-separation [50])**.** Every path connecting sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ in a DAG G, ignoring direction of the arcs, is considered to be *d-separated* by a set of of nodes $\mathbf{Z}$, if and only if the following holds:

- every path between nodes in $\mathbf{X}$ and $\mathbf{Y}$ includes a non-collider node in $\mathbf{Z}$, or
- none of the collider nodes on a path and none of their descendants occur in $\mathbf{Z}$

written as $\mathbf{X} \perp\!\!\!\perp_G^d \mathbf{Y} \mid \mathbf{Z}$.

An example of d-separation is given in Fig. 2. In this case, X_1, X_7 are d-separated, given that X_4, X_5 block the only existing path between X_1, and X_7; they are both non-colliders: the $X_1 \perp\!\!\!\perp_d X_7 \mid \{X_4, X_5\}$ holds. On the other hand, $X_1 \perp\!\!\!\perp_G^d X_7 \mid X_2$ and $X_1 \perp\!\!\!\perp_G^d X_7 \mid X_6$ are *false*, given that the highlighted nodes X_2 and X_6 are colliders as shown in the 2. We call in both cases X_1 and X_7 *d-connected*, written as: $X_1 \not\perp\!\!\!\perp_G^d X_7 \mid X_2$ and $X_1 \not\perp\!\!\!\perp_G^d X_7 \mid X_6$.

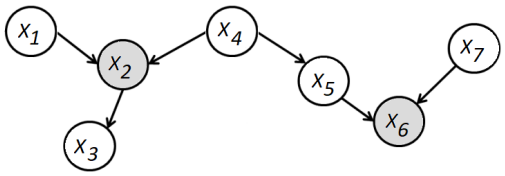


Figure 2: D-separation example [17]

As a Bayesian network is both a representation of a joint probability distribution in the form of a graph and a specification of the associated probabilistic parameters, the two, network and parameters, are closely connected. One of the properties of a Bayesian network is that all the independence expressed by means of d-separation are also implied by the joint probability distribution. Formally it holds for any Bayesian network $\mathcal{B} = (G, P)$ for any disjoint sets of nodes (variables) $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ that

$$\mathbf{X} \perp\!\!\!\perp_G^d \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$$

It is said that G is an *I-map* of P [50].

2.3.1 Bayesian networks in prognosis

Prognosis is the process of reaching a specific prediction about the outcome of a disease, given information about the patient and based on a therapeutic approach [73]. The primary focus of such a model is to be able to predict the patient's situation in the future. It is possible that the focus is not a person but a group of people [2]. In the former case, the physicians interact with these models to get some conclusions about the patient's outcome in the future. As a consequence, the patient reacts to the predictions/decisions and has a better understanding of the situation and the available options. In the latter case, where more patients are the focus, the Bayesian models support the medical doctors in coordinating better the treatment process and perform better resources management [77].

Time is also critical in these types of models. Bayesian networks enable doctors to predict a specific outcome in the future when time is also incorporated in the model. The prognostic Bayesian models have a significant advantage; the prediction process takes place whenever it is required by the physician, combining each time all available knowledge for a specific patient

[75]. What needs to be clarified is that after interacting with a Bayesian model, the doctor can obtain an idea regarding all variables of interest and not just outcome variables (e.g., death).

The goal is to be able to query the Bayesian network and answer questions as accurately as possible about new data, that are not used for learning the probabilities and train the prognostic BN.

In detail, training is considered to be the procedure in which a dataset is used to build the prognostic Bayesian network. The network is created in such a way, that the final Bayesian network yields probabilities that approach the training data. It is uncommon to produce a Bayesian network that is capable of precisely recreating the training dataset. The training process consists of two essential steps. First off the structure of the network is created, and then the learning of the parameters takes place. The knowledge that lies within each Bayesian network is the structure and the probabilities between the selected variables.

2.3.2 Advantages and challenges in Bayesian networks

The current section highlights the significant *advantages* of Bayesian networks. The accuracy of Bayesian network can be good even when learned from a small dataset, but in that case the graphical structure needs to be sparse, as otherwise the network will act as a database giving rise to overfitting. In some applications, the number of missing records can be high. Even in these cases, algorithms such as the Expectation-Maximization (EM) algorithm enable to calculate conditional probabilities [71].

A feature of great importance is the fact that BNs are able to combine multiple sources of information for obtaining a suitable graphical structure and probability distribution. As also happened in the current thesis, researchers often use expert knowledge in designing the model structure together with available data to build the complete model (structure and distribution). It is even possible to combine available data concerning variables with variables for which data is not available at all, and use expert knowledge or information from the literature instead to assess probability distributions [39]. This makes the formalism rather unique in its flexibility.

On the other hand, there are some *challenges* to tackle when working with Bayesian networks. In many cases, the available datasets contain both discrete and continuous variables. To combine the two types of variable, usually the decision is made to discretize the continuous variables and to proceed with working in the discrete domain. As a consequence, the discretized variables may lose some information and possibly also statistical strength. Although there are some known automatic methods to pursue discretization, the option to take into consideration expert knowledge remains an optimal way to deal with the challenge [30, 81, 16].

A challenging problem occurs when a network consists of a large number of nodes and it is in that case likely that the Bayesian network has some nodes having many parent nodes, i.e., the network is large, complex, and dense. If only a low number of data points is available for a particular node as a function of the parent-node values, the resulting model will not fit the data well. In general, the size of a conditional probability table (CPT) of a node in a Bayesian network is exponential in its number of parent nodes, i.e., $O(n \cdot d^{k+1})$, where k is the maximum number of parent nodes, d is the maximum number of values each node can take, and n the number of nodes present in the Bayesian network. Clearly, the number of parent nodes influences severely the number of needed BN parameters. In cases of small datasets and dense Bayesian networks, Bayesian networks lose their key advantage and alternative methods are needed [71].

2.3.3 Incomplete data

The presence of incomplete observations in the dataset leads to multiple problems. Firstly, missing information decreases the size of the sample and as a consequence the statistical tests are much less accurate. Additionally, from a more practical point of view, the majority of

computational mechanisms prefers complete information. This is why the available statistical tools offer different options to create complete datasets by imputing the missing observations.

It is quite often the case that the reason behind this missing information is unknown. The statisticians usually perform comparison of their analysis under different assumptions. In order to be able to make such assumptions for the missing data and decide about how to tackle this challenge, the statisticians categorize the multiple mechanisms based on how random they appear to be.

- Missing completely at random (MCAR): In this situation the probability of a variable missing is the same for all variables in the dataset. If, indeed, this is a valid assumption to make, then the statisticians do not have bias in their inferences.
- Missing at random (MAR): In most cases, the information is not missing completely at random from a dataset. Thus, the statistician proceeds to a slightly more complex assumption. When data fall under this category then a variable is missing with a probability that is based only on existing knowledge and not dependent on the missing data. If a predictor variable is MAR, then it can be treated as NA, but it is necessary to have a study with enough data on this variable to avoid any bias.
- Missing not at random (MNAR): In this case, the probability of a variable missing is always related to unobserved information and the statistician does not have any background information in respect to the missingness of these data.

In any case, the researchers' goal is to understand and remove the bias if it is introduced in their analysis. Each time, the statisticians choose among various missing data methods to impute or delete incomplete data. The *imputation* methods are used to deal with the missing data and create a complete version of the datasets. These methods replace the NA's with estimated values given the knownn information. There is a great variety of available methods such as single imputation, multiple imputation and model-based estimation. Maintaining the initial dataset size can improve effectively the accuracy of the analysis and reduce bias. On the other hand, bias can be also introduced if imputation is performed in a wrong way. In practice, the deletion of the missing information is possible too, but it should be performed really carefully. Discarding data can be quite inefficient and can increase standard errors given that the number of observations is reduced.

In more detail, *single imputation* methods are considered to be the ones that produce a complete dataset by inserting values where the dataset has NAs, without having determined a specific model for the incomplete data. A procedure that is used quite often is mean imputation, in which the incomplete observations are filled in with the mean value of the existing values for each variable. Please refer to [34] for other options of single imputation. One of the issues occurring when single procedures are used is that the replaced observations are perceived by the analysis as real observations but with standard errors. *Multiple imputation* addresses this issue by catering for the uncertainty of the estimates. In general, these methods attempt to fill in each missing observation with multiple imputed values instead of one. Each value comes from a slightly altered model. The goal is to acquire different complete datasets, for which the researcher performs a standard analysis and later on combines inferences for the total number of the imputed datasets. However, as already mentioned, there are deletion procedures that address the challenge of missing data. The most frequent are the listwise or complete case deletion and pairwise or available case deletion. The former one refers to deleting the observations that have missing data for any variable in the dataset and considering only complete observations. Pairwise deletion describes the method in which an observation is deleted when there is a missing value for a variable needed for a specific analysis, but incorporating this observation in analysis for which all required variables have values. [34]

Structure	*Observability*	*Learning Method*
Known	Fully known	Maximum-Likelihood estimation
Known	Partially known	EM, MCMC
Unknown	Fully known	Search model space
Unknown	Partially known	EM & Search model space

Table 1: Different cases in learning BN

2.3.4 Learning Bayesian networks

In lots of different applications, the underlying Bayesian network needs to be determined by using the given dataset. The process requires the construction of the graph representation, given that someone has the necessary prior knowledge and data at their disposal (e.g., knowledge by experts, cause-effect relations between variables). Following this, the estimation of the parameters of the joint probability distribution in the Bayesian network takes place. In practice, the latter is called fitting a Bayesian network to the given data. The construction of the graph, in the absence of expert knowledge, is realized by using appropriate structure learning algorithms.

The following equation describes the joint probability of a Bayesian network with structure G with its associated parameters Θ given a database D.

In detail, the term $P(G, \Theta \mid D)$ is the product of $P(G \mid D)$ and $P(\Theta \mid G, D)$. $P(G \mid D)$ describes how likely a particular Bayesian network is, given data D. This term show that there are many different graphs G that are needed to calculate it. In this case, a search algorithm could be used to reveal the graph that maximizes the term, when D is fixed. $P(\Theta \mid G, D)$ expresses how likely specific probability tables are for a Bayesian network, given a graph structure G and data D. This equation reflects the fact that the learning of the structure and the parameters of a Bayesian network are two different processes.

$$P(G, \Theta \mid D) = P(G \mid D)P(\Theta \mid G, D) \tag{6}$$

Structure learning focuses on identifying the DAG G, which reveals the existing dependencies among the variables. The process becomes more challenging when there are either additionally hidden nodes (partially known) or missing data. The table below indicates different cases of learning a Bayesian network [46].

The difference between structure and parameter learning is the assumption that parameters in separate local distributions should be independent, and as a result, they are learned effectively and simultaneously for every variable [57]. In contrast, research has shown that structure learning is a challenging procedure, and multiple algorithms try to deal with it.

When it comes to parameter learning, it is also important to distinguish if the purpose is to come up with a point parameter estimation (best selected unique model), or a Bayesian parameter estimation (posterior distribution over parameters). Parameter learning as a procedure fits a model to the data by producing an estimation of the parameters of the global probability distribution. Given a structure that is known to the user, either through proper structure learning algorithm or prior knowledge, an estimation of the parameters of the local distribution is computed effectively. In detail, each one of the nodes in the network has a corresponding CPT (Conditional Probability Table). This reflects the node's conditional probability distribution, according to the values of the parent nodes.

In respect to learning the structure of a Bayesian network, two major categories exist. These are the constraint-based algorithms and the score-based algorithms. There are also hybrid algorithms. The best scenario is to construct a graph that is the minimal I-map of the dependence structure of the dataset. [47]. Even if this is not the case, the resulted distribution should approach the correct distribution in the probability space.

Constraint-based algorithms approach the problem through various statistical conditional independence tests in order to produce the dependencies between the variables. The number of dependencies and independencies is then illustrated in a DAG. More specifically, these algorithms calculate the conditional independencies among the variables. The conditional independence constraints are distributed then across the DAG. After this step, the incompatible ones are entirely excluded. These algorithms yield only the I-equivalent graphs, which are the ones with structures that illustrate identical independence relations [63]. The basis of constraint-based algorithms is Pearl's Inductive Causation algorithm [70] Examples of these algorithms are PC [18], gs, iamb [68], fast.iamb [78]. In Bayesian networks based on smaller datasets, structure learning algorithms give better results than score-based ones. In more detail, the steps that are executed by these algorithms are as follows [55]: The underlying structure of the network is learned. Specifically, the result is the undirected graph, which is the skeleton of the network. The complexity of the exhaustive search, even for small datasets, means that the algorithms limit the search, that they perform, to each node's Markov blanket. The Markov blanket consists of the parents, the children, and the total number of nodes that have a common child with this specific node. The connections, which participate in a u-structure, become directed ($X_i \rightarrow X_k \leftarrow X_j$). Finally, the remaining arcs become also directed, by making sure that the network is a directed acyclic graph.

Score-based algorithms have as their purpose to provide an optimized graph and should be preferred when large datasets are available. The basis of this type of algorithms is the fact that in each of the models, one network score is assigned, revealing this way how well the model fits the data. The goal is then to maximize a scoring function [53][19]. The first part of the score-based algorithms is a scoring criterion capturing how well a graph G fits the data D. This score enables us to order different Bayesian networks based on their quality. The two score-based algorithms considered in the present thesis are Hc (Hill-climbing search), and tabu search.

Hill-climbing search is a heuristic search algorithm that falls under the local search algorithms category. Given the pseudocode of the hill-climbing algorithm, a dataset and a heuristic function are used and an attempt is made to come up with the most suitable solution to a problem in a reasonable period of time. More specifically, the algorithm chooses the best successor node under some heuristic function denoted by *f*. This function sorts all possible options at every step of the search. Every time a successor node is found, the algorithm commits the search to it and continues with the most suitable step in the search space. The process continues until no more improvement is feasible and a local maximum (or minimum) is achieved. It is the case that the mentioned solution may not be the global optimal solution. In the current thesis, we use arc-strength in the Bayesian network, to help the algorithm avoid being stuck in local maxima, which of course is not the best and most optimal solution [14].

On the other hand, *tabu search* algorithm is similar to hill-climbing, but better behavior is expected. More specifically, the algorithm's notable feature is that it uses memory structures, which are called tabu-lists. A short-term group of past results is saved in the tabu-list, in order to better filter the resulted models and come up with the most optimal one, avoiding this way being stuck in the local optimum. Through the use of the tabu-list the search does not return the recently visited nodes of the search area, unless the number of moves in this alternation is greater than the length of the tabu-list. One simple way to update the tabu-list is to include any step that has been revisited in the last k steps of the search [14].

Hybrid algorithms characteristic is that they have two phases. The constraint-based approach is used for the first one in order to minimize the space of the number of graphs G (restrict phase). The second phase, which is a score-based approach, focuses on unveiling the optimal solution in the limited space. A really good example in this category is the Max-Min hill-climbing (MMHC) [57] [68]

Algorithm 1 Hill-Climbing search

```
1: Input: State space min. problem with initial state s and neighbor relation Succ
2: Output: State with low evaluation
3: u ⇐ v ⇐ s
4: do
5:    Succ(u) ⇐ Expand(u)
6:    for v ∈ Succ(u) do
7:       if (f(v) < f(u)) then
8:          u ⇐ v
9: while (u ≠ v)
10: return u
```

Figure 3: Pseudocode for hill-climbing algorithm

Algorithm 2 Tabu Search

```
1: Input: State space min. problem
2: Output: State with low evaluation
3: Tabu ⇐ {s}
4: best ⇐ s
5: Terminate ⇐ false
6: u ⇐ s
7: while (¬Terminate) do v ⇐ Select(Succ(u)\Tabu)
8:    if (f(v) < f(u)) then
9:       best ⇐ u
10:      Tabu ⇐ Refine(Tabu)
11:      Terminate ⇐ Update(Terminate)
12:      v ⇐ u
13: return best
```

Figure 4: Pseudocode for tabu search algorithm

2.3.5 Querying the Bayesian Networks

Inference is the process of getting answers from a Bayesian network after its implementation. The graph provides information about a specific subset of the variables when the values of different variables are known. These variables are also called evidence variables. By defining the values of these evidence variables, the Bayesian network is instantiated, and the process of acquiring answers, which is known as probabilistic reasoning, starts. The unique feature of BN is the fact that there are no input or output variables. Input variables are considered the ones for which the values are known. This way any variable can be either an input or an output variable. After their determination, the information propagates in both sides in each BN, resulting in the desired probabilities. To sum up, the probabilistic inference can be defined as the procedure that produces the posterior distribution of variables based on specific evidence.

There are currently multiple different algorithms performing probabilistic inference in Bayesian networks. These algorithms differ in their characteristics and what they offer. Others are less complex but not that fast, while some are not that accurate and vague while being fast. Some of them attempt to cluster dependent nodes and treat them as super-nodes. Others incorporate message propagation to exchange probabilistic information, whereas some algorithms

eliminate nodes that are not necessary and absorb the probabilistic information in the rest of the nodes. Based on the need, the user proceeds with choosing the appropriate probabilistic inference algorithm taking into consideration the different trade-offs. These algorithms are divided into exact and approximate algorithms. Inference becomes a challenging task when performed in extensive networks. Additionally, Bayesian networks with very small probabilities increase the difficulty, and they cannot be easily queried. In case there is a causal relationship among the nodes in a Bayesian network, then inference is interpreted differently. The arcs between the nodes reflect causality, and the queries assess the probability of identified causes, given the outcome or vice versa [47].

2.3.6 Goodness-of-fit

In the process of discovering the best possible fit among a list of similar models, a proper comparison has to be performed. The goal is to find what is the model that explains the dataset in the best possible way. As in several types of mathematical models, Bayesian models that have minor differences (e.g., an extra node, an additional arc, an opposite direction for specific arc) need to be effectively compared.

The essential quantity in this process is the likelihood of the dataset D given the DAG structure G. We also will use $P(G)$: the probability distribution over the DAG structures. The *likelihood* is simply the probability of the data D given the graph G, i.e. $P(D \mid G)$.

In this thesis we often wish to compare the fit of two different DAGs, G and G', on the same dataset D. This is done by means of the *likelihood ratio*:

$$q = \frac{P(D \mid G)}{P(D \mid G')} \tag{7}$$

also called the *Bayes factor*.

The log-likelihood is often easier to compute:

$$\log q = \log P(D \mid G) - \log P(D \mid G') \tag{8}$$

After the computation of a node's log-likelihood, all the values for each node are added together. The Bayesian score for a BN $\mathcal{B}$ is thus defined by summing all the logs together for the participant nodes. This way, the result is more numerically tractable. However, the mentioned step is performed under the assumption that the observations are independent and that they come from identical distributions.

Thus, the model selection takes place by checking the log-likelihood (equation 8) of the available dataset D under a model G and under a model G'. By using this approach, the explanation is straight-forward; In case G yields a better score, then, it is considered to be a more representative picture of the dataset, in comparison with G'. The log-likelihood values can be lower than zero. Thus, the best possible value for this score is the least negative. In terms of absolute numbers, the lower it gets, the better the result is. As already discussed, careful consideration of the number of variables that are present to the model is needed. In general, the increase of the variables in a Bayesian network brings a much more convenient score in the end. As a result, the focus should be on maintaining the right balance between the number of nodes in the model and the available data. In our case, the log-likelihood ratio is used as a means of comparison between the different models.

In the case that we have to cater to more complicated Bayesian models (e.g., larger BN), two options are possible. One way to tackle this challenge is to include background information about the Bayesian network. Alternatively, the user can accept the assumption of all BNs being evenly possible.

2.4 Survival analysis

2.4.1 Introduction

Survival analysis describes the methods used for the data analysis which focuses on the time until a specific event. According to the theory of survival analysis, time acts as a parameter which allows modeling the time until a particular event takes place. It is important to clarify that there is no assumption that the rates of occurrences remain the same. The model, in this case, reflects one of the following three cases:

- the time until a specific event takes place,
- the comparison of the time until a specific event among multiple groups,
- the correlation between the time until an event occurs with quantitative data.

The term event might refer to recurrence after treatment, recovery, any other specified point of interest in the life of one of the individuals of the study or death. In general the method of survival analysis supports model-ling the time of the occurrence of any event. However only one event is the point of interest of a specific statistical analysis and as a result the probability distribution describes only one variable which is the event variable (usually survival).

2.4.2 Statistics

A *survival function* expresses the probability of a participant in the study to survive and specifically the probability that the event of interest (death) will not happen until a specific point in time t. T corresponds to the time when the event of interest takes place and $P(T > t)$ corresponds to the probability that the survival time comes later than the moment t.

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx \tag{9}$$

Where,

$$f(t) = -\frac{dS(t)}{dt} \tag{10}$$

$f(t)dt$ can be roughly interpreted as the probability that the event of interest will take place at time t, where $f(t) > 0$ and the area underneath is equal to 1.

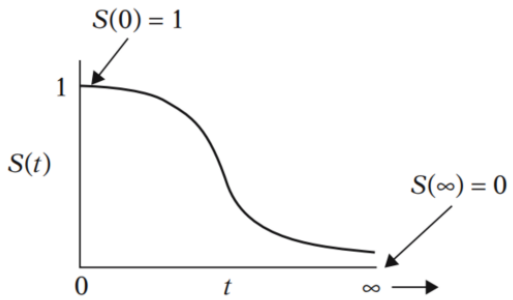


Figure 5: survival function [17]

There are numerous survival curves mentioned in the literature. However, they have the same essential characteristics: they are monotone functions which are equal to 1 at time $t = 0$ and approach 0 as time increases to infinity, as figure 5 shows [17].

The *hazard function* $h(t)$ corresponds to the probability at a specific moment, per portion of time, for the event of interest to occur, given that the individual has not experienced it up to time t [17].

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{11}$$

In some sense, the two functions describe the same thing from different perspectives. While the survival function illustrates the positive aspect, by computing the *not* failing scenario, the Hazard function gives the failing scenario [28]. If someone knows one of them, then it is easy to come up with the corresponding form of the other function. More specifically these functions can be written as: [17]

$$S(t) = \exp\left[-\int_0^t h(u)du\right] \tag{12}$$

$$h(t) = -\left[\frac{\frac{dS(t)}{dt}}{S(t)}\right] = \frac{f(t)}{S(t)} \tag{13}$$

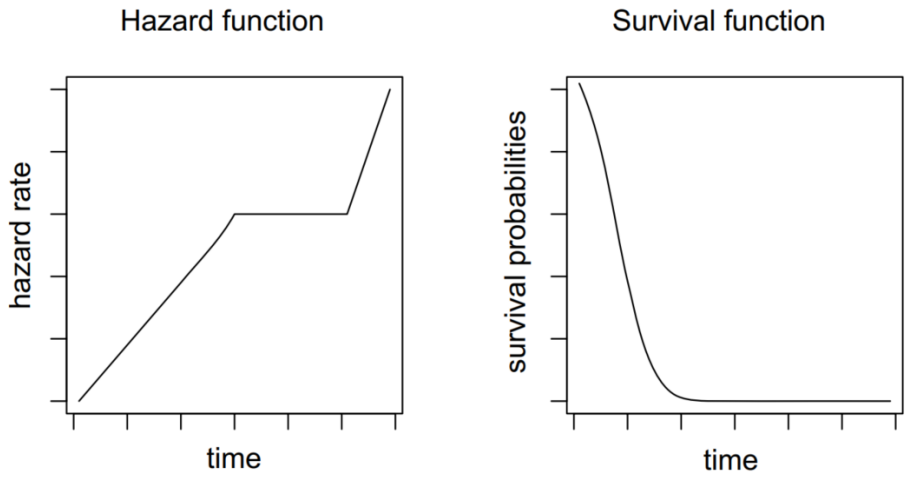


Figure 6: Comparison of Survival and hazard function to the same data [17]

The idea is that when we perform data analysis, through any programming language, the one emerges from the other, through a smooth transformation.

Figure 7: Relationship of S(t) and h(t)

The goals of the statistical analysis of survival data are listed below.

- Calculation and explanation of Survival and hazard functions

- Comparison of Survival and hazard functions: Comparison of the survival functions of different groups of individuals in a study, by taking into consideration a specific variable for which the groups have opposing values. The former group could be patients that use placebo pills and the latter patients that use treatment.
- The assessment of how the variables in a survival dataset influence the survival time. In this case, proper mathematical models are necessary.

To sum up, both these functions provide insights into survival and contribute to the analysis of survival data. Nevertheless, the hazard function gives a more sensitive and intuitive representation of risk.

2.4.3 Censoring

An interesting topic related to survival analysis is censoring. The presence of partially known (censored information) in a study complicates the analysis. Many researchers consider this to be a challenging subject to tackle. In medical studies, patients might stop visiting the doctor and as a result there is no follow up information available. The study may finish, without knowing if the event of interest happens or not. Moreover, another event might take place in an individual's life (e.g., death, when death is not the event of interest). Scientists should handle these data differently. It is possible to ignore these missing observations and focus on the analysis of the uncensored complete data. However, this approach leads to less efficiency, especially when e.g. in the medical world half of the observations are incomplete. Additionally, ignoring the missing information introduces estimation bias to the analysis.

Researchers take into consideration the following types of censored data:

- *Right Censoring* happens when an individual drops out of the study before an event takes place, or the study's end occurs before the event of interest. E.g. In case the event of interest is death, then the data are censored if the patients are still alive when the study finishes. This is a very common type of censoring.
- *Left Censoring* occurs when the event of interest happens before the actual starting date of the study. In a study that is about following people until they get a specific disease, it is possible to record the existence of the disease when the first examination takes place at time t; It is only known then that the event occurred before t and not exactly when. This is a quite rare situation.
- *Interval Censoring* happens when an individual can have good results in the first examination but a negative second examination. In a situation like this, the patient gets the disease between the two moments in time (time interval), but we are not able to determine exactly when.
- *Type I censoring* arises when a study consists of a specific number of individuals and stops at a planned moment in time, when all remaining subjects are right-censored. This type of censoring also happens when there is a completely random reason for which the subjects drop out of the study.
- *Type II censoring* occurs when the study, consisting of a selected number of subjects, finishes after a pre-specified number of individuals participating in the experiment fails. In this case, the remaining individuals are right-censored.
- *Random (or non-informative) censoring* appears when every individual that participates in the study has a censoring time, which is stochastically independent of the time to experience the event of interest. The individuals who have failure time larger than their censoring time belong to the right-censored subjects.

2.4.4 The different approaches in survival analysis

In this section, we attempt to describe the three approaches to model fitting in survival analysis. The purpose is to explain the basics before diving into the methods used in the current thesis. The first one is the parametric approach which assumes the statistical distribution of the survival curve. By carefully considering the Hazard and survival function, the suitable distribution is selected. Parametric models are considered easy to understand and explain. An additional benefit of these models is that they are fast to learn from a given dataset. Even if a parametric model does not fit that well the data, it can still be effectively used and provide results. However, these models also have some limitations. The selected functional form for the baseline hazard poses a constraint for the survival analysis. The use of a parametric model is more suitable for less complicated cases. In real cases, it is highly possible, that there is no good fit. In general, not all distributions are equally suitable for each scenario. *Normal* Distribution is preferred when the risk of failure rises significantly with time. *Uniform* Distribution is not often assumed in real-life applications. In this case the hazard increases exponentially with time. Additionally, *Exponential* Distribution is assumed really often in survival analysis. The risk of failure is constant over time in this case. Moreover, both *Weibull* and *Log-normal* Distribution can be optimized. The former one has a parameter gamma that can be improved in a way that produces multiple hazard behavior over time. This flexibility makes this distribution applicable to lots of different real-life scenarios. The same holds for *Log-normal* Distribution in which the alteration of sigma leads to non-monotonic shape for hazard function. In this case *Weibull* Distribution is not advisable. However their common flexibility and the fact that they are somehow complementary to each other, makes these distributions applicable to almost all situations.

On the other hand, non-parametric algorithms do not assume much about the form of baseline hazard and focus on calculating the regression coefficients. This aspect provides them with the flexibility to learn any functional form from the given data. These models are preferable in case of large datasets and no previous knowledge.

Non-parametric models, as said, are flexible and robust given the lack of assumptions for the functional forms. Their use can lead to models with better performance, especially for predicting purposes. On the other hand, these algorithms pose limitations too. The usage of more massive datasets is necessary for the estimation of baseline hazard. They also require more time to be trained, and there is a high risk of over-fitting the data, without being able to explain why some predictions are made.

The last approach is the semi-parametric models, where the baseline hazard is partially assumed. In these models, there are both parametric and non-parametric components. More specifically, the component which is associated with the covariates is parametric. On the other hand, the part concerning the estimation of survival function is completely non-parametric and as a result no assumption is being made for its distribution. The most well-known semi-parametric method is the Cox proportional hazards model which is used in the current thesis and analyzed later on. It is considered to be the cornerstone of survival analysis in recent years.

2.4.5 Product limit estimation

The Kaplan-Meier survival curve corresponds to the survival probability in a specified period, when time is divided into multiple time intervals. The method belongs to non-parametric estimators and in comparison with the parametric methods, it has higher flexibility and can cater for medical applications [3].

The mentioned analysis takes place based on three fundamental assumptions. First of all, the censored patients are considered to have the same chances for survival as the ones that continue to be under surveillance. Additionally, the survival probabilities are assumed to be

equal both for patients that joined the study quite early but also later. Last but not least, this type of analysis assumes that the event takes place at the indicated time.

The Kaplan-Meier survival analysis method is also known as product-limit estimator. Given N survival times in a dataset D, where d are uncensored and c are censored, we assume that at time t_i, the number of deaths is equal to d_i. In this case, the following holds: $d_1+d_2+...+d_k = d$, in which $t_1 < t_2 < ... < t_k$.

An alternative description is obtained by assuming that at a time interval t_{i-1}, t_i, n_i patients are at risk and censoring happens at time $t_1, t_2, ..., t_k$. Given that the hazard is not changing in the mentioned period, the maximum likelihood of this hazard h_i is expressed as follows:

$$\widehat{h_i} = \frac{NumberOfDeathsIn(t_{i-1}, t_i)}{TotalTimeSurvived} = \frac{d_i}{n_i(t_i - t_{i-1})} \tag{14}$$

$$S(t) = P(T > t) = \exp\left\{-\int_0^t h(u)du\right\} = \exp\sum_i\left\{-\int_{t_{i-1}}^{t_i} h_i dt\right\} = \exp\left\{-\sum_i (t_i - t_{i-1})h_i\right\} \tag{15}$$

As a result,

$$\widehat{S}(t) = \exp\left\{-\sum_i (t_i - t_{i-1})\widehat{h_i}\right\} = \exp\left\{-\sum_i \frac{d_i}{n_i}\right\} \tag{16}$$

where the following holds: $t_i \leq t$.

For large values of n_i,

$$\exp\left(-\frac{d_i}{n_i}\right) = 1 - \frac{d_i}{n_i} + \frac{1}{2!}\left(\frac{d_i}{n_i}\right)^2 - \frac{1}{3!}\left(\frac{d_i}{n_i}\right)^3 + ... \approx 1 - \frac{d_i}{n_i} \tag{17}$$

The product of all i for $t_i \leq t$ is calculated and known as Kaplan-Meier estimate. As a result,

$$\widehat{S}(t) = \prod_{i:t_i \leq t}^{N}\left(1 - \frac{d_i}{n_i}\right) \tag{18}$$

given that in the previous equation, we considered the product over all i, for $t_i \leq t$.

The calculation of survival probabilities by using the Kaplan-Meier estimate is essential to investigate survival data and perform a successful analysis. In this calculation, no assumption is made about the population that generated the sample of survival times. They are completely determined by the characteristics of the sampled data [28].

2.4.6 Cox proportional hazards model

As already said, instead of determining the statistical structure of a given population entirely and having a parametric hazards model in place, an alternative approach can be used. In this case, the estimation of how the predictor variables affect survival does not have as a prerequisite the definition of a baseline hazard function [28]. Additionally, no other assumption for the probability distribution function of the sampled data is necessary, making this approach an essential method for survival data analysis.

The Cox proportional hazards (CPH) model describes the conditional hazard function which is interpreted as the risk of a specific individual to experience a specific event at a specific point in time based on present conditions:

$$h(t|X) = h_0(t) \times c_i = h_0(t)\exp(\beta^T X) = h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_p X_p) \tag{19}$$

In this equation, $\beta = (\beta_1, \ldots, \beta p)^T$ represents the unknown regression coefficients of the model in a p-dimensional vector. These coefficients measure how the covariates X_1,...,X_p affect survival. Their estimation happens with no assumptions about the baseline hazard function h_0. In other words, the way that survival times are distributed is not at all associated with the estimation process. The baseline hazard function h_0 is an unknown non-negative function and it is interpreted as the hazard of an individual to experience a specific event for which all covariates are zero.

As mentioned earlier, no other assumption is made regarding the population. The Cox method is widely applicable to a variety of different situations, and it is generally preferred. However, the procedure requires the hazard functions to be proportional over time. It is necessary to state the core ideas of the Cox method. To be more specific, the hazards are assumed by the model to be proportional. This means that there is an assumption regarding a consistent relationship between the dependent variable and the predictor variables.

In this method, the risk of a specific event in any group of observations is a fixed multiple of the risk in any other. The consequence of this is that the hazard curves for these groups should be proportional and can not intersect. However, this assumption needs to be checked. The proportionality can be evaluated through statistical analysis and graphical diagnostic methods. Furthermore, the method entails the assessment of how well the model fits the sampled data (goodness-of-fit). In the presence of censored data, the method also expects non-informative censoring.

Hazard ratio: In the equation below, β_l is explained with regard to relative risk when X_l rises by 1 and all covariates are constant:

$$\frac{h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_l X_{l+1} + \cdots + \beta_p X_p)}{h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_l X_l + \cdots + \beta_p X_p)} = \exp(\beta_l) \tag{20}$$

In case $\beta_l < 0$, then $\exp(\beta_l) < 1$ and the risk of experiencing an event decreases for a subject as X_l rises. On the other hand, if $\beta_l > 0$, the risk of an event rises as X_l increases. Consider $\mathbf{X}$ and $\mathbf{X}'$ as two covariate vectors, the hazard ratio is described by the following equation and as shown it is not dependent on time:

$$\frac{h(t \mid \mathbf{X})}{h(t \mid \mathbf{X}')} = \frac{h_0(t)\exp(\beta^T \mathbf{X})}{h_0(t)\exp(\beta^T \mathbf{X}')} = \exp(\beta^T(\mathbf{X} - \mathbf{X}')) \tag{21}$$

As it can be concluded $h(t \mid \mathbf{X})$ and $h(t \mid \mathbf{X}')$ are proportional, which of course explains the name of these models. It is necessary to elaborate more on the possible situations on top of the details mentioned above. When the hazard ratio is less than 1, it shows that as the value of a single covariate increases the length of survival increases, indicating a decrease in the hazard. In literature, the predictor variables that have hazard ratio higher than 1, they are known as bad prognostic factors. On the other hand, the covariates with a hazard ratio of less than 1, they are known as good prognostic factors.

Earlier, the Cox proportional hazards model is mentioned as a non-parametric approach. However, we need to be more specific and correct about its categorization. The method includes a parametric part which is the regression parameter $\beta \in \mathbb{R}^p$ and a non-parametric part which corresponds to the baseline hazard function $h_0(\cdot)$. As a result, it is more accurate to consider the method as a semi-parametric approach.

Model validation: While using the Cox proportional hazards model, the *goodness-of-fit* is closely associated with the concept of proportionality. It is necessary to prove that the impact of explanatory variables is not at all related to survival time. In case the hazard functions are not proportional, the estimated regression coefficients (X) cannot describe sufficiently how the

explanatory variables affect the event of interest. In other words, the impact of a single variable is not reflected by one regression coefficient [48] [28].

The evaluation of the goodness-of-fit for a specific model might fail due to multiple reasons. Let us mention some of them. First of all, it is possible to perform an incorrect determination of the functional form of the covariates, e.g., instead of incorporating $\log(X_j)$, X_j is used in 19. Additionally, the PH assumption might not be valid for the specific case because of regression coefficients β that vary with time. Moreover, the link function exp might be specified inaccurately. In other words, there may not exist a log-linear association between the hazard and the linear predictor $\beta^T X$ [28].

The assessment of a model like this 19 takes place by using various techniques. There are different goodness-of-fit tests, graphical techniques and residual methods that have been developed for this purpose. Some example methods are the following. Regarding the proportionality assumption, it can be checked through a score-type hypothesis test and Schoenfeld residuals. The deviation from this assumption takes place when the regression coefficients are dependent on time and as a result equation 19 changes to:

$$h(t \mid X) = h_0(t) \exp(\beta(t)^T X) \tag{22}$$

The assessment of PH, in this case, corresponds to examining:

$$H_0 : \beta(t) \equiv \beta \tag{23}$$

where $\beta(t) := (\beta_1(t), \ldots, \beta p(t))^T$ A likelihood ratio test or the Wald test can also be used for this purpose. Martingale residuals can also assist in examining the functional form of a covariate and the proportionality assumption. It is also essential to evaluate the influence of individual subjects in the study on the estimation of β.

In case the Cox proportional hazards model does not fit the data, it is possible to refine the model through several techniques. Some of them are combining the Cox model with stratification. In this case, time-dependency is integrated to the model by making sure the regression-coefficients influenced by time, are present to the model. Other technique to tackle this issue is additive hazards model and of course much more.

2.5 Bayesian networks vs survival analysis

Survival analysis and especially Cox proportional hazards models are widely used in the medical world. The purpose of the method is to estimate how different risk factors influence survival. As part of this thesis, we implement both a CPH model and a Bayesian network. In the current section we compare the two methods qualitatively, before proceeding with the implementation.

An essential advantage of Bayesian networks is their ability to capture causality which is explicitly reflected. This can not be modeled in Cox regression models. As already explained hazard ratio indicates the impact of a single risk factor to survival. Based on the literature, this ratio is determined as the fraction of the hazard in a particular risk group to the hazard in an another group. The latter one is considered as a group of individuals to whom we do not observe any of the risk factors. According to a common assumption of these models, HR remains consistent over time [11]. However, it makes sense that this is not always the case. An extension of CPH models is thus introduced to deal with this assumption. In general, HRs can quite possibly be altered depending on the existing risk factors. On the other hand, Bayesian networks do not depend on mandatory assumptions that set constraints to their use, apart from conditional independence assumptions and they are much more flexible. Bayesian networks are effectively learnt from data, while it is is also possible to be modeled solely based on expert's knowledge. Alternatively, they can be formulated by combining both approaches. On the contrary, this is not the case for regression models.

Moreover, BNs can be used for multi-risk assessment by integrating different risk models. These models can be initialized by inserting the values for different risk factors and observe how the probability for a specific outcome e.g. death is affected. These models can reflect multiple outcomes in a single network. This allows optimal treatment of patients by applying individual level prediction. There is also the possibility to broaden such a model in order to use it as a decision model, by introducing appropriate utility variables. On the other hand, it is not clear how the extension of the model takes place in CPHs, without performing again parameter learning [29]. In general, one model is required per outcome under a research study. In order to understand how this works in BNs, please refer to figure 14 and its explanation. Bayesian networks are not that widely known in health sector, however regression models and specifically CPHs are preferred. A possible explanation is that a good understanding of Bayesian statistics is necessary to perceive the concepts and apply them efficiently. On top of that, they are computationally demanding models, while CPHs can be used with almost every available statistical package.

3 The problem domain: endometrial cancer

3.1 Clinical description

Endometrial cancer is the result of an affection that occurs in the uterus. More specifically, this type of cancer grows into the endometrium, which is the inner epithelial layer of the uterus[76].It is considered to be more common in developed countries, where more risk factors are present [4] [21]. In general, there has been an increase in incidence rates around the world. More specifically, in countries with accelerated development generated by different socioeconomic factors, e.g., North America and Europe, the increase is multiplied by ten in comparison with less developed countries [24, 35, 36]. This increase relates to a proportionate increase in life expectancy [64]. A study, published on October 16, 2017, in the Journal of the National Cancer Institute, reveals that there is a rapid rise in rates in the entire age spectrum in 26 out of the 43 populations. The same study indicates that this increase takes place in 27 postmenopausal populations and 15 pre-menopausal ones. This information agrees with the dataset used in the current thesis, as women with postmenopausal status have a higher risk of developing endometrial cancer in comparison to pre-menopausal women [35]. During 2017, around 61,380 women were diagnosed with this type of cancer in the USA. Endometrial cancer is the fourth most common cancer and the sixth most common cause of death in the United States. In Europe, approximately 9000 patients die from this disease [7] [4] and one in every twenty women with cancer in Europe develops the disease into the uterus. Despite the progress that has been made during the last years not only relatively to the early detection of the disease but also treatment methodologies, there are not promising results regarding mortality [9, 61]. Endometrial cancer (EC) has become the most frequent type of cancer in the female reproductive system in developed countries. Some risk factors as high-fat consumption and in general poor diet increase the number of incidents [25] [8]. More than half of the cancers developed have a good prognosis. However, there is a proportion of 20% that lead to a bad clinical result [41, 42] The classification based on histopathological factors creates two categories [42, 26]:

- Endometrioid endometrial cancer (EEC);
- Non Endometrioid endometrial cancer (NEEC).

EEC of grade one or two, which is considered to be low-risk cancer [41]-[60], is treated by a hysterectomy with two-sided salpingo-oophorectomy. On the other hand, the treatment for grade three of the former type and also for high-risk NEEC is advised to be a complete surgical staging. The step mentioned above is a matter of crucial importance since, in these carcinomas, the disease tends to spread. More specifically, lymph node dissection has to take place, given that they tend to spread to lymph nodes. This staging also consists of omentum biopsies and random biopsies throughout abdomen [32, 79, 44, 65, 5]. The current method of establishing a preoperative diagnosis of endometrial cancer takes place in a specific way by doctors all over the world, using an endometrial sample. The most important thing is the fact that this way, preoperative diagnosis differs from the final pathology by 40%. The substantial result of this is that patients do not receive the proper treatment. More specifically, 25% of the patients in the Netherlands receive more treatment than they should. On the contrary, 15% of patients receive less treatment. The most crucial point is that nowadays, patients do or do not receive surgical staging because of the under-grading or over-grading [20]. Besides that, there is another resulting issue related to the prediction of lymph node metastasis that happens based on preoperative grading only. In detail, patients that have preoperative low-risk histology (grade 1 or 2) have a 5-8% risk of lymph node metastasis [67], on the other hand, patients with high-risk histology (grade 3 EEC and NEEC) have a 24% risk of lymph node metastasis [67]. Staging is then based only on histology, an essential part of low-risk patients is not taken into consideration for lymph

node metastasis. It is essential to highlight that 5-8% remains a significant proportion, given that low-risk patients are approximately 80% of the total population. At the same time, the overtreatment of high-risk patients, without the presence of lymph node metastasis, is a fact. The need for molecular markers is highly crucial because, by using them, the assessment of each patient's risk happens preoperatively and becomes much more effective. The research that has been done so far regarding markers such as estrogen receptor (ER) and progesterone receptor (PR), has shown that loss of both of these receptors in tumor biopsies relates to lymph node metastasis and can increase the preoperative risk selection [31, 67] Prognosis, in this case, is inferior too. Another marker is p53, the over-expression of which is connected to lymph node metastasis too (MoMaTEC study). The dataset used in the current thesis also includes the L1 cell adhesion molecule (L1CAM), which is considered a valuable and powerful marker for the prediction of lymph node metastasis and patient's survival [23, 80, 72].

3.2 Risk factors

A risk factor can be everything that can cause a person's higher chance of developing a disease. It is necessary to clarify that the presence of one or more risk factors does not necessarily mean that the woman develops the disease. It is also possible that a woman with no risk factors present may develop endometrial cancer. It is also the case that the doctors are not in a position to be sure about whether or not a specific risk factor or risk factors present to a patient caused the disease in the first place[62]. Multiple factors can influence the risk of endometrial cancer, either positively or negatively [62]. At this point, the analysis of crucial risk factors takes place. First of all, a body-mass index (BMI) that is higher than 30 kg/m^2 triples the risk of the disease. High BMI leads to higher levels of estrogen, given that estrogen is produced in fat tissues before the patients are even considered patients at risk[10]. Additionally, unopposed estrogen is used to improve/treat postmenopausal symptoms. Methods that affect the hormone levels in this way are considered to create a high risk of getting the disease. Specifically, there is a correlation between EEC (80%) and high levels of hormones. The latter increases the risk of EEC significantly. Moreover, women with the starting date of their menstrual cycle earlier than the age of twelve and menopause that comes later than usual have a higher risk in comparison to rest. It is a fact that every woman in her menstrual cycle experiences a high level of estrogen. The higher the number of menstrual cycles, the greater the risk of cancer becomes. Studies have shown that the drug tamoxifen is related to breast cancer treatment. It can also cause the growth of the endometrium, and when it is used during menopause, it increases the risk of the disease. Furthermore, the nulliparity condition is among the risk factors of the disease. Women that are unable to get pregnant or have never been pregnant have a higher risk of developing endometrial cancer. This is because pregnancy causes a decrease in estrogen. The risk is additionally higher, in case a patient has a previous occurrence of cancer, in which the therapy consisted of radiation treatment to the pelvis or family members have also experienced the disease in the past. On the other hand, some factors help reduce the risk of the disease. Firstly, the grand multiparity condition leads to the increased production of progestagens. As a result, pregnancies protect in a way women from endometrial cancer by decreasing the risk. Smoking contributes to reducing the risk too, by influencing estrogen and the body's metabolism. Other self-explanatory factors in decreasing this risk are regular exercise, birth control pills, and phytoestrogen diet [4].

3.3 Prognosis and Survival

The factors that enable better prognosis of the disease are the surgical FIGO stage, myometrial invasion, histology, and the grade of differentiation. The separate FIGO stages represent 5-year survival. Based on [4], stage I has approximately 85% survival, stage II has 75%, and stage III has 45%, while one out of four people with stage IV disease has 5-year survival [45][1]. Myometrial

invasion affects 5-year survival significantly too. FIGO surgical staging, based on myometrial invasion, is altered more by tumor grade. The percentage is higher for low-grade tumors, which are affected by approximately 95% when deep myometrial invasion is present. On the other hand, some research has indicated lymph-vascular space invasion as an independent prognostic factor, even though there is a correlation with tumor grade and myometrial invasion. More specifically, more than a single vascular cross section should participate, although LVSI is expressed in around 37% of endometrial cancers[4][1]. Being able to recognize lymph node metastasis risk in each patient is crucial for prognosis and survival too. It is a fact that surgical findings provide valuable prognostic insights. As a result, the selection of treatment is facilitated. This way, physicians can assess lymph nodes effectively and come up with a tailored therapeutic approach [6]. Achieving successful prognostification through the use of such models preoperatively can potentially lead to improved individualized treatment. By properly classifying the patients with low risk endometrial cancer, that do not require surgery, the individuals do not undertake unnecessary treatment and healthcare costs can also be reduced.

4 Description of the data used in the research

4.1 Introduction

The research for the present thesis is conducted as part of an ENITEC study and with close collaboration with gynecologic oncologists in the Department of Obstetrics and Gynecology, at Radboud university medical centre. The initial study cohort consisted of patients that received treatment for International Federation of Gynecology and Obstetrics (FIGO) stage I-IV endometrioid endometrial carcinoma (EEC), or non-endometrioid endometrial carcinoma (NEEC) at one of the European Network for Individualized Treatment of Endometrial Cancer (ENITEC) centres that participated in the study. The patients were treated between 1995 and 2003 and they had minimum 36 months follow-up time. Patients treatment provided expert physicians with a complete set of clinical and pathological data. In total, 1199 individuals were part of this study, for whom preoperative endometrial biopsy tissue was used for analysis purposes. Additionally, preoperative endometrial biopsy slides were obtained so as the list of chosen molecular biomarkers is evaluated. The final selected cohort used for the present thesis is 763 patients that received treatment for EC. Only individuals that received their diagnosis by an expert physician in the field of gynecologic pathology were included [51] [72].

4.2 Preoperative variables

Age is incorporated in the dataset. The risk is getting higher as age increases, and it is also related to lymph node metastasis and poor survival. The age at the time of diagnosis and the patient's birth date are both known. The dataset also consists of information regarding death dates. Additionally, we are also aware of whether or not the patient's decease is a result of endometrial cancer. As body mass index (BMI) gets higher, the risk of developing endometrial cancer rises. More specifically, values over $25kg/m^2$ double the risk. Physicians consider this when examining a patient's primary state [10]. A higher level of BMI in pre-menopausal women leads to insulin resistance and severe hormone changes such as progesterone deficiency. On the other hand, in *postmenopausal* women bio-available oestradiol and testosterone result in a high increase in the number of endometrial-cells[4] Both of these variables are essential when evaluating the primary situation and possibly affect patient's risk. *Cervical cytology* is a screening process that provides useful information about the cervix. More specifically, it shows if malignant cervical cells are present or not. It is not expected to see endometrial cells in the pap smear. The presence of endometrial cells is considered to be a sign of carcinoma progression in the endometrium. It could be a consequence of the epithelial-mesenchymal transition (EMT) mechanism. The mentioned mechanism is responsible for the loss of cell adhesion, which can drive endometrial cells to the cervix [4].

Serum markers: Ca-125 is a serum marker. It is likely to have Ca-125 expressed when malignancy is not present. It is frequent when cancer becomes metastatic. According to [27], Ca-125, among other proposed predictors, is established as an effective indicator for nodal metastasis with cut-off value: 35 IU/mL.

tumor characteristics: The dataset consists of four protein indicators: tumor suppressor p53, PR (progesterone receptor), ER (estrogen receptor), and L1 cell adhesion molecule (L1CAM). According to some research, they are independent prognostic markers in the case of primary tumors [67][54]. The majority of endometrial cancers show preoperatively expression of the estrogen receptor. More specifically, a low level of *ER* has been proven to be an independent predictor for lymph node metastasis and recurrence of the disease. The same holds for *PR*. Both estrogen and progesterone receptors are expressed mostly in endometrioid endometrial cancers giving a quite good prognosis[72]. The normal behavior of the tumor suppressor gene (*P53*) protects the cells. In case it is mutated, this essential attribute vanishes. It is thus a

protein that does not work correctly, leading this way to its over-expression (especially in serous carcinomas[72]). This situation is related to poor prognosis. These markers can be present at the same time. This fact is mainly related to specific values of some variables: the patient's higher age, advanced FIGO stage, grade 3 tumor, lymph node metastasis, and non-endometrioid endometrial cancer histological behavior[67]. *L1CAM* is a protein closely associated with the molecular mechanisms of epithelial-mesenchymal transition (EMT). The over-expression of this protein leads to cell mobility and a higher risk of spreading cancer. Especially an over-expression greater than 10% has proven to be connected to the recurrence of the disease and mortality. In comparison with the other proteins, L1CAM is a far better prognostic indicator. Especially in endometrioid cancers, research has shown that the stated protein is capable of identifying the group of tumors with aggressive character and poor outcome. L1CAM should be assessed for all EC of type I, given that when increased, the protein provides much better insights than other available factors (e.g., cancer's histological grade). Furthermore, L1CAM does not associate with BMI, patient's age at the time of diagnosis, diabetes, obesity, nulliparity, hypertension, and exposure to unopposed estrogens. The physicians should perform the risk analysis based on L1CAM by testing curettage material before surgery[80]. The dataset also includes three variables corresponding to the serum markers level of white blood cells (*leukocytes*), *hemoglobin*, and *platelets*. The immune response-associated production of these markers is related to the primary tumor. *Preoperative grade* is ranging from grade 1-3. This assessment is basically about how aggressive a tumor looks. Grade 1 is well-differentiated and low-risk cancer. The second one is more aggressive and moderate differentiated, while grade 3 is the most aggressive and poorly differentiated cancer. It is assessed with an endometrial biopsy. An assessment of *histology* also takes place. It refers to the structure of the cells. There are multiple different types of carcinomas but two main categories: endometrioid and non-endometrioid. The former type can be grade 1,2 or 3 while the latter is always aggressive (grade 3 carcinomas). It is the case that some carcinomas can be both endometrioid and non-endometrioid when different parts have different cell architecture. A carcinoma is stated to be non-endometrioid when the non-endometrioid part is more than 10% of the carcinoma.

Imaging variables: The dataset consists of essential imaging variables too. Computed Tomography (*CT*), X-thorax, Magnetic Resonance Imaging (*MRI*) and Positron Emission Tomography (*PET-CT*) implement this information. Their purpose is to provide a more precise staging of the disease. The use of different types of imaging is because none of them provides as clear results as physicians would like to. The imaging methods assess the enlargement of lymph nodes. In more detail, CT's purpose is to reveal the existence of enlarged lymph nodes in the pelvis and para-aortic area. It is also used for the detection of distant metastasis, taking place either in the liver or the lungs. X-thorax's object is to detect if the cancer is metastatic and present in the lungs. MRI and PET-CT detect similar patterns with CT. However, both of them are considered to be better and more accurate methods than CT. MRI findings are significant for a more tailor-made treatment. The physicians also prefer MRI, because, it enables them to proceed with a successful selection of the group of patients, that can benefit from para-aortic lymph node dissection[43].

4.3 Postoperative variables

There is also a postoperative version of *grade* variable. It refers to the tumor grade after the surgery and reflects how aggressive it is. The same holds for *histology*. It is the result of the assessment of the cell architecture after the surgery. Additionally, *myometrial invasion* represents the depth of tumor invasion inside myometrium, which is the middle layer of the uterus. In case the tumor is present in the uterus, the FIGO stage corresponds to stage IA or IB. It is considered to be a significant postoperative prognostic factor. However, the depth of the

invasion is not known precisely. *Lymphovascular space invasion* refers to the fact that the tumor can grow into the blood vessels or lymph vessels. It is a crucial postoperative variable and a sign that the patient may have a high risk of metastatic disease [4]. *Cervical invasion* refers to the fact that the tumor can grow inside the cervix. As a critical postoperative indicator, it shows if the tumor is endocervical or stromal (grows outside the cervix). In case, the latter holds, the FIGO stage is at least stage II. *FIGO stage* shows precisely where the tumor is localized, reflecting this way the stage after surgery. In total, there are eight stages. Please refer to [4] for more information. Additional *therapy/treatment variables* are present in the dataset, i.e. variables that describe the use of hormonal suppletion therapy (either hormonal contraception or estrogen suppletion for postmenopausal complaints) and also the use of hormonal therapy for medical conditions (mostly breast cancer), which is either tamoxifen or variant. Furthermore, some of the individuals of the study receive chemotherapy, radiotherapy or a combination of them after surgery. Experience has revealed that using these methods before surgery does not yield better results in comparison with postoperative use.*Follow-up time* is also a crucial variable. The dataset contains information regarding the moment in time when the patient had the last contact with her physicians. By following an individual throughout time, it is likewise expected that information regarding *recurrence* of the disease is also stored. More specifically the type can be one f the following types. Local recurrence takes place on the top of the vagina, where the uterus is located. Additionally, regional recurrence is the one that happens inside the pelvis, including the pelvic lymph nodes. Finally, distant recurrence is the one that appears outside the pelvis area.

4.4 Selection of variables

The dataset, selected for the current thesis [51], consists of variables corresponding to biomarkers such as Ca-125, platelets while imaging methods and results are also present. Additionally, estrogen receptor (ER), progesterone receptor (PR), L1 cell adhesion molecule (L1CAM) and p53 were chosen for immunohistochemical staining on biopsy samples that were received preoperatively. The loss of ER and PR is confirmed as independent prognostic variables for the prediction of lymph node metastasis (LNM) [67]. The loss of ER is closely correlated to epithelial-mesenchymal transition (EMT) mechanism [4]. L1CAM has been also confirmed as powerful prognostic marker in endometrial cancer, which is related to EMT as analysed earlier [22] [72]. On top of that, research has shown that p53 is closely related to patients that received poor prognosis [33]. Additionally, the individuals received adjuvant therapy based on existing protocols, applicable in each hospital and this information is taken also into consideration here. With respect to outcome variables, the data describe the presence of lymph node metastasis (pelvic, para-aortic), disease recurrence and disease-specific survival at one, three and five years. The disease recurrence is incorporated in the data in three different categories; local recurrence which corresponds to vaginal vault, regional recurrence involving pelvic structures and distant recurrence in case of other types of recurrence present in each patient. Moreover, patients' results of preoperative processes such as cervical cytology are listed too, supporting effectively the process of building the prediction model later on. The selection of variables with potential prognostic power took place by the physicians after careful and systematic research and review of existing literature [52].

Variable Names	Cutoff value(s)
Preoperative Variables	
Age	$< 70; \geq 70 years$
Body Mass Index (BMI)	$< 25; \geq 25 kg/m^2$
Hemoglobin	$< 12; \geq 12 g/dl$
Leucocyte counts	$\leq 10 \times 10^9; > 10 \times 10^9/l$
Thrombocyte counts	$< 400 \times 10^9; \geq 400 \times 10^9/l$
Ca-125 serum levels	$< 35; \geq 35 IU/ml$
Lymphadenopathy on MRI or CT	No; Yes ($\geq 10 mm$ short axis diameter)
Cervical cytology	No; Yes (atypical endometrial cells present)
Tumor grade	1; 2; 3
Tumor histology	Endometrioid; Non-endometriod
Preoperative molecular biomarkers	
ER expression	$< 10; \geq 10\%$ of tumor cells with nuclear staining
PR expression	$< 10; \geq 10\%$ of tumor cells with nuclear staining
L1CAM expression	$< 10; \geq 10\%$ of tumor cells with membranous staining
p53 expression	Wild type; Overexpression with cutoff value 40-50%
Postoperative Variables	
Myometrial invasion	No invasion; Invasion <50%; Invasion ≥50%
Lymphovascular space invasion (LVSI)	No; Yes
Cervical invasion	No; Yes
FIGO stage	IA; IB; II; IIIA; IIIB; IIIC; IV
Tumor grade	1; 2; 3
Tumor histological subtype	Endometrioid; Non-Endometrioid
Adjuvant therapy	None; Radiotherapy; Chemotherapy; Chemoradiation; Other

Table 2: Candidate predictor variables for constructing the Bayesian network

5 Methods and results

5.1 Tools for data analysis & visualization

Bnlearn is an R package for performing learning of Bayesian networks. It is used for parameters estimation and inference procedures. It started being used in 2007 and it has been under development since then [58].

More specifically, the package helps us manually set up the structure of the network and learn the parameters by providing the user with joint conditional probability tables. The process and the details of the steps used are explained in a different part of the thesis. The model is implemented by using two structure learning algorithms that are incorporated in this package i.e. Hill-climbing and Tabu search algorithms. As it is analyzed later on, a large number of structures are explored and by using as a resampling method, bootstrapping, the effect of local optimum BNs on learning is reduced. Different functions of the package are used to acquire the averaged BN and measure the strength of the arcs incorporated in the learnt structures. Given this strength, we are able to draw conclusions in respect to the significance of the arcs and the necessity of their presence in the Bayesian network. Validation of the models is also supported by the package and finally inference can be also demonstrated. However, *SAMIAM* software supports us in the inference process in this thesis project. It is an easy to use tool for modeling and reasoning, developed in Java. Visualization of the network is done with SAMIAM and by interacting with it, we are able to perform inference and explore its behavior, based on multiple inputs. The software offers an interface where the user can build and interact with the Bayesian networks, which can be also saved in multiple formats. The reasoning process incorporates inference, parameter estimation, time-space trade-offs, sensitivity analysis, and explanation-generation based on MAP and MPE [37]. Additionally *GeNIe Modeler* is also used for the visualization part. It has graphical user interface and also supports the user during the interactive modelling process and learning.

5.2 Summary of the steps

The main idea is that given the limited size of the dataset and the considerable number of missing values, expert knowledge of cause-effect relationships is used to support and guide the structure learning process. First of all, we attempt to reveal the direction of the arcs, pointing from cause to effect, that connect the different variables. Additionally, it is checked if the results obtained from the structure learning process can be explained sufficiently. Experts knowledge supports us in reducing the search space of structure learning by indicating relationships between variables that are really strong or the ones that should not exist. It is important to incorporate expert knowledge in the model given their understanding of the fundamental elements of the domain and the causal relationships between multiple factors. However, it is worth noting that determining the model solely according to physicians knowledge can lead to biased results that do not reflect adequately the problem area and the dataset. Hence, the contribution of causal graphs as tool for knowledge elicitation is vital. This chapter analyses in depth how the data is being prepared for the analysis, how we deal with incomplete data and perform parameter learning to calculate the joint probability distributions. In general, all the different steps involved in the Bayesian model implementation are defined here. It is also explained how the user can interact with the model and perform inference to extract valuable information in the form of posterior distributions. A simple cox model is also implemented and Brier scores are calculated for both methods. A detailed qualitative comparison is also attempted to highlight how beneficial BNs can be as prognostic models in the medical world.

5.3 Visual representation

The first step of the process consists of discussions focusing on the relationships between the available variables of the dataset. The goal is to discover a logical way to connect the variables, in order to obtain the first visual representation of the problem and not an actual Bayesian network.

In this network model representation, we attempt to demonstrate all the possible logical connections and cause-effect relationships from a biological point of view. This helps us build a starting representation to visualize the problem and understand more about the domain and the available dataset. At the beginning, the aim is to gain some insights into how the gynecologic oncologists deal with endometrial cancer as a disease and their thinking process when they need to understand patients condition and proceed to their treatment. The first important point in which we should focus on is the immune response. The reaction of human body towards the existence of a substance that is not identified as part of the body itself causes the variability of the serum levels of Hemoglobin, Platelets and white blood cells (Leukocytes) and this relationship is represented by arcs connecting the different nodes.

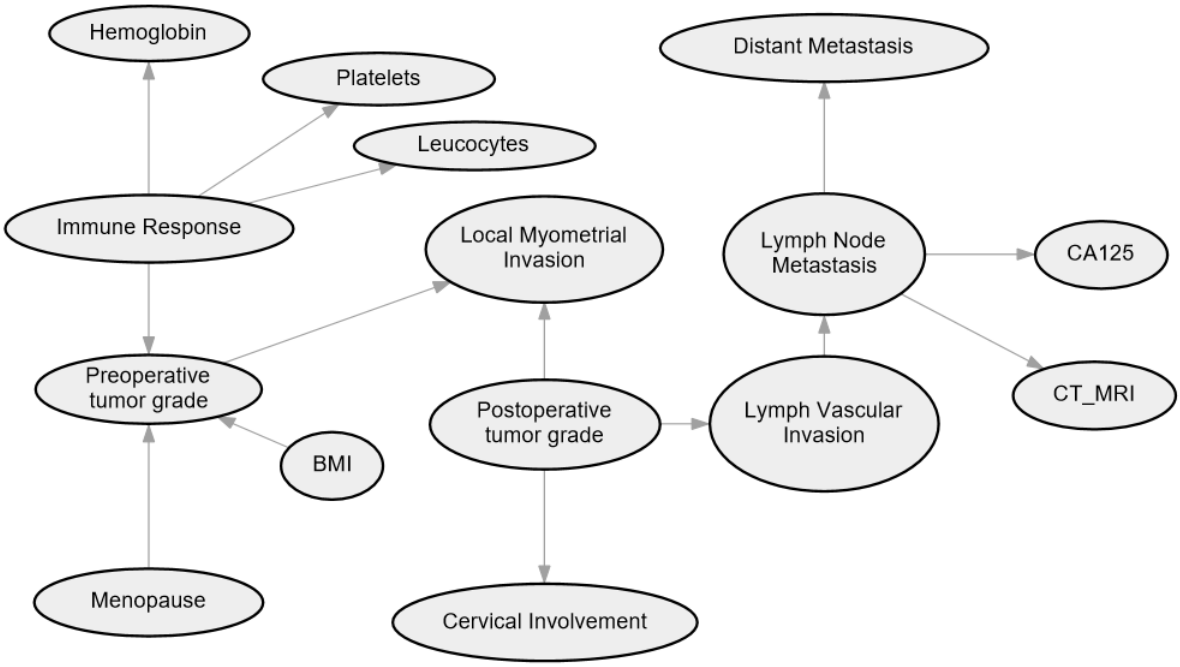


Figure 8: Initial causal network model based on clinical evidence

On top of this, the preoperative grade assessment in a patient is done via an endometrial biopsy. The mentioned core action that physicians perform is present in this representation as the *preoperative tumor grade* node and it is closely related to the immune response. The way that the body of the patient reacts helps physicians assess the situation preoperatively. This is why a direct connection exists between these two. Moreover, a connection between *menopause* and *preoperative tumor grade* reflects the fact that the closer the patient is to menopause the higher the risk gets. Likewise, the *preoperative tumor grade* node is connected to *BMI*, given that according to physicians the risk changes drastically relatively to patient's BMI. When a physician performs the assessment of a grade, s/he is not able to exactly determine the depth of the carcinoma, which is known as local myometrial invasion. However, according to the assessment, s/he understands how s/he should proceed with the patient. Thus, the *preoperative tumor grade* is directly connected to *local myometrial invasion*. More specifically, the higher

the preoperative grade is, the higher the risk of deep invasion to myometrium gets. Something similar holds for the postoperative grade which is represented here by *postoperative tumor grade* node. Both preoperative and postoperative grade assessments are directly related to myometrial invasion according to literature and this is why the graph reflects the same. Being able to provide prognosis preoperatively is highly crucial for the patients. However, the physicians are able to determine if the carcinoma grows inside or outside of the cervix only postoperatively.

Given that this is an important postoperative prognostic factor, a connection between *postoperative tumor grade* and *cervical involvement* is present too. The postoperative grade is additionally closely associated with lymphovascular space invasion (LVSI). The physicians can only postoperatively see if the carcinoma grows into the blood-vessels or lymph vessels, revealing this way tumor's metastatic trend. As a result, logical connections of *LVSI* and both *lymph node metastasis* and *distant metastasis* nodes are drawn. If the physicians consider it necessary for the patient, lymphadenectomy is performed and the lymph nodes in the pelvis and around the aorta are removed during surgery. This way the doctors can see if the removed lymph nodes are positive for carcinoma or not. Hence, this is considered to be an important postoperative prognostic factor.

Furthermore, recent literature reveals the importance of different serum markers; one of them is *Ca-125*. The node representing this serum level is placed in a direct connection to *lymph node metastasis*. This happens because this tumor marker is elevated in multiple malignant conditions, such as endometrial cancer. Additionally, It is noticed that it is connected to *lymph node metastasis* and not *preoperative* or *postoperative tumor grade* nodes; This is because Ca-125 is mostly elevated when a metastatic disease is present and not when the presence of carcinoma is inside the uterus only. The last node of this representation, *CT-MRI*, indicates the importance of Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scan to detect mostly metastatic disease in lymph nodes.

5.4 Data processing

The first step before the analysis is the pre-processing of the data. Based on expert knowledge a subset of the variables is selected and renamed. The purpose it to have meaningful names for the nodes of the Bayesian model, when the visual representation is created. To continue, we proceed with careful transformation of the data. The dataset is checked for values that do not make sense e.g., 99. In this case they are considered to be missing values and they are replaced with *NA*. Categorical variables are also transformed to factors. On top of that, the values of the variables are replaced to more meaningful ones. For this reason, we define the states of the variables in a convenient way e.g., values of *Primarytumor* variable: 1 ,2 ,3 are replaced by *grade1, grade2, grade3* respectively. This way, we can properly interprete the differences in the probabilities that are reflected to the Bayesian network. Please refer to the appendix for the code used to clean and refine the dataset.

5.5 The Bayesian network development

5.5.1 Introduction

The Bayesian network is constructed step by step by using multiple iterations. First of all, the most important variables for tumor progression are considered. In detail, postoperative tumor grade, myometrial invasion (MI) and lymphovascular space invasion (LVSI) are among the primitive nodes of the network. The mentioned variables are not used for predicting the outcomes and as a result, a list of preoperative predictors are also introduced to the model. The table below demonstrates the baseline attributes of the development cohort. It is worth mentioning that 215 individuals provided the experts with information on the entire list of the

variables in the data.

After discussing with the clinical doctors about the domain and creating the logical representation, the problem becomes more clear. Given its complexity, it is expected that a lot of changes and adjustments are made during this step-wise modelling approach. The variables are illustrated as nodes in the network and their relationships are shown as arcs. The arcs are directed, which reveals causality. The improvement of the network takes place by manually interacting with the structure, after careful data-driven thinking.

By adding or removing arcs according to expert knowledge, we are able to refine the model a bit further. It is important to mention that we evaluate several times during the implementation process how well each network fits the data by calculating its log-likelihood. The closer this value is to zero, the better the model gets. Additionally, we manage to improve the model by incorporating structure learning algorithms results. This concept and its implementation is analysed later on.

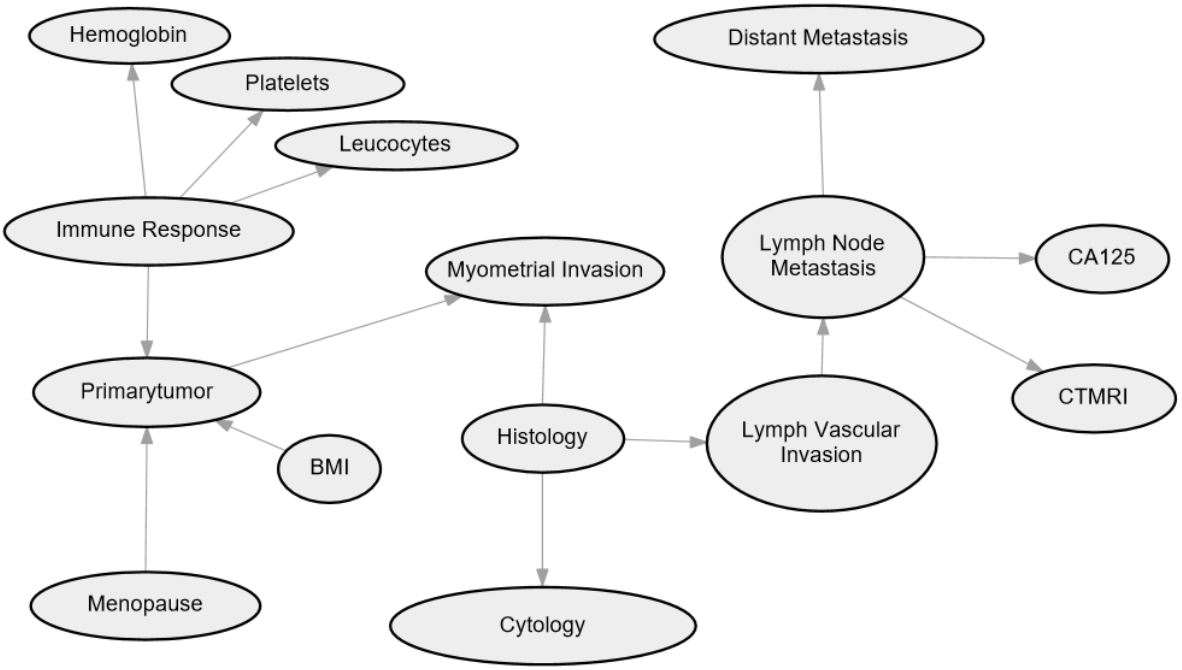


Figure 9: The first Bayesian network model

At this point, after translating the initial representation to the first Bayesian model, the DAG looks as illustrated in figure 9 After the first discussions with the clinical doctors and before evaluating the structure by using log-likelihood, we have been requested to incorporate more variables that are crucial for the progress of the disease. Specifically, as it is reflected in the graph that follows, connections between *ER, PR, p53, L1CAM* and *Primarytumor* are drawn. This is suggested by the physicians given that they can get an idea of what the tumor grade might be by looking into these metrics preoperatively. Of course, these first graphical representations express solely the doctor's perspective, based on their everyday way of working. An extra arc is additionally placed between *Cytology* and *Primarytumor*, given that a cervical swab can assist doctors in the preoperative examination of the tumor. The same holds for the extra connection between *Primarytumor* and *Lymph node metastasis* node. Concerning the connection between *Histology* and *Lymph Vascular invasion*, since physicians were in doubt about whether or not a direct connection makes sense, at this point it is left out. We re-examine different scenarios regarding this pair of nodes later on.

Variable Name	Development Cohort (N=763)
Age	65 (58-71)
Body Mass Index (kg/m^2)	29 (26-33)
Follow-up time (months)	60 (45–74)
Tumor grade, preoperative	
1	372 (48.8%)
2	173 (22.7%)
3	110 (14.4%)
Unknown	108 (14.2%)
ER expression	
Positive	686 (89.9%)
Negative	76 (10.0%)
Unknown	1 (0.1%)
PR expression	
Positive	620 (81.3%)
Negative	137 (18.0%)
Unknown	6 (0.8%)
L1CAM expression	
Positive	79 (10.4%)
Negative	665 (87.2%)
Unknown	19 (2.5%)
p53 expression	
Wildtype	584 (76.5%)
Mutant	112 (14.7%)
Unknown	67 (8.8%)
Ca-125	
$\leq 35IU/ml$	318 (41.7%)
$\geq 35IU/ml$	90 (11.8%)
Unknown	355 (46.5%)
Thrombocytes	
$< 400times10^9/l$	557 (73.0%)
$\geq 400times10^9/l$	25 (3.3%)
Unknown	181 (23.7%)
Imaging Results	
No lymphadenopathy	460 (60.3%)
Lymphadenopathy	38 (5.0%)
Unknown	265 (34.7%)
Cervical Cytology	
Normal	406 (53.2%)
Abnormal	27 (3.5%)
Unknown	330 (43.3%)
Tumor grade	
1	317 (41.5%)
2	289 (37.9%)
3	157 (20.6%)

Table 3: Development cohort baseline attributes

Variable Name	Development Cohort (N=763)
Histological subtype	
EEC	714 (93.6%)
NEEC	49 (6.4%)
Unknown	
Myometrial Invasion (MI)	
$< 50\%$	477 (62.5%)
$\geq 50\%$	283 (37.1%)
Unknown	3 (0.4%)
Cervical invasion	
No	591 (77.5%)
Yes	86 (11.3%)
Unknown	86 (11.3%)
Lymphovascular space invasion	
No	435 (57.0%)
Yes	96 (12.6%)
Unknown	232 (30.4%)
Enlarged lymph nodes	
Negative	440 (57.7%)
Positive	53 (6.9%)
Unknown	270 (35.4%)
Treatment	
None	415 (54.4%)
Radiotherapy	283 (37.1%)
Chemotherapy	38 (5.0%)
Chemoradiation	26 (3.4%)
Hormonal	0 (0%)
Unknown	1 (0.1%)

Table 4: Development cohort baseline attributes (continued)

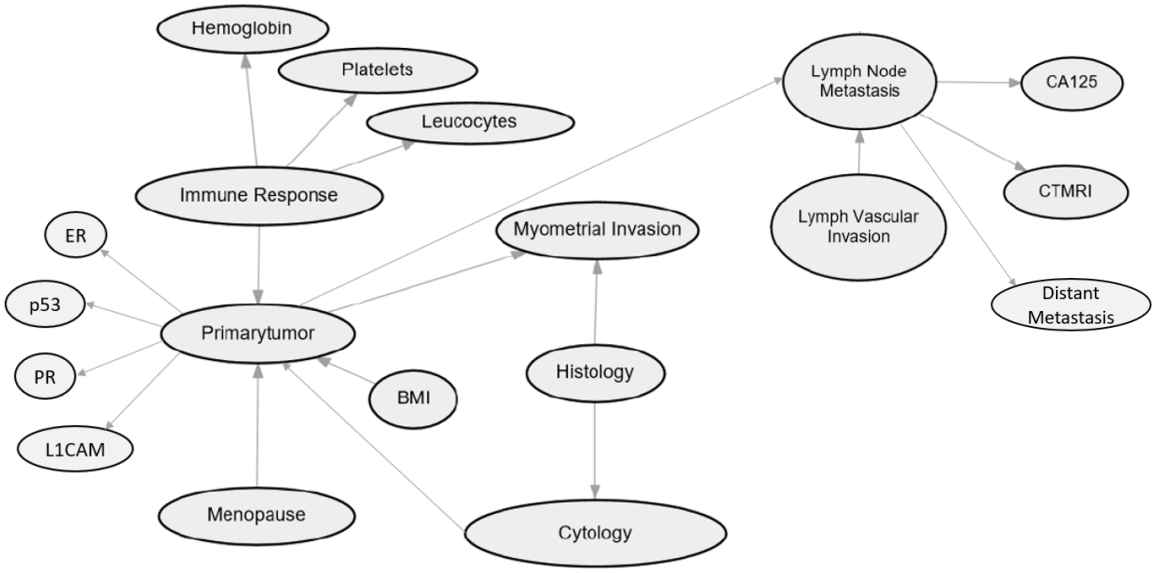


Figure 10: Enriched first Bayesian network model based on expert knowledge

5.5.2 Learning parameters & dealing with missing data

The first step to build the Bayesian network is to create a model string in R. The direction of the arcs is defined and we incorporate cause-effect relationships in the model. After this, the model string is transformed to a directed acyclic graph (DAG). Based on this DAG, each node's conditional probability table (CPT) is computed as explained earlier. In detail, the *bn.fit* function of the package uses the EM algorithm to fit the parameters of local distributions given the available data and the structure of the DAG. It is important to note that given the definition of a BN, the parameter learning process takes place only when the total number of arcs, that are present in the network, are directed, and no plain connections exist between the nodes. The imputation is done with the *impute* function, which uses as compulsory arguments the fitted BN object and the subset of the dataset, with the exact number of variables that are present in the model. There are two possible methods to perform imputation with this function: the *parents* method which is the default one and the *bayes-lw* method. The former method calculates the missing values of the parent nodes of a node in the local probability distribution incorporated in the fitted object. The latter method, which is the one used in the experiments of the current thesis, computes the missing values by using all the nodes of the DAG as evidence (without the node for which the missing values are calculated). Specifically, the missing values are calculated by incorporating the average likelihood weight measures for the total number of nodes in the model. The user is able to define the number of the randomly selected samples n, that are averaged, producing this way the new observation. Given that the subset of the dataset that is used consists of discrete variables, the missing value is replaced by the level of the variable that has the highest conditional probability. There is also the option to limit the number of nodes used to compute the missing values, however we use all the available nodes in the model. After imputing the data, the dataframe is checked for missing values before proceeding. A new fitted object is now created with the imputed data as input and the SAMIAM software helps us visualize the result and interact with it.

5.5.3 Goodness-of-fit & model enrichment

A really important point is to check if the model fits the data. In the current thesis, as stated earlier, many different scenarios are checked and the development of the Bayesian model follows a step-wise implementation approach before reaching to its final state. The various adjustments are discussed in the current section. The reasons driving these changes are the collaboration with the gynecologic oncologists and the effort to incorporate their knowledge with the best possible way into the model. Each resulted version of the Bayesian network is evaluated by initializing the network and examining the way that the model behaves. On top of this, the log-likelihood of each BN is computed. This happens by using the fitted object that "carries" the parameters of the Bayesian model. After discussing in depth the arcs in terms of causality and evaluating the BN, a great number of changes takes place. Specifically, for the computation of the log-likelihood and the comparison of the different models, we use Bnlearn package in R and *loglik* function. It has to be stated that log-likelihood calculation makes sense only in terms of model comparison. The closest a model's *loglik* is to zero the better the model fits the data.

As a starting point, after introducing the initial Bayesian model the *loglik* is equal to *-4796,561*. In the steps that follow, the goal is to refine the model as much as possible, using medical expert knowledge, before we proceed with using structure learning algorithms. First off, the connection between *Primary tumor* and *Lymph node metastasis* nodes is deleted as the preoperative assessment of the tumor grade through endometrial biopsy does not have direct relationship with *Lymph node metastasis*. Given that the goal is to include only the real meaningful connections to the Bayesian model, by keeping the number of arcs as limited as possible, the connection between *Primary tumor* and *Cytology* is also removed. On the other hand, after careful analysis, an arc from *Cytology* to *Histology* is added. The architecture of the cells, expressed by *Histology* variable, can be more easily related to *Cytology*, given that the latter represents the information driven by a cervical swab, which basically examines the existence of malignant endometrial cells in the cervix. Additionally, the connection between the preoperative grade, expressed by *Primary tumor* variable, and local myometrial invasion; *MI* is also deleted. The underlying reason for this is the fact that the doctors are not able to identify preoperatively, how deep the tumor grows inside the middle layer of the uterus. Furthermore, a direct connection between *Ca-125* marker and *CT-MRI* should be present. As a tumor biomarker, *Ca-125* is increased in multiple malignant conditions such as endometrial cancer. Given that both Computed Tomography (CT) scan and Magnetic Resonance Imaging (MRI) scan are used to detect either distant metastasis and/or the presence of enlarged lymph nodes, we decide to combine both variables in one, due to the limited information available for each. Regarding the connections of *ER*, *PR*, *L1CAM*, *p53* with *Primary tumor* variable that are introduced in the first Bayesian network model, they should be deleted. Instead, these should be connected to *Histology* and *Lymph node metastasis* nodes. This occurs because, by studying these variables, the experts are able to get more clarity on the architecture of the cells, which corresponds to *Histology* variable and not gain insights into how the tumor looks like preoperatively, which is what *Primary tumor* variable actually reflects. The direct connection between *Lymph node metastasis* and the mentioned markers, is a change driven by the fact that these are crucial predictors for the disease spreading to lymph nodes. After performing these adjustments to the Bayesian model, an increase of the log-likelihood confirms that the new version fits better the data. Specifically log-likelihood is now equal to *-4765,79*. Further modifications, requested by the physicians, are the elimination of the following nodes: *BMI, Distant Metastasis, Menopause*. The addition of these variables does not influence the behavior of the model in any way, that would be interesting or useful for the patients treatment. The goal is to include only the variables that can actually contribute to the model. Similarly, we chose to exclude *hemoglobin, leukocytes*, the presence of which does not yield improved performance [52]. Moreover, *Recurrence* variable is added. Connections are established with the following variables: *Lymph node*

metastasis, Ca-125, Histology, Lymph Vascular invasion. This is due to the fact that these variables are considered to be the most important predictors for disease recurrence according to literature. After performing these modifications the log-likelihood of the Bayesian model is calculated to evaluate how well the updated version fits the data. Indeed, the log-likelihood is *-4333,137*. After the mentioned steps of extending the model and refining the structure solely based on experts knowledge, the BN looks as shown in the graph below. In the following section, the survival variables are introduced and explained. The structure learning process takes them into consideration and thus further adjustments are implemented into the model. The clinical doctors request us the addition of *Therapy* variable too. More in depth details about this step are given later on.

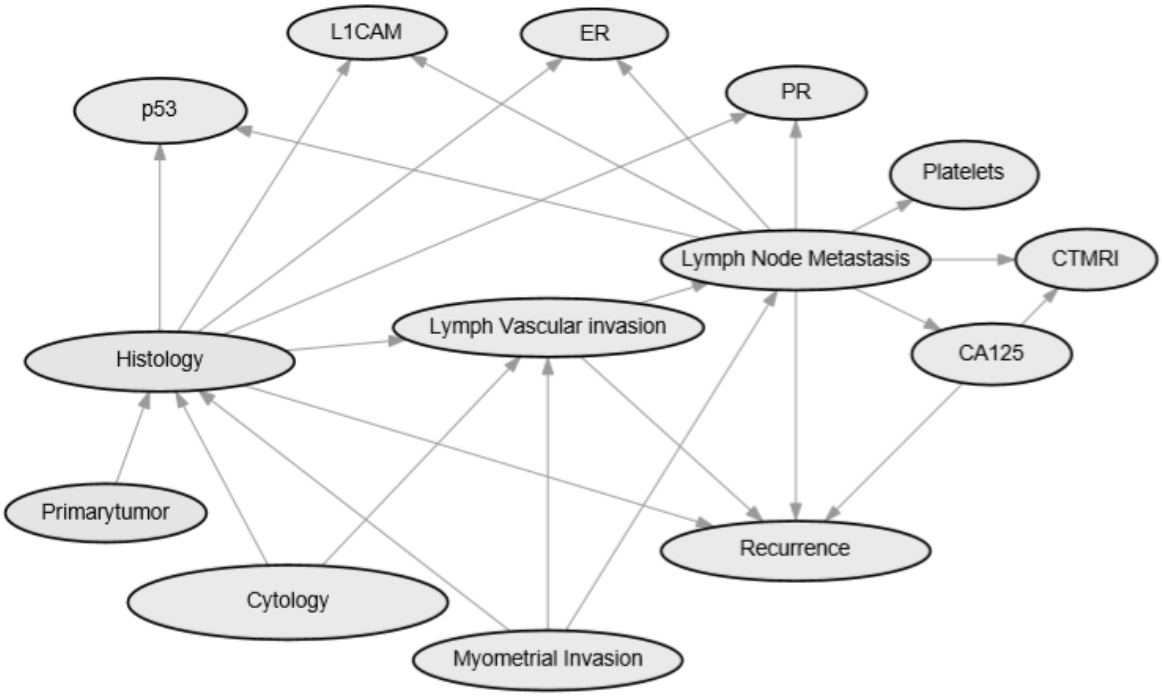


Figure 11: Bayesian model, after improving the structure and before applying structure learning algorithms

5.5.4 Survival variables

As already explained the modeling process consists of multiple in-between steps in which the model is refined and extended. After performing multiple changes, we extend the model by adding the survival variables too. The survival probabilities are calculated by using Kaplan-Meier estimator which is described in detail in a previous chapter. The goal is to evaluate if the survival probabilities produced in the Bayesian model are approximately the same as the resulted probabilities after using the Kaplan-Meier estimator method. An important point is that the dataset consists of patients for whom information is known up until three years after the starting point of the research (right censoring). The three survival variables are generated by using:

- the variable that reflects if the patient died because of endometrial cancer or not,
- the variable describing the follow-up time and
- the time of death

In the following table, information for six sample patients are reflected. As it makes sense, the patient that is reported dead after 14 months (second patient), has three-year survival and five-year survival variables which are both equal to *No*. Additionally, the fourth, fifth and sixth patients have an unknown outcome and as a result they are considered censored observations.

Patients	*Death by EC*	*1 Year Survival*	*3 Year Survival*	*5 Year Survival*	*Follow-up*	*Censored*
1	Yes	No	No	No	11 months	No
2	Yes	Yes	No	No	14 months	No
3	Yes	Yes	Yes	No	40 months	No
4	No	Yes	Yes	Unknown	41 months	Yes
5	No	Yes	Unknown	Unknown	15 months	Yes
6	No	Unknown	Unknown	Unknown	6 months	Yes

Table 5: Survival variables for sample data

Product limit estimator of the survival function S, allows us to calculate the probability that someone survived for a period longer than t. In this equation, t_i corresponds to the time, when at least one event took place, d_i reflects the number of deaths or in general the number of patients that experience the event, at a specified time t_i. Finally n_i corresponds to the patients that do not experience the event or are censored at time t_i

$$S(t) = \prod_{i:t_i \leqslant t} \left(1 - \frac{d_i}{n_i}\right) \tag{24}$$

The calculations of survival probabilities by using the mentioned estimator are as shown in the table below. It is worth noting, that these calculations produce similar results as the ones generated in the Bayesian model. This is further shown later on.

Period	*At risk*	*Censored*	*Died*	*Survived*	*Kaplan-Meier Survival Probability Estimate*
0-1year	763	4	16	747	0,979030144
1-3years	743	17	27	716	0,943453006
3-5years	699	294	16	683	0,921857515

Table 6: Survival probabilities using Kaplan-Meier estimator

5.5.5 Structure learning algorithms

Score-based structure learning algorithms are used in the present thesis, to enrich the Bayesian model and investigate the strength of the connections between the variables. Specifically, we use *Hill-climbing* and *Tabu* algorithms. The aspiration is to come up with the best possible network, given the limited number of data available, the missing values and existing noise. In order to deal with these points, the strength of the arcs in the model are measured by using the bootstrapping technique. From the multiple networks that come out of the structure learning algorithms, we are interested in the strongest arcs. These are the ones that are incorporated to the model which has been built so far.

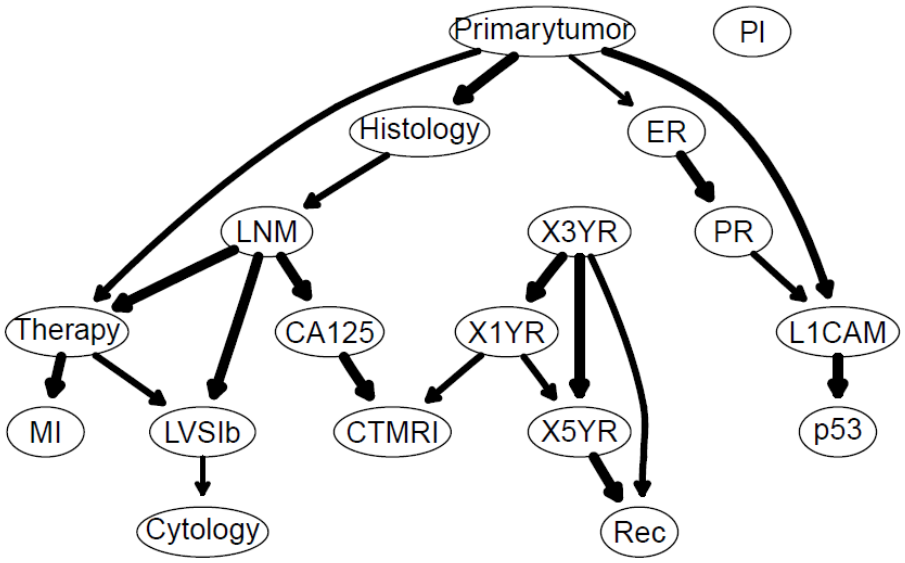
Bnlearn package in R consists of functions that facilitate the steps that are executed for both structure learning algorithms. First off, *boot.strength* performs non-parametric bootstrapping and is used for the assessment of arc strength and direction. In detail, each arc's strength is calculated as the number of times the arc is present in a network divided by the total number of Bayesian networks generated from bootstrap samples. The function uses the imputed dataset

containing only the variables that are present in the model. The probability of each arc is then calculated and on top of this the function computes the probability of each arc's direction based on the present arcs in the model [55]. Given that the *bn.strength* result of this function is not a single Bayesian network, the next step is to perform model averaging. For this goal *averaged.network* function is used. Hence, the result is a network consisted of a set of arcs, which are present in more than a pre-specified percentage d of the resulted Bayesian networks. In example, if the threshold d is set to be equal to *60%*, then the resulted graph reflects the arcs that are present in at least *60%* of the BNs produced. The result of this step is drawn below by *strength.plot* function that shows the generated networks both for Hill-climbing and Tabu algorithm. In these representations the strongest arcs are shown with wider arcs. According to literature, a set of arcs is considered significant in case it is present in at least 85% of the networks and with the most common direction (i.e. the direction that is present in more than 50% of the networs) [56]. However, in the present thesis, and due to the limited number of data, different thresholds were examined. The results were discussed thoroughly with the physicians. The aim is to end up with a list of arcs that could be used to successfully enrich the initial Bayesian network. These algorithms assign a specific score to each candidate network and the expected outcome is a network that maximizes this score. Given the resemblance of the two structure learning algorithms, which is analyzed in a previous section, they behave as expected in a similar way. In the following graphs, the result of both algorithms are shown. For the calculations, the threshold is set to 70%, to reflect the connections that are present to at least 70% of the models produced through the process.

By carefully examining the result of the algorithms (figure 12) and discussing with the doctors, multiple extra adjustments are made to the previously created Bayesian network model.

In detail, the first connection from *Histology* to *LNM* is necessary, given the causal relationship that exists between the architecture of the cells and the presence of enlarged lymph nodes. Additionally, we notice that the existing arc between *MI* and *LNM* is highly important. The deeper the local invasion is, the higher the risk of lymph node metastasis gets. When the tumor is not limited to the uterus, it might grow to the cervix and spread. As an important postoperative prognostic factor, the knowledge of the depth of myometrial invasion supports physicians to estimate the risk of lymph node metastasis.

Additionally, in respect to estrogen receptor expression (*ER*) connecting to progesterone receptor expression (*PR*), most endometrial cancers are hormone-dependent, which basically means that they are driven by increased hormone levels. This is why females with high hormone levels (e.g. females with high BMI or females with hormone-replacement therapy) have higher chance of developing the disease. For this reason, the majority of endometrial cancer cases are ER and PR positive, which basically means that they express receptors for both these hormones. The loss of hormone receptors is related to a previously mentioned process, *Epithelial to Mesenchymal transition*, which relates to cells' more aggressive phenotype. An aggressive phenotype means that the connection between the cells gets lost, as they become more irregular, which leads to an easier metastasis. Strong connection is established between *L1CAM* and *p53* too. The markers are frequently expressed in the same way, by being either both positive or both negative. In non-endometrioid tumors L1CAM is positive and almost all of them express p53 mutation. On the other hand, endometrioid tumors express L1CAM and p53 less frequently [38] [74]. Additionally a less important connection is revealed via both structured learning algorithms; Between *PR* and *L1CAM*, which extended physicians perspective and it is therefore included.



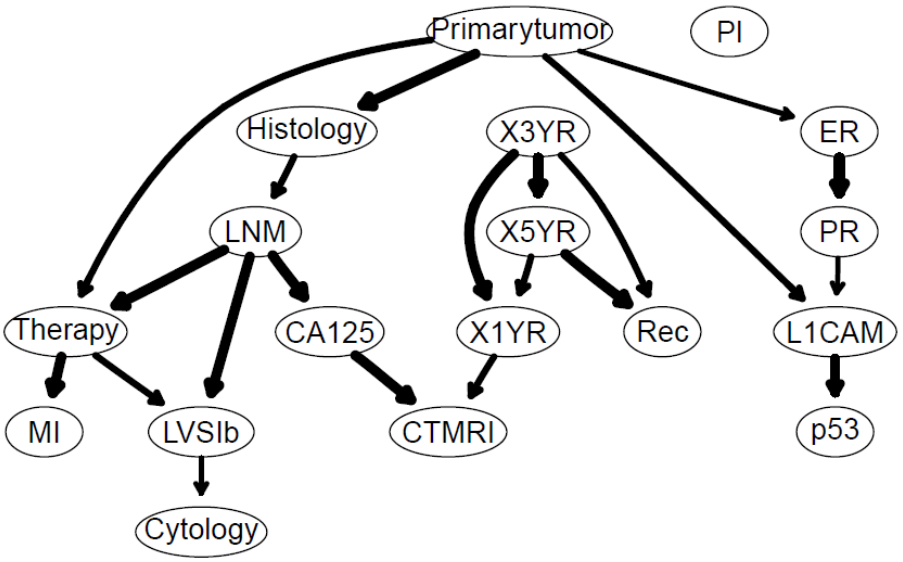


Figure 12: Structure learning algorithm results using bootstrapping

Furthermore, when doctors examine the architecture of the cells are able to determine a type of therapy. This is why an arc between *Histology* and *Therapy* is added, even though the connection exists indirectly in the Hill-climbing and Tabu results. Both algorithms reveal also a strong correlation between *Therapy* and *LNM*, which of course is something expected. The choice of the therapy affects the risk of metastasis.

Last but not least, the three survival variables are introduced to the Bayesian model and they are connected to *Recurrence* and *Therapy* variables after the physicians identified causality among them. The selection of treatment that is used for each patient can affect their survival. Additionally, the type of recurrence, if any, affects the risk of death. The final Bayesian model after all the necessary changes is as shown in the figure 13.

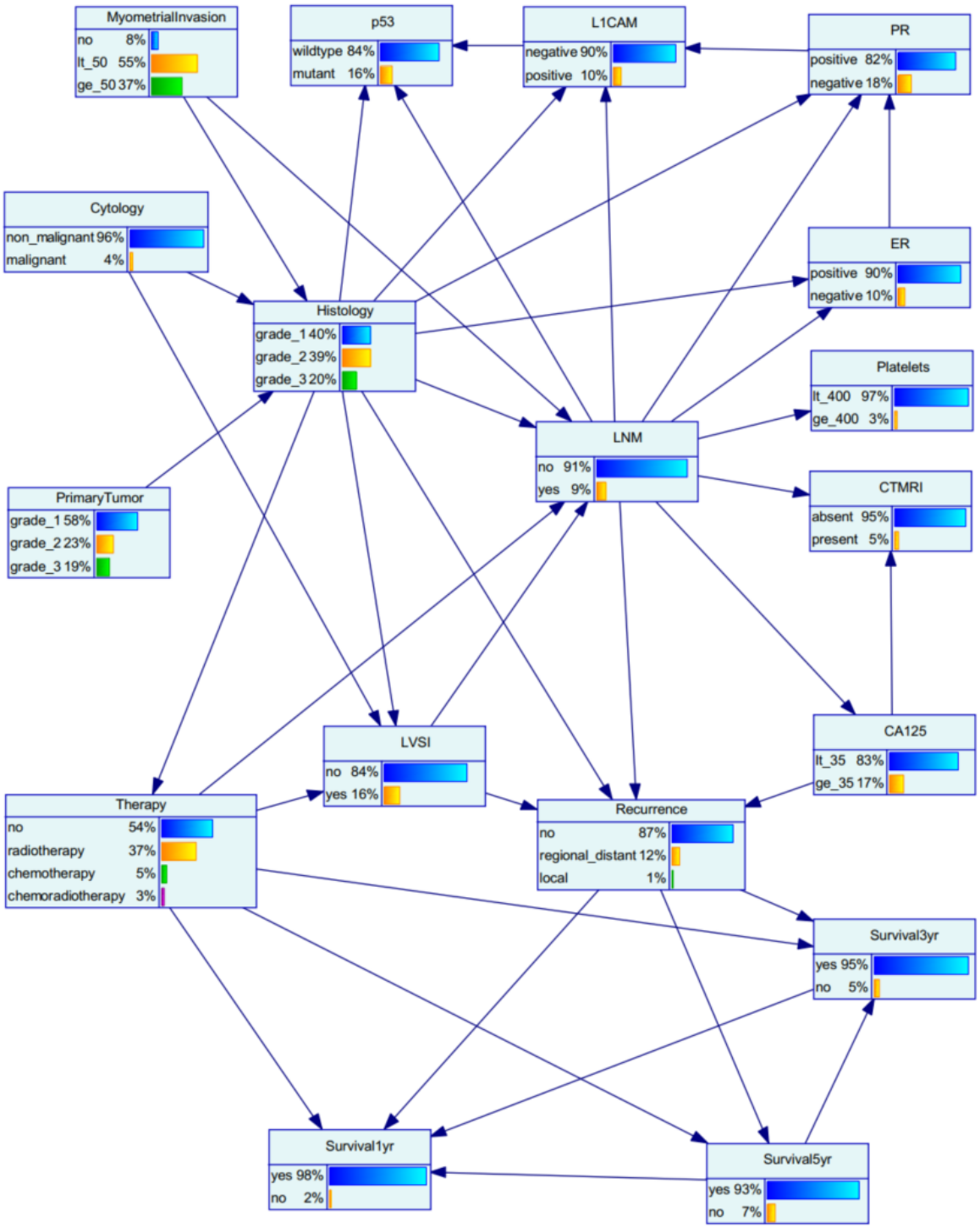


Figure 13: Final Bayesian model, including survival variables, after incorporating structure learning outcome, without any evidence inserted

5.5.6 Inference

At this point, the focus is the interaction with the Bayesian model and how it can be effectively used. As previously mentioned, Bayesian models can be queried and provide information based on knowledge about specific variables. We attempt to show how the Bayesian model can provide us with insightful answers and demonstrate how much the answers agree with expert knowledge

and available research on this subject, being a promising tool as a prognostic decision support system.

The user can initialize selected variables, by providing a specific input as evidence and observe model's behavior. In the case shown below (Figure 14), the experts are able to choose specific variables of interest e.g. *Histology* and *ER* and inspect how lymph node metastasis probability distribution is affected, how the survival variables change or acquire insights about possible future recurrence of the disease. The probability distributions are reflected in each node and variable dependencies are demonstrated by the arrows. When no direct or indirect connection is present between the nodes, they are considered independent. The variables in which one state is equal to 100% and the remaining 0% are the initialized variables, for which evidence has been inserted to the model. When input is given to the network, the probability distributions of the remaining nodes are affected accordingly. More specifically, given a grade 3 histology and negative ER, the probability of enlarged lymph nodes is increased by 22%, while the probability of negative PR is high and equal to 89%, which is something expected, given that *ER* and *PR* are usually co-expressed. Additionally, probability estimates of the entire set of the present nodes are derived each time, as said. This is why, differences are noticed in other nodes of the network, in comparison with figure 13, where no input has been provided to the network.

In general, the clinical doctors provided us with some sample results based on the model and their expectations. Some of them are listed in the table 7. The first column corresponds to the selected variables that are initialized and their chosen values. The column *Expectation*, indicates the values that should be reflected to the BN based on expert knowledge around the respective variables. The *Observation* column represents the answers that the BN provides to the users. It has to be noted that given the nature of the Bayesian model, it is not completely correct to expect exactly the same results as the second column indicates and based on existing literature. This is due to the fact that the BN model incorporates the cause-effect relationships of multiple variables and the reflected probability distributions are learnt based on all variables in the network. It makes sense that performing inference in such models provides more sophisticated results as the entire BN contributes to them. It is still valid, though, to look into the general patterns reflected on the model and compare with the expected values. In case, a patient has *Histology* of grade 3, it is expected a 20-25% probability of *LNM*, while the BN indicates a probability of 26%. The preoperative evaluation and prediction of enlarged lymph nodes is not really accurate nowadays. From the literature, it is known that the architecture of the cells, on top of myometrial invasion and cervical involvement, has high predictive power in respect to lymph node metastasis [40][13]. This shows that indeed this variable could be used effectively to reveal high potential risk of lymph node metastasis preoperatively. The BN reflects many connections of the said variable with multiple others, demonstrating direct cause-effect relationships. Moreover, the positive expression of *L1CAM* is associated with poor outcome of the disease and proven to be a good predictor of lymph node metastasis [66]. This is shown in the network, given that the probability of *LNM* is increased by 16%, when *L1CAM* is positively expressed. *L1CAM* is also correlated to the presence of more aggressive disease, loss of hormones *ER* and *PR* and reduced survival. The study [72] is in line with this fact, by revealing the association of the mentioned variable with enlarged lymph nodes, the presence of high grade disease and metastasis. The incorporation of *ER* and *PR* in the network is really important in terms of lymph node metastasis prediction and poor outcome prognosis. These proteins are co-expressed and predict independently lymph node metastasis [67]. It is indeed noticed a 14% increase of *LNM* probability and an 8% increase of the probability to have poor disease specific outcome in 5 years. The combination of information regarding these two proteins, together with *p53* status provides better prognosis of metastatic behavior and poor outcome preoperatively. By initializing *Primarytumor* to grade 2 and defining the *ER* and *PR* status to be positive, we notice that patients have low risk of lymph node metastasis which corresponds to 7% and 93% 5-year disease specific survival.

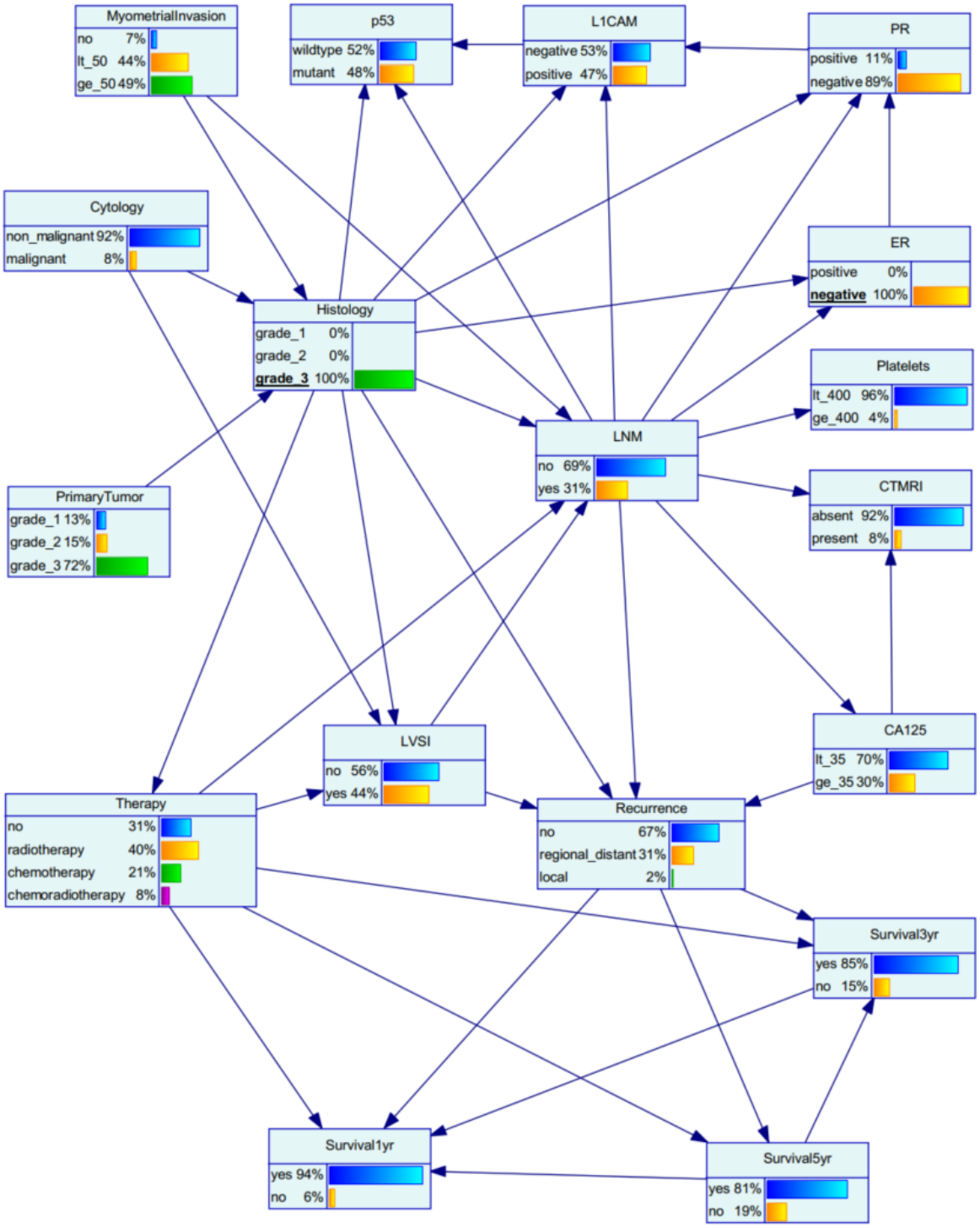


Figure 14: Final Bayesian model with *Histology* and *ER* variables initialized

This fact is also illustrated in [67]. Furthermore, research has classified *p53* receptor as an important predictor of patients survival [15]. In case of loss of this receptor, the BN points out a 21% probability of having lymph node metastasis and 6% higher probability of having poor 5-year survival outcome. Additionally, according to [27] and experts input, *Ca-125* plays an important role in lymph node metastasis. Elevated status of *Ca-125*, corresponds to 36% probability of having enlarged lymph nodes, while a normal level of *Ca-125* decreases this probability by 6% and yields to a 95% positive 5-year disease specific survival.

Evidence	*Expectation*	*Observation*
Histology: grade 3	LNM: 20-25%	LNM: 26%
ER: negative	PR: negative 80% - positive 20%	PR: negative 81% - positive 19%
Primarytumor: grade 3 ER: negative	LNM: 25-30%	LNM: 28%
Histology: grade 3 L1CAM: positive	LNM: 30%	LNM: 32%
LNM: positive	Ca-125: elevated 80% - normal 20%	Ca-125: elevated 79% - normal 21%
Ca-125: elevated	LNM: 25%	LNM: 36%
Ca-125: normal	LNM: 5-10%	LNM: 3%
LVSI: yes	LNM: 20-30%	LNM: 35%
L1CAM: positive	ER: negative 40% - positive 60%	ER: negative 32% - positive 68%
L1CAM: positive	PR: negative 60% - positive 40%	PR: negative 58% - positive 42%
L1CAM: positive	P53: negative 50% - positive 50%	P53: negative 57% - positive 43%
L1CAM: positive	LNM: 20-30%	LNM: 25%
PR: negative	L1CAM: negative 70% - positive 30%	L1CAM: negative 66% - positive 34%

Table 7: Performing inference to the BN - Expert's expectation vs BN observation

5.6 Cox proportional hazards model

First of all, the implementation of the Cox model requires the presence of *survival* and *survminer* packages in R. As a basis, we use the dataset created via bnlearn function in Imputation.R script. Refer to the appendix for the complete survival analysis script. The dataset is extended with two variables, which are necessary for Cox model implementation: *DeathEC* and *FUtime*. The former one illustrates whether or not the patient experienced the event of interest (i.e. death) and the latter one reflects how much time after joining the first visit, the patient had her last follow-up with the clinical doctors. In this extended version of the dataset, there are 18 missing values for DeathEC variable and 11 missing values for FUtime. These observations are omitted from the data and only the complete cases remain in the dataset.

At this point, we create the censor flag *cens* to indicate the observations that are censored. We initialize the flag to be equal to 1 for the complete list of the observations that are present in the data. This value indicates that they are initialized to be censored. On the other hand, based on three specific conditions, the non-censored observation are flagged with cens equal to 0. In detail, the non-censored observations are:

- the patients that have 3-year disease-specific survival, the follow-up time is between 3 and

5 years and the last follow-up do not represent the event of interest (DeathEC is equal to 0),

- the patients that have 1-year disease-specific survival, the follow-up time is between 1 and 3 years and at the moment of the last follow-up they are still alive,

- the patients that have follow-up time less than 12 months and they do not experience the event of interest at the moment of the last follow-up.

Based on these variables, we calculate the survival object by *Surv* function in R. This object can be explained as a matrix, with a single row for each case, the first column representing the number of the observation and the second one the last follow-up time. A plus sign + complements the value of the last follow-up time, specifying the censored observations. At first, we considered including in the Cox proportional hazards model the following two variables: *Therapy* and *Recurrence* of the disease. These two variables participate in the Markov blanket of the survival nodes (e.g. 5-year disease-specific survival) in the Bayesian network. As a consequence, it would make sense to implement such a model and investigate how they jointly impact survival. However there is a limitation posed by the number of data for Recurrence variable. Specifically the 86.2% of patients do not get the disease again, while 7.4% and 0.4% of the population die from regional/distant recurrence and local recurrence of endometrial cancer respectively. Due to this fact, by including Recurrence variable to the model, we observe extreme coefficients and we can not get such a model to work. In a previous section, we analyse why *L1CAM* protein is closely correlated with the disease and considered to be an important protein for the prediction of patient's survival in endometrial cancer. The Cox model in the present thesis incorporates *L1CAM* variable, together with *Therapy* variable. The function *coxph* is used to fit a Cox proportional hazards regression model. The survival object which is previously calculated is used as input for fitting the survival model. After fitting the model and before examining the coefficients, it is vital to assess the validity of the Cox model by measuring whether or not the proportionality assumption holds. In other words, we calculate if the ratio of the hazards for any two observations remains constant over time. Please refer to the theory in 2.4.6, discussing in detail the proportional hazards assumption. In practice, we use function *cox.zph*, present in survival package to measure this. Unfortunately, the p-values for both variables are much smaller than 0.05, which reveals that the proportionality assumption is violated. To address this issue, we decide to divide the time. Based on the histogram in figure 15, we split the time as follows: *(0,60],(60,228]*

At this point, after fitting the model for a subset of data (with follow-up time up to 60 months (5 years)), we test proportional hazards assumption. We notice that the test is not statistically significant for the two covariates and more specifically p-values are equal to 0.37 for Therapy variable and 0.72 for L1CAM variable. Additionally, the global test is not statistically significant , which allows us to accept that the proportional hazards assumption holds. The next step is to evaluate the coefficients for this model. We use *summary()* R function to produce a complete report for the model.

- The column marked z provides the Wald statistic value. It represents the ratio of each regression coefficient to its standard error i.e. $z = coef/se(coef)$. The Wald statistic assesses if the difference of the beta coefficient of a specific variable from 0 is statistically significant or not. From the report, we understand that this is the case for both variables.

- With respect to the *regression coefficients* ($coef$) of the model, a value which is greater than 0 corresponds to a higher hazard for the mentioned group. Because of the higher risk of death for this group, the prognosis is worse in comparison with the group of reference. E.g. The regression coefficient value for the group of patients who are positive for L1CAM

is 0.63. In order to make the example easier to grasp, we consider this to be group 1. The R summary for the model produces the hazard ratio for group 1 relative to the observations that have negative value when measured for L1CAM (group 2). The regression coefficient 0.63 for group 1 demonstrates that patients with over-expression of L1CAM have higher risk of death (higher survival rates) in comparison with group 2 in these data.

- The Hazard ratios (HR), which in practice are calculated as: *exp(coef)* measure the impact for the model covariates. In detail, being part of group 1, increases the hazard by a factor of 0.879 or 87.9%. Having over-expression of L1CAM is correlated to worse prognosis as mentioned earlier.

- *Summary()* function produces the global statistical significance of the Cox model too, which reveals that the model is statistically significant. In detail, it provides three separate alternative test results: The likelihood-ratio test, the Wald test and the score (logrank) test. The mentioned tests are asymptotically similar. For large datasets, they provide equivalent p-values. However for smaller datasets they tend to differ moderately. In case the dataset is relatively small, the likelihood-ratio method displays improved performance, so it is usually a better option.

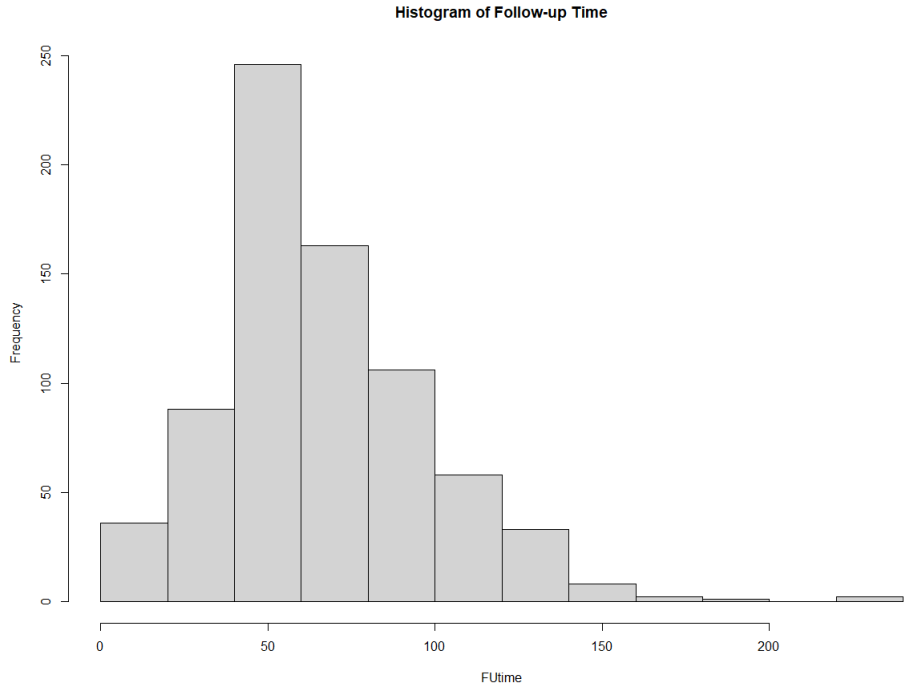


Figure 15: Histogram providing an approximate representation of the distribution of the last follow-up time for the set of patients participated in the study

We can see below an illustration that provides us with an overview of how the mortality of the different groups evolves over time. The figure 16 represents the survival function in the form of a step function. Time is displayed on the x-axis and survival on the y-axis. Each time the curve drops there are patients experiencing the event. E.g. By checking the survival curve in figure 17 for the group of patients with positive L1CAM expression, that is treated by radiotherapy, we see that there is one patient at risk when time is 60 months. This patient is a censored observation for whom the outcome is not known and as a result the representation shows the plot dropping to zero at that point. Let's focus on the survival curve reflecting the group of patients that

are treated with chemotherapy and the are positive for L1CAM expression. We notice that the median survival of the specific subset of the data is less than 40 months, which means that 50% of the specific population survives approximately 37 months. The 95% confidence interval is also defined for each group of patients. This estimate can support effectively the application of the model to larger population of patients. By checking the survival curve of a different group of patients for this model, we can extract information in a similar way. On top of this, we notice vertical lines intersect the survival curve, demonstrating the censored observations. This way, even if we are not aware of the outcome of specific patients, CPHs allow us incorporate them in the analysis. The moment when this vertical line appears, the patient of this specific group of patients drops out of the study. This means that we follow them until the moment that they stop participating in the study and this information is included in the model. The figure 17 reflects the number of patients at risk for the specific group of observations for this subset of the data, at different points in time.

Figure 16: Survival curves for the different groups of patients

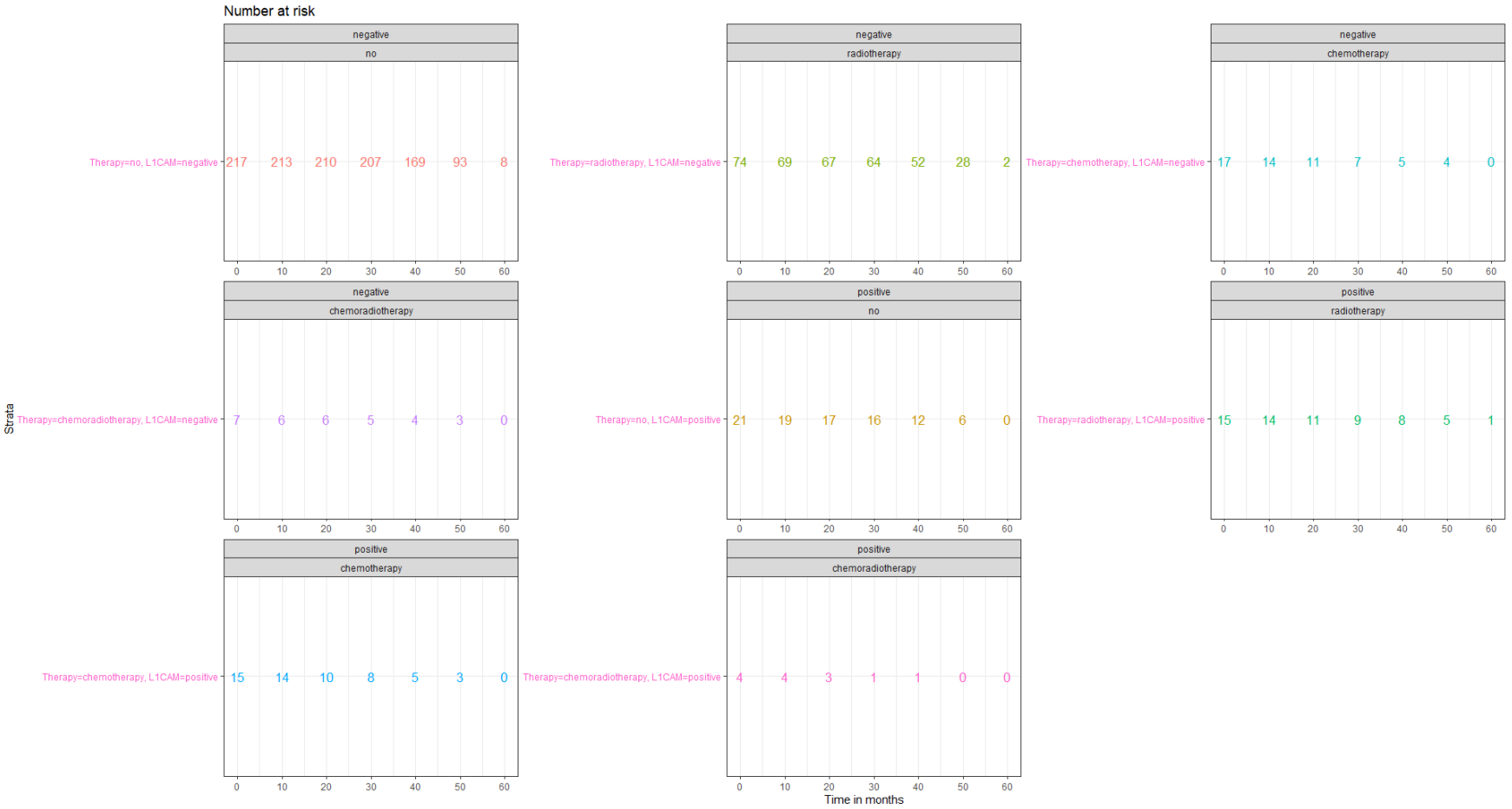


Figure 17: Patients at risk for the different groups of patients

5.7 Quantitative comparison

The *Brier Score* is considered to be a really good performance measure, extensively used to evaluate predictive models in medicine. This approach verifies how accurate a probability prediction is, by determining how close probability forecast is to the actual outcome for each individual. It can be effectively used for binary outcomes, where two possible events exist. This statistical approach is also used for categorical outcomes. However they should be adjusted so as they are in a binary format e.g. true, false. Brier score is always a number between 0 and 1. A perfect model would yield a score equal to 0 while brier score 0.25 corresponds to a much less beneficial model with 50% overall incidence of the outcome. On the contrary, a model with brier score equal to 1, is a model which provides completely inaccurate predictions to its users. The higher the score is the harder the interpretation of such a model becomes.

The formula for brier score calculation is shown below and it is the mean squared error of a forecast:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{25}$$

In this equation N is the number of observation participating in the model for which the calculation takes place. On top of that, p_i corresponds to the probability which is calculated by the prediction model. In this thesis, this refers to the survival probability. o_i is determined by the variable *DeathEC* in the dataset and it is the outcome for each patient which is either 1, in case the individual experiences the event of interest, or 0, in case she does not. Please refer to the appendix for the R script used for this calculation. By focusing on 5-year survival, the computation for the Bayesian model across multiple bootstrap samples generates a brier score equal to *0.034*, better than the Cox model, which yields a value equal to *0.078*. Below the histograms for both scores are displayed.

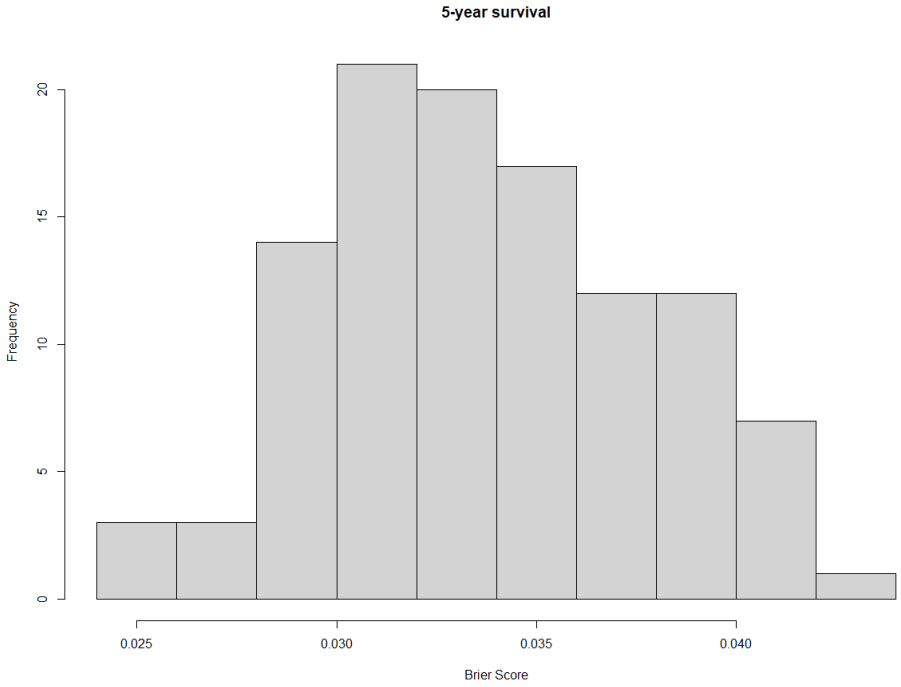


Figure 18: Brier score histogram for ENDORISK model

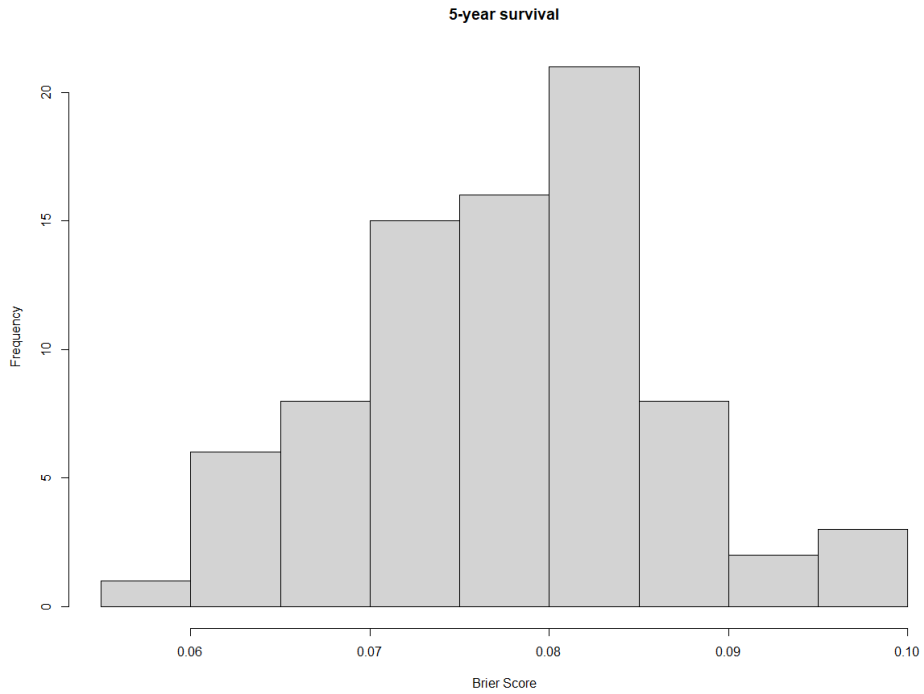


Figure 19: Brier score histogram for Cox model

6 Discussion

6.1 Conclusions

The outcome of interest for this thesis is 5-year disease-specific survival. We list the values of cumulative survival based on both methods in the table 8. However it is really important to mention that we should not look at these values in the same way. These methods differ and their results should not be interpreted similarly. We are diving into their differences in the current section based on the models that are demonstrated earlier.

Type of Therapy	*L1CAM protein level*	*5-year dss Cox model*	*5-year dss BN model*
No	Negative	0.90	0.95
Radiotherapy	Negative	0.91	0.94
Chemotherapy	Negative	0.73	0.63
Chemoradiotherapy	Negative	0.88	0.88
No	Positive	0.87	0.91
Radiotherapy	Positive	0.88	0.90
Chemotherapy	Positive	0.65	0.58
Chemoradiotherapy	Positive	0.84	0.84

Table 8: Survival probabilities of the two models

First, we focus on explaining the CPHs results about 5-year disease specific survival. Typically in a clinical study, CPHs models produce a risk equation to estimate how great is the risk of an event given the hazard function. In our case, there are two chosen predictor variables. To derive the probabilities from Cox model, we specify the combinations of values of these covariates for which the probabilities have to be generated. After defining these in a dataset, we calculate the median survival time for the same. In that case we can evaluate, how these predictors jointly impact survival based on this subset of data. It becomes clear that each outcome of interest has to be evaluated and investigated on a specific model, under a static number of observations and that is a limitation posed by all regression-based models.

Next to these, BNs results for the 5-year dss variable are listed too. As discussed previously, Bayesian networks provide a graphical demonstration of joint probability distributions (JPDs) that reflect the causal interactions of the participated variables and allow risk modelling. These networks represent in a way personal belief in combination with new evidence that can be incorporated later. Additionally, more than one outcomes can be examined under the same model. In the case of the present thesis, the clinical doctors were also interested in lymph node metastasis as an event of interest. This is really straightforward to achieve in BNs and an advantage over CPHs models. More information can be found in [51] about these results. In general, the same clinical model can provide answers to multiple questions being this way a really powerful tool for clinicians.

The figure 13 highlights the conditional dependencies among the variables closely related to endometrial cancer. Given no prior evidence, an observation's chance of 5-year disease specific survival is 93%. Once clinical doctors treat the patient with Chemotherapy after determining a positive expression of L1CAM protein, the probability drops to 58%. In case the only known evidence is L1CAM and its positive expression, we can use the model to actually investigate what is the most suitable treatment. The more evidence the clinicians collect the more individualized analysis is conducted. We can also explore different scenarios e.g. how the 5-year dss is affected if the patient experiences a specific type of recurrence or increased levels of other biomarkers such as Ca-125. This way, the prior probabilities in the network are updated and the different nodes therefore reflect the updated distributions (posterior probabilities) according to Bayes's

theorem. A BN has the ability not only to remain relevant and applicable even if the population is adjusted but also it can get more and more refined over time as new information is available.

Earlier we have also analyzed the coefficient for overexpression of L1CAM protein in Cox model. This demonstrates a worse prognosis for this group of patients. When similar evidence is inserted into the BN, we notice that 5-year disease-specific survival drops rapidly. In general it is possible to perform individualised risk reasoning by using CPHs. However this is not that simple and it requires the application of cumulative hazard and then combine coefficients of multiple model predictor variables. The difficulty level increases, specifically when there is a continuous variable involved in the model with a reference state which matches a mean value. Interpretation becomes a challenging task then [49].

A Bayesian model is used to reason from cause to effect but also vice-versa, which is not possible in CPHs models. E.g. in case the event of interest is lymph node metastasis, by initializing the BN with 'yes' level of this variable, the doctor can evaluate the updated distributions of the rest of the nodes and reason about how patient's hormones are affected. Of course it is also possible to insert any evidence available based on medical test results and estimate this way the probability of lymph node metastasis by using the exact same model. To sum up information can propagate from the initialised nodes towards all directions, through the links established clearly in the DAG during model implementation. It is worth mentioning that when a model's focus is predicting risk, it is not necessary to have causal relationships among the present variables.

On the other hand, CPHs model supports scientists in observing how survival probabilities evolve during the period of the study for specific groups of patients. E.g. when visualising the survival curve of the Cox model, we can literally check what is the survival probability of a specific group at a specific moment in time. In order to have something similar in a Bayesian network, time has to be introduced in the form of a variable and parameter learning should be therefore performed. After taking this step, we are able to evaluate survival after 1, 3 and 5 years in the present thesis.

An extra challenge of Cox models is the proportionality assumption. Especially if the hazard of almost all variables that we want to incorporate into the model changes over time. In order to have a valid model, this assumption has to hold. While working with Cox models, we dealt with this challenge. In case the assumption is violated, different workarounds have to be explored to overcome this barrier. Going into depth about it is not in scope for the current thesis but there is great research and examples available. Bayesian networks do not pose such limitations to the users. On the contrary, they create a transparent visualization of the dependencies, yielding a more compact, flexible and intuitive model. Last but not least, a possible difficulty when using BNs is the need to to have a tool (such as SAMIAM or GeNIe) after implementing it, in order to be able to to manipulate and configure it in a straightforward way.

6.2 Future steps

As mentioned earlier, the medical doctors from Radboud University Medical Center have already received more funding to continue with their research based on the ENDORISK model. The plan is to implement a study in which oncologists start actively using the model in their medical practise. More analysis is performed in order to extend the available data and include potentially useful variables and then validation in a large cohort of patients in Norway (10,000 patients) will take place. A possible future step would be the use of different structure learning algorithms during the structure learning process with more emphasis on modern techniques for causal discovery. Especially after introducing new variables in the data, it might be beneficial to explore whether or not different algorithms (e.g. constraint based algorithms) can yield better results and support the clinical doctors in discovering meaningful connections. After diving into the fundamentals of Bayesian models, it might be also really interesting to investigate if

Dynamic Bayesian networks (DNB) could be also used for such purposes and practically support medical experts that focus on different types of cancer.

Appendices

A Data Processing

```
# ----------------------
# Working on the dataset
# ----------------------

# VARIABLE NAMES AND POSSIBLE VALUES:

# INPUT VARIABLES
# ---------------
# Cytology: preoperative cervical cytology
#           "non_malignant": no malignant endometrial cells present
#           "malignant": malginant endometrial cells present

# Histology: postoperative tumor grade
#            "grade_1"
#            "grade_2"
#            "grade_3"

# MI:  Myometrial invasion
#      "no-invasion"": no invasion of the myometrial by tumor cells
#      "less-50": 0% < invasive tumor cells < 50% of wall
#      "equalorgreater-50": invasive tumor cells >= 50% of wall

# PrimaryTumor: preoperative tumor grade
#                "grade_1"
#                "grade_2"
#                "grade_3"

# Therapy
#                "no"
#                "radiotherapy"
#                "chemotherapy"
#                "chemo_and_radiotherapy"

# LVSI: Does the carcinoma grow into the bloodvessels or lymph vessels
#       "no"
#       "yes"

# CA125: CA-125 (Cancer Antigen 125) serum levels
#        "less-35"
#        "equalorgreater-35"

# CTMRI: CT or MRI imaging.
#        The presence of lymphadenopathy or distant metastasis
#        "no" : absent
#        "yes":present

# ER: Estrogen receptor levels
#     "negative"
#     "positive"

# PR: Progesteron receptor levels
#     "negative"
#     "positive"

# L1CAM: L1CAM is an intracellular protein which promotes cell motility,
# and thus might cause the carcinoma to spread faster.
#        "negative"
```

```
#        "positive"

# p53 = p53 is a tumor suppressor gene
#       "wildtype"
#       "mutant"

# Pl: number of platelets in blood
#           "no"
#           "yes"

# Rec: Recurrence of the disease
#          "no"
#          "regional_distant"
#          "local"

# OUTPUT VARIABLES
# ----------------

# LNM: lymphnode metastases
#      "no"
#      "yes"

# X1YR: disease specific disease survival of at least 1 year
#               "no"
#               "yes"

# X3YR: disease specific disease survival of at least 3 years
#               "no"
#               "yes"

# X5YR: disease specific disease survival of at least 5 years
#               "no"
#               "yes"

# Load Data
myData <- read.csv2(file="../Dataset/L1CAM_2_database_11122018_clean.csv",
                    header=TRUE, sep = ";", #colClasses = c(rep("factor")),
                    na.strings=c("","  ","NA"))

attach(myData)
# -----------------------------------------
# Rename selected variables of initial dataset
# -----------------------------------------
names(myData)[names(myData) == 'Grade_PREOP'] <- 'Primarytumor'
names(myData)[names(myData) == 'Grade'] <- 'Histology'
names(myData)[names(myData) == 'LVSI_bi1'] <- 'LVSIb'
names(myData)[names(myData) == 'Positive_nodes_bi1'] <- 'LNM'
names(myData)[names(myData) == 'CA125_PREOP_bi'] <- 'CA125'
names(myData)[names(myData) == 'CT_or_MRI_LNM'] <- 'CTMRI'
names(myData)[names(myData) == 'Platelets_bi'] <- 'Pl'
names(myData)[names(myData) == 'ER_expression_preop'] <- 'ER'
names(myData)[names(myData) == 'PR_expression_preop'] <- 'PR'
names(myData)[names(myData) == 'L1CAM_expression_preop'] <- 'L1CAM'
names(myData)[names(myData) == 'p53_expression_preop'] <- 'p53'
names(myData)[names(myData) == 'Recurrence_location'] <- 'Rec'
names(myData)[names(myData) == 'Adjuvanttherapy'] <- 'Therapy'
names(myData)[names(myData) == 'Death_by_EC'] <- 'DeathEC'
names(myData)[names(myData) == 'Duration_followup'] <- 'FUtime'
names(myData)[names(myData) == 'one_year_survival'] <- 'X1YR'
names(myData)[names(myData) == 'three_year_survival'] <- 'X3YR'
names(myData)[names(myData) == 'five_year_survival'] <- 'X5YR'

```

```
120
121 # ---------------------------
122 # Transformation of variables
123 # ---------------------------
124 attach(myData)
125 # Recode values with NA
126 for (i in 1:length(myData[,1])) {
127   if ((!is.na(myData$Cytology[i]) &&
128        myData$Cytology[i] == 99)) {
129     myData$Cytology[i] = NA
130   }
131 }
132
133 for (i in 1:length(myData[,1])) {
134   if ((!is.na(myData$CA125[i]) &&
135        myData$CA125[i] == 99)) {
136     myData$CA125[i] = NA
137   }
138 }
139
140 for (i in 1:length(myData[,1])) {
141   if ((!is.na(myData$LVSIb[i]) &&
142        myData$LVSIb[i] == 99)) {
143     myData$LVSIb[i] = NA
144   }
145 }
146
147 for (i in 1:length(myData[,1])) {
148   if ((!is.na(myData$MI[i]) &&
149        myData$MI[i] == 99)) {
150     myData$MI[i] = NA
151   }
152 }
153
154 for (i in 1:length(myData[,1])) {
155   if ((!is.na(myData$DeathEC[i]) &&
156        myData$DeathEC[i] == 99)) {
157     myData$DeathEC[i] = NA
158   }
159 }
160
161 for (i in 1:length(myData[,1])) {
162   if ((!is.na(myData$Primarytumor[i]) &&
163        myData$Primarytumor[i] == 99)) {
164     myData$Primarytumor[i] = NA
165   }
166 }
167
168 for (i in 1:length(myData[,1])){
169   if(myData$Cytology[i] == 3 && !is.na(myData$Cytology[i])){
170     myData$Cytology[i] = NA
171   }
172 }
173
174 for (i in 1:length(myData[,1])){
175   if(myData$Cytology[i] == 0 && !is.na(myData$Cytology[i])){
176     myData$Cytology[i] = NA
177   }
178 }
179
180 # -------------------------------------------
181 # Create the necessary subsets of the dataset
```

```r
# ------------------------------------------
attach(myData)

subsetDAG.o <- data.frame(MI, CTMRI, LNM, LVSIb,
                          CA125, Primarytumor, Histology, Pl,
                          Cytology, ER, PR, L1CAM, p53,
                          Rec, Therapy,
                          X1YR, X3YR, X5YR)
summary(subsetDAG.o)
# ------------------------------------------
# Transform the variables
# Categorical to factors, Continuous to numeric
# ------------------------------------------
library(dplyr)
attach(subsetDAG.o)
subsetDAG.o <- subsetDAG.o %>%
  mutate(
    MI = as.factor(MI),
    CTMRI = as.factor(CTMRI),
    LNM = as.factor(LNM),
    LVSIb = as.factor(LVSIb),
    Primarytumor = as.factor(Primarytumor),
    Histology = as.factor(Histology),
    Pl = as.factor(Pl),
    Cytology = as.factor(Cytology),
    ER = as.factor(ER),
    PR = as.factor(PR),
    L1CAM = as.factor(L1CAM),
    p53 = as.factor(p53),
    Rec = as.factor(Rec),
    Therapy = as.factor(Therapy),
    CA125 = as.factor(CA125),
    X1YR = as.factor(X1YR),
    X3YR = as.factor(X3YR),
    X5YR = as.factor(X5YR)
)

# ------------------------------------------------------------
# Change the values of the variables of subsetDAG.o to more
# meaningful ones, attach value labels to factors levels
# ------------------------------------------------------------
attach(subsetDAG.o)
summary(subsetDAG.o)
subsetDAG.o$ER <- factor(subsetDAG.o$ER,
                         levels = c(0,1),
                         labels = c("positive", "negative"))
summary(subsetDAG.o$ER)

subsetDAG.o$PR <- factor(subsetDAG.o$PR,
                         levels = c(0,1),
                         labels = c("positive", "negative"))
summary(subsetDAG.o$PR)

subsetDAG.o$L1CAM <- factor(subsetDAG.o$L1CAM,
                            levels = c(0,1),
                            labels = c("negative", "positive"))
summary(subsetDAG.o$L1CAM)

subsetDAG.o$p53 <- factor(subsetDAG.o$p53,
                          levels = c(0,1),
                          labels = c("wildtype", "mutant"))
summary(subsetDAG.o$p53)
```

```
244
245 subsetDAG.o$MI <- factor(subsetDAG.o$MI,
246                          levels = c(0, 1, 2),
247                          labels = c("no-invasion", "less-50", "equalorgreater-50
    "))
248 summary(subsetDAG.o$MI)
249
250 subsetDAG.o$Primarytumor <- factor(subsetDAG.o$Primarytumor,
251                                    levels = c(1,2,3),
252                                    labels = c("grade1", "grade2", "grade3"))
253 summary(subsetDAG.o$Primarytumor)
254
255 subsetDAG.o$Histology <- factor(subsetDAG.o$Histology,
256                                 levels = c(1,2,3),
257                                 labels = c("grade1", "grade2", "grade3"))
258 summary(subsetDAG.o$Histology)
259
260 subsetDAG.o$LVSIb <- factor(subsetDAG.o$LVSIb,
261                             levels = c(0,1),
262                             labels = c("no", "yes"))
263 summary(subsetDAG.o$LVSIb)
264
265 subsetDAG.o$LNM <- factor(subsetDAG.o$LNM,
266                           levels = c(0,1),
267                           labels = c("negative", "positive"))
268 summary(subsetDAG.o$LNM)
269
270 subsetDAG.o$CA125 <- factor(subsetDAG.o$CA125,
271                             levels = c(0,1),
272                             labels = c("less-35", "equalorgreater-35"))
273 summary(subsetDAG.o$CA125)
274
275 subsetDAG.o$CTMRI <- factor(subsetDAG.o$CTMRI,
276                             levels = c(0,1),
277                             labels = c("no", "yes"))
278 summary(subsetDAG.o$CTMRI)
279
280 subsetDAG.o$Pl <- factor(subsetDAG.o$Pl,
281                          levels = c(0,1),
282                          labels = c("no", "yes"))
283 summary(subsetDAG.o$Pl)
284
285 subsetDAG.o$Cytology <- factor(subsetDAG.o$Cytology,
286                                levels = c(1,2),
287                                labels = c("non-malignant", "malignant"))
288 summary(subsetDAG.o$Cytology)
289
290 subsetDAG.o$X1YR <- factor(subsetDAG.o$X1YR,
291                            levels = c(0,1),
292                            labels = c("yes", "no"))
293 summary(subsetDAG.o$X1YR)
294
295 subsetDAG.o$X3YR <- factor(subsetDAG.o$X3YR,
296                            levels = c(0,1),
297                            labels = c("yes", "no"))
298 summary(subsetDAG.o$X3YR)
299
300 subsetDAG.o$X5YR <- factor(subsetDAG.o$X5YR,
301                            levels = c(0,1),
302                            labels = c("yes", "no"))
303 summary(subsetDAG.o$X5YR)
304
```

```r
subsetDAG.o$Rec <- factor(subsetDAG.o$Rec,
                          levels = c(0,1,2,3),
                          labels = c("no", "yes_distant", "yes_local", "yes_
    regional"))
summary(subsetDAG.o$Rec)

subsetDAG.o$Therapy <- factor(subsetDAG.o$Therapy,
                              levels = c(0,1,2,3),
                              labels = c("no", "radiotherapy", "chemotherapy", "
    chemoradiotherapy"))
summary(subsetDAG.o$Therapy)
```

B Modeling Process - Dealing with Incomplete & Missing Data

```r
# ----------------------------------------  #
# Build a Bayesian network from a modelstring
# ----------------------------------------  #
# Multiple models were explored during the implementation process;
# Only the first and the final ones are shown below.
# Properly adjusted subsets of the main dataset, are used later on,
# in order to perform parameter learning on the different
# Directed Acyclic Graphs (DAGs)
#-------------------------------------------------------------
# Prior to structure learning and after improving the network
#-------------------------------------------------------------
modelstringDAG.o <- paste("[MI][Cytology]",
                          "[Pl|LNM][Primarytumor]",
                          "[Histology|Primarytumor:MI:Cytology]",
                          "[LNM|LVSIb:MI]",
                          "[CA125|LNM][CTMRI|LNM:CA125]",
                          "[LVSIb|Histology:Cytology:MI]",
                          "[p53|LNM:Histology]",
                          "[ER|LNM:Histology][PR|LNM:Histology]",
                          "[L1CAM|LNM:Histology]",
                          "[Rec|CA125:LNM:LVSIb:Histology]",sep="")
```

```r
#-----------------------------------------------
# Final model, after incorporating structure
# learning algorithms results and survival nodes
#-----------------------------------------------
modelstringDAG.o <- paste("[MI][Cytology]",
                          "[Pl|LNM][Primarytumor]",
                          "[Histology|Primarytumor:MI:Cytology]",
                          "[LNM|LVSIb:MI:Histology:Therapy]",
                          "[CA125|LNM][CTMRI|LNM:CA125]",
                          "[LVSIb|Histology:Cytology:Therapy]",
                          "[p53|LNM:Histology:L1CAM]",
                          "[ER|LNM:Histology][PR|LNM:Histology:ER]",
                          "[L1CAM|LNM:Histology:PR]",
                          "[Rec|CA125:LNM:LVSIb:Histology]",
                          "[Therapy|Histology]", "[X1YR|Rec:Therapy:X3YR:X5YR]",
                          "[X3YR|Rec:Therapy:X5YR]",
                          "[X5YR|Rec:Therapy]",sep="")
```

```r
# ------------------------------------
# BN by Hand - Imputation with bnlearn
# ------------------------------------
library(bnlearn)

# ------------------------------------------------------------
# The necessary subsets are created in the cleaning.R
```

```r
# Load a bayesian network (modelstring) from multiplemodels.R
# ------------------------------------------------------------
# Transform the modelstring to Directed Acyclic Graph
DAG.o <- model2network(modelstringDAG.o)

# Plot the Graphs
graphviz.plot(DAG.o, shape = "ellipse")

# ------------------------------------------------
# Create the bn.fit object (i.e. with parameters)
# Both the number and names of nodes should be the
# same in "Dag" and the dataset
# ------------------------------------------------
# fit the parameters of the local distributions given
# its structure and a data set -> in a form of
# conditional probability tables.
# ------------------------------------------------
attach(subsetDAG.o)
fittedDAG.o <- bn.fit(DAG.o, data = subsetDAG.o, method = "bayes")

# bn.net returns the structure underlying a fitted Bayesian network.
# fittedDAG.o.net <- bn.net(fittedDAG.o)

# --------------------
# Imputation
# --------------------
imputed.o <- impute(fittedDAG.o, subsetDAG.o, method = "bayes-lw")
class(imputed.o)

# Double-Check whether NA or not
sapply(imputed.o, function(x) sum(is.na(x)))

summary(imputed.o)

# Save the imputed dataset
write.csv(imputed.o, file = "imputed.o.csv")

# ------------------------------------------------------------------------
# Create the bn.fit object (i.e. with parameters) with imputed.o datasets
# ------------------------------------------------------------------------
fittedDAGi.o <- bn.fit(DAG.o, data = imputed.o)
fittedDAGi.o.net <- bn.net(fittedDAGi.o)

# Create the .net file for SAMIAM use for imputed datasets
write.net("BN.net", fittedDAGi.o)

# ----------------------------
# Check if models fit the data
# ----------------------------
# complete cases based on initial dataset: subsetDAG.O for score of unimputed BN
sum(complete.cases(subsetDAG.o))
cmpcases.o <- subsetDAG.o[complete.cases(subsetDAG.o), ]

# Compute logLikelihoods to compare the BN, how well the model fits the data
logLik(fittedDAG.o, cmpcases.o)  #-987.2534

# Check what happens to the likelihood
# How well the model fits complete cases data
# (using fitted object, created with imputed data)
logLik(fittedDAGi.o, cmpcases.o)
# (using fitted object, created with imputed data)
logLik(fittedDAGi.o, imputed.o)
```

C Structure learning process

```
# ------------------------------
# Score-based Structure Learning
# ------------------------------
# Compute arc strength for hc and tabu, a way of avoid getting
# stuck in local maxima: restart for the number of random
# restarts & perturb for the number of perturbed
# arcs in the new starting DAG

library(bnlearn)

#------------------------
# Hill-climbing algorithm
#------------------------
hc.strength <- boot.strength(imputed.o, R = 500, algorithm = "hc",
                             algorithm.args = list(score = "bde",
                             iss = 10, restart = 5, perturb = 10))

# We choose the threshold for an arc to be considered strong enough and
# added to the the averaged network
head(hc.strength[(hc.strength$strength > 0.85)
                 & (hc.strength$direction >= 0.5), ], n = 3)
hc.strongest <- hc.strength[(hc.strength$strength > 0.85)
                 & (hc.strength$direction >= 0.5), ]

# With d <- 0, all arcs are present, and their strength is obvious by the
# width of the arcs. Adjust the threshold d:
d <- 0.5

hc.avg <- averaged.network(hc.strength, threshold = d)

hc.strplot <- strength.plot(hc.avg, hc.strength,
              sub = paste("Algorithm = hc search; Threshold =",as.character(d)),
              shape = "ellipse") #highlight = list(arcs = arcs())

dev.copy2pdf(file = "endomcancer-hc-d05outcome.pdf")

#---------------------
#Tabu search algorithm
#---------------------
tabu.strength <- boot.strength(imputed.o, R = 500, algorithm = "tabu",
                 algorithm.args = list(score = "bde", iss = 10))  #debug = TRUE


# We choose the threshold for an arc to be considered strong enough and
# added to the the averaged network
head(tabu.strength[(tabu.strength$strength > 0.85)
                   & (tabu.strength$direction >= 0.5), ], n = 3)
# Arcs are considered significant if they appear in at least 85% of the networks
# and in the most frequent direction >0.5

tabu.strongest <- tabu.strength[(tabu.strength$strength > 0.85)
                   & (tabu.strength$direction >= 0.5), ]

# With d <- 0, all arcs are present, and their strength is obvious by the width
    of the arcs
d <- 0.5

tabu.avg <- averaged.network(tabu.strength, threshold = d)

```

```r
tabu.strplot <- strength.plot(tabu.avg, tabu.strength,
                sub = paste("Algorithm = Tabu search; Threshold =",
                            as.character(d)),
                shape = "ellipse") #highlight = list(arcs = arcs())

dev.copy2pdf(file = "endomcancer-tabu-d07outcome.pdf")
```

D Survival analysis

```r
# ------------------
# Survival Analysis
# ------------------
sessioninfo::session_info()

# Load necessary libraries
library(survival)
#library(survreg)
#library(ggplot2)
#library(ggfortify)
library(survminer)
#library(pec)


# Use of imputed.o dataset (created via bnlearn function in Imputation.R script)
# Extension of dataset with two important variables for survival analysis:
# Deathec and FUtime
# The extended version: subsetDAG.ext


attach(myData)
subsetDAG.ext <- data.frame(imputed.o, DeathEC, FUtime)
subsetDAG.ext$FUtime <- as.numeric(FUtime)
summary(subsetDAG.ext$FUtime)
subsetDAG.ext$DeathEC <- as.factor(DeathEC)
summary(subsetDAG.ext$DeathEC)

# Double-Check whether NA or not
sapply(subsetDAG.ext, function(x) sum(is.na(x)))
summary(subsetDAG.ext)

subsetDAG.ext.cmp <- na.omit(subsetDAG.ext)
subsetDAG.ext.i <- subsetDAG.ext.cmp


# Create the censor flag variable for patients for whom the outcome is
# not known according to the three ourcome variables X1TY, X3YR, X5YR

subsetDAG.ext.i$cens <- 1

subsetDAG.ext.i[(subsetDAG.ext.i$X3YR == "yes" & subsetDAG.ext.i$FUtime < 60
                  & subsetDAG.ext.i$FUtime >= 36
                  & subsetDAG.ext.i$DeathEC == 0) |
                 (subsetDAG.ext.i$X1YR == "yes" & subsetDAG.ext.i$FUtime < 36
                  & subsetDAG.ext.i$FUtime >= 12
                  & subsetDAG.ext.i$DeathEC == 0) |
                 (subsetDAG.ext.i$FUtime < 12 & subsetDAG.ext.i$DeathEC == 0),
               "cens"] <- 0

subsetDAG.ext.i$cens <- as.factor(subsetDAG.ext.i$cens)


```

```r

# Fit a cox model containing two variables Therapy and LICAM

cox.fit<- coxph(Surv(time = FUtime, event = as.numeric(cens))
                              ~ Therapy +L1CAM, data = subsetDAG.ext.i, x= TRUE)

summary(cox.fit)
cox.zph(cox.fit)

# Violation of proportionality assumption, p values for both Therapy
# and L1CAM are <0.05. To address this issue, we divide the survival
# analysis into two time periods

# Follow up time histogram
attach(subsetDAG.ext.i)
hist(FUtime, main = paste("Histogram of Follow-up Time"))

# Based on the histogram it is reasonable to divide time to the following
# periods (0, 60], (60, 228).

# Time period: (0, 60]
subsetDAG.FU <- subsetDAG.ext.i[which(subsetDAG.ext.i$FUtime<=60) ,]
attach(subsetDAG.FU)

cox.fit<- coxph(Surv(time = FUtime, event = as.numeric(cens))
                ~ Therapy + L1CAM, data = subsetDAG.FU, x= TRUE)
cox.fit
summary(cox.fit)

# Check proportionality assumption
cox.zph(cox.fit) #it holds

# Simple tabulatios to understand how many observations experienced
# the event in the subset of the data based on the two variables
table(subsetDAG.FU$Therapy, subsetDAG.FU$DeathEC)
table(subsetDAG.FU$L1CAM, subsetDAG.FU$DeathEC)

# The variables Therapy, L1CAM are not significant
# the model is not significant p-value=0.8
# We focus only on the model fitted in the first
# subset of the data to measure the 5 year disease specific survival

# Data used: subsetDAG.FU for Followup time: (0, 60]
require(survival)

fit <- survfit(Surv(time = FUtime, event = as.numeric(cens))
               ~ Therapy+L1CAM, data = subsetDAG.FU)

ggsurvival <- ggsurvplot(fit, conf.int = TRUE,
                     risk.table = TRUE,
                     risk.table.col="strata",
                     xlab = "Time in months",
                     ggtheme = theme_bw(),
                     data = subsetDAG.FU)
ggsurvival
curv_facet <- ggsurvival$plot + facet_wrap(L1CAM ~ Therapy)
curv_facet

# Facet risk tables
ggsurvival$table + facet_wrap(~ L1CAM + Therapy, scales = "free")+
  theme(legend.position = "none")

```

```r
# Focus on each facet columns and produce the relevant risk table
tb_fac <- ggsurvival$table + facet_grid(.~ Therapy, scales = "free")
tb_fac + theme(legend.position = "none")

tb_fac <- ggsurvival$table + facet_grid(.~ L1CAM, scales = "free")
tb_fac + theme(legend.position = "none")
```

E Models comparison - Brier Score calculation

```r
library(bnlearn)
library(pec)
library(survival)

#----------------------------
# 5 year survival as outcome
#----------------------------

# Calculation for Bayesian network

nbs <- 110  #number of bootstrap samples
ss <- 650   #sample size
bs=matrix(nrow=nbs,ncol=1)

for (i in 1:nbs){
  set.seed(i)
  ss.rows <- sample(1:nrow(subsetDAG.o),ss,replace=TRUE)
  ss.subset <- subsetDAG.o[ss.rows,]
  ss.isubset <- bnlearn::impute(fittedDAG.o, ss.subset, method = "bayes-lw")
  fittedDAGi.o <- bn.fit(DAG.o, data = ss.isubset)
  pred <- predict(fittedDAGi.o, node="X5YR" ,ss.isubset, prob=TRUE)
  res.prob <- data.frame(t(attributes(pred)$prob))
  ss.isubset$X5YR.binary<-ifelse(ss.isubset$X5YR=="yes",1,0)
  df <- as.data.frame(cbind(pred, res.prob,X5YR=ss.isubset$X5YR,X5YR.binary=ss.
    isubset$X5YR.binary))

  #Brier score computation
  bs[i,1] <- sum((df$yes-df$X5YR.binary)^2)/ss

}

#Brier score across bootstrap samples
bs
mean(bs)  #0.0336
hist(bs,xlab = "Brier Score", main="5-year survival")


# Calculation for Cox model

nbs <- 80  #number of bootstrap samples
ss <- 650   #sample size
bs=matrix(nrow=nbs,ncol=1)

for (i in 1:nbs){
  set.seed(i)
  subsetDAG.FU <- subsetDAG.ext.i[which(subsetDAG.ext.i$FUtime<=60) ,]
  ss.rows <- sample(1:nrow(subsetDAG.FU),ss,replace=TRUE)
  ss.subset.train <- subsetDAG.FU[ss.rows,]
  ss.subset.test <- subsetDAG.FU[-ss.rows,]
  cox.fit<- coxph(Surv(time = FUtime, event = as.numeric(cens)) ~
                    Therapy + Rec, data = ss.subset.train, x=TRUE)
  pred <- predictSurvProb(cox.fit,times=60,newdata=ss.subset.test)
```

```
  ss.subset.test$X5YR.binary=ifelse(ss.subset.test$X5YR=="yes",1,0)
  df <- as.data.frame(cbind(pred, X5YR.binary=ss.subset.test$X5YR.binary))

  #Brier score computation
  bs[i,1] <- sum((df$V1-df$X5YR.binary)^2)/ss

}


#Brier score across bootstrap samples
bs
mean(bs) #0.078
sd(bs)
hist(bs,xlab = "Brier Score", main="5-year survival")
```

References

[1] Vera M Abeler and Kjell E Kjørstad. Endometrial adenocarcinoma in norway. a study of a total population. *Cancer*, 67(12):3093–3103, 1991.

[2] Ameen Abu-Hanna and Peter JF Lucas. Prognostic models in medicine. *Methods of information in medicine*, 40(01):1–5, 2001.

[3] DG Altman. Analysis of survival times. *Practical statistics for medical research*, 1, 1991.

[4] Frederic Amant, Philippe Moerman, Patrick Neven, Dirk Timmerman, Erik Van Limbergen, and Ignace Vergote. Endometrial cancer. *The Lancet*, 366(9484):491–505, 2005.

[5] Tina A Ayeni, Jamie N Bakkum-Gamez, Andrea Mariani, Michaela E McGree, Amy L Weaver, Michael G Haddock, Gary L Keeney, Harry J Long, Sean C Dowdy, and Karl C Podratz. Comparative outcomes assessment of uterine grade 3 endometrioid, serous, and clear cell carcinomas. *Gynecologic oncology*, 129(3):478–485, 2013.

[6] Michael L Berman, Samuel C Ballon, Leo D Lagasse, and Watson G Watring. Prognosis and treatment of endometrial cancer. *American journal of obstetrics and gynecology*, 136(5):679–688, 1980.

[7] Cancer.Net Editorial Board. Uterine cancer: Statistics. `https://www.cancer.net/cancer-types/uterine-cancer/statistics`. Accessed: June 2017.

[8] Dorry Boll, HE Karim-Kos, RHA Verhoeven, CW Burger, JW Coebergh, LV van de Poll-Franse, and HC Van Doorn. Increased incidence and improved survival in endometrioid endometrial cancer diagnosed since 1989 in the netherlands: a population based study. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 166(2):209–214, 2013.

[9] Moller H Bray F1, Dos Santos Silva I and Weiderpass E. Endometrial cancer incidence trends in europe: underlying determinants and prospects for prevention. *Cancer Epidemiology, Biomarkers and Prevention*, 14(5):1132–42, 2005.

[10] Eugenia E Calle, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of us adults. *New England Journal of Medicine*, 348(17):1625–1638, 2003.

[11] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[12] David Roxbee Cox. *Analysis of survival data*. Routledge, 2018.

[13] William T Creasman, C Paul Morrow, Brian N Bundy, Howard D Homesley, James E Graham, and Paul B Heller. Surgical pathologic spread patterns of endometrial cancer: a gynecologic oncology group study. *Cancer*, 60(S8):2035–2041, 1987.

[14] Stefan Edelkamp and Stefan Schroedl. *Heuristic search: theory and applications*. Elsevier, 2011.

[15] Ingeborg B Engelsen, Lars A Akslen, and Helga B Salvesen. Biologic markers in endometrial cancer treatment. *Apmis*, 117(10):693–707, 2009.

[16] Nir Friedman, Moises Goldszmidt, et al. Discretizing continuous attributes while learning bayesian networks. In *ICML*, pages 157–165, 1996.

[17] Mitchell Gail, Klaus Krickeberg, J Samet, Anastasios Tsiatis, and Wing Wong. Statistics for biology and health, 2007.

[18] Clark Glymour, Richard Scheines, and Peter Spirtes. *Causation, prediction, and search*. MIT Press, 2001.

[19] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[20] L Helpman, R Kupets, A Covens, RS Saad, MA Khalifa, N Ismiil, Z Ghorab, V Dubé, and S Nofech-Mozes. Assessment of endometrial sampling as a predictor of final surgical pathology in endometrial cancer. *British journal of cancer*, 110(3):609, 2014.

[21] Holly A Hill, J William Eley, Linda C Harlan, Raymond S Greenberg, Rolland J Barrett II, and Vivien W Chen. Racial differences in endometrial cancer survival: the black/white cancer survival study. *Obstetrics & Gynecology*, 88(6):919–926, 1996.

[22] Monica Huszar, Marco Pfeifer, Uwe Schirmer, Helena Kiefel, Gottfried E Konecny, Alon Ben-Arie, Lutz Edler, Maria Münch, Elisabeth Müller-Holzner, Susanne Jerabek-Klestil, et al. Up-regulation of l1cam is linked to loss of hormone receptors and e-cadherin in aggressive subtypes of endometrial carcinomas. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 220(5):551–561, 2010.

[23] Mylonas I1. Prognostic significance and clinical importance of estrogen receptor alpha and beta in human endometrioid adenocarcinomas. *Oncol Rep.*, 24(2):385–93, 2010 Aug.

[24] Nationa Cancer Institute. Endometrial cancer incidence rising in the us and worldwide. `https://www.cancer.gov/news-events/cancer-currents-blog/2017/endometrial-cancer-incidence-rising`. Accessed: 2017-11-20.

[25] Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90, 2011.

[26] Bokhman JV. Two pathogenetic types of endometrial carcinoma. *Gynecol Oncol.*, pages 96–104, 1983.

[27] Sokbom Kang, Woo Dae Kang, Hyun Hoon Chung, Dae Hoon Jeong, Sang-Soo Seo, Jong-Min Lee, Jae-Kwan Lee, Jae Weon Kim, Seok-Mo Kim, Sang-Yoon Park, et al. Preoperative identification of a low-risk group for lymph node metastasis in endometrial cancer: a korean gynecologic oncology group study. *Journal of Clinical Oncology*, 30(12):1329–1334, 2012.

[28] Lois Kim. Survival analysis for epidemiologic and medical research. steve selvin., 2008.

[29] Jidapa Kraisangka, Marek J Druzdzel, Lisa C Lohmueller, Manreet K Kanwar, James F Antaki, and Raymond L Benza. Bayesian network vs. cox's proportional hazard model of pah risk: A comparison. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 139–149. Springer, 2019.

[30] Lukasz Kurgan and Krzysztof J Cios. Discretization algorithm that uses class-attribute interdependence maximization. In *Proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI 2001)*, pages 980–987, 2001.

[31] Mario M Leitao, Siobhan Kehoe, Richard R Barakat, Kaled Alektiar, Leda P Gattoc, Catherine Rabbitt, Dennis S Chi, Robert A Soslow, and Nadeem R Abu-Rustum. Accuracy of preoperative endometrial sampling diagnosis of figo grade 1 endometrial adenocarcinoma. *Gynecologic oncology*, 111(2):244–248, 2008.

[32] Douglas A Levine, Cancer Genome Atlas Research Network, et al. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67, 2013.

[33] Douglas A Levine, Cancer Genome Atlas Research Network, et al. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

[34] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[35] Joannie Lortet-Tieulent, Jacques Ferlay, Freddie Bray, and Ahmedin Jemal. International patterns and trends in endometrial cancer incidence, 1978–2013. *JNCI: Journal of the National Cancer Institute*, 2017.

[36] Terri Madison, David Schottenfeld, Sherman A James, Ann G Schwartz, and Stephen B Gruber. Endometrial cancer: socioeconomic status and racial/ethnic differences in stage at diagnosis, treatment, and survival. *American journal of public health*, 94(12):2104–2111, 2004.

[37] Mohamed Ali Mahjoub and Karim Kalti. Software comparison dealing with bayesian networks. In *International Symposium on Neural Networks*, pages 168–177. Springer, 2011.

[38] Annu Makker and Madhu Mati Goel. Tumor progression, metastasis, and modulators of epithelial-mesenchymal transition in endometrioid endometrial carcinoma: an update. *Endocr Relat Cancer*, 23(2):R85–R111, 2016.

[39] Bruce G Marcot, Richard S Holthausen, Martin G Raphael, Mary M Rowland, and Michael J Wisdom. Using bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest ecology and management*, 153(1-3):29–42, 2001.

[40] Andrea Mariani, Thomas J Sebo, Jerry A Katzmann, Patrick C Roche, Gary L Keeney, Timothy G Lesnick, and Karl C Podratz. Endometrial cancer: can nodal status be predicted with curettage? *Gynecologic oncology*, 96(3):594–600, 2005.

[41] Jessica N McAlpine, Sarah M Temkin, and Helen J Mackay. Endometrial cancer: not your grandmother's cancer. *Cancer*, 122(18):2787–2798, 2016.

[42] Mackay HJ3. McAlpine JN1, Temkin SM2. Endometrial cancer: Not your grandmother's cancer. *Int J Gynaecol Obstet*, pages 96–104, 2015 Oct.

[43] Matthias Meissnitzer and Rosemarie Forstne. Mri of endometrium cancer – how we do it. *Cancer Imaging*, 30(12):1329–34, 2016.

[44] McConechy MK, Ding J, Cheang MC, Wiegand K, Senz J, Tone A, Yang W, Prentice L, Tse K, Zeng T, McDonald H, Schmidt AP, Mutch DG, McAlpine JN, Hirst M, Shah SP, Lee CH, Goodfellow PJ, Gilks CB, and Huntsman DG. Use of mutation profiles to refine the classification of endometrial carcinomas. *J Pathol.*, 22(1):20–30, 2012 Sep.

[45] C Paul Morrow, Brian N Bundy, Robert J Kurman, William T Creasman, Paul Heller, Howard D Homesley, and James E Graham. Relationship between surgical-pathological risk factors and outcome in clinical stage i and ii carcinoma of the endometrium: a gynecologic oncology group study. *Gynecologic oncology*, 40(1):55–65, 1991.

[46] Kevin Murphy. An introduction to graphical models. *Rap. tech*, pages 1–19, 2001.

[47] Radhakrishnan Nagarajan, Marco Scutari, and Sophie Lèbre. Bayesian networks in r. *Springer*, 122:125–127, 2013.

[48] Mikhail Nikulin and Hong-Dar Isaac Wu. The cox proportional hazards model. In *The Cox Model and Its Applications*, pages 35–51. Springer, 2016.

[49] Agnieszka Onisko, Marek J Druzdzel, and R Marshall Austin. How to interpret the results of medical time series data analysis: classical statistical approaches versus dynamic bayesian network modeling. *Journal of pathology informatics*, 7, 2016.

[50] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference (revised 2nd printing)*. Elsevier, 1988.

[51] C. Reijnen, E. Gogou, others, P.J.F. Lucas, and J.M.A. Pijnenborg. Preoperative risk stratification in endometrial cancer (endorisk) by a Bayesian network model: A development and validation study. *PLOS Medicine*, 17(5):e1003111, 2020.

[52] Casper Reijnen, Joanna IntHout, Leon FAG Massuger, Fleur Strobbe, Heidi VN Küsters-Vandevelde, Ingfrid S Haldorsen, Marc PLM Snijders, and Johanna Pijnenborg. Diagnostic accuracy of clinical biomarkers for preoperative prediction of lymph node metastasis in endometrial carcinoma: A systematic review and meta-analysis. *Oncologist*, 24(9), 2019.

[53] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[54] Helga B Salvesen, Ingfrid S Haldorsen, and Jone Trovik. Markers for individualised therapy in endometrial carcinoma. *The lancet oncology*, 13(8):e353–e361, 2012.

[55] Marco Scutari. Bayesian network structure learning, parameter learning and inference, 2011.

[56] Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. Chapman and Hall/CRC, 2014.

[57] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms? *arXiv preprint arXiv:1805.11908*, 2018.

[58] Marco Scutari and Maintainer Marco Scutari. The bnlearn package. *compare*, 18:1, 2007.

[59] Laurie H Sehn, Brian Berry, Mukesh Chhanabhai, Catherine Fitzgerald, Karamjit Gill, Paul Hoskins, Richard Klasa, Kerry J Savage, Tamara Shenkier, Judy Sutherland, et al. The revised international prognostic index (r-ipi) is a better predictor of outcome than the standard ipi for patients with diffuse large b-cell lymphoma treated with r-chop. *Blood*, 109(5):1857–1861, 2007.

[60] Lax SF1. Molecular genetic pathways in various types of endometrial carcinoma: from a phenotypical to a molecular-based classification. *Virchows Arch*, 444(3):213–23, 2004.

[61] Eyre HJ. Smith RA, Cokkinides V. American cancer society guidelines for the early detection of cancer. *American Cancer Society*, 53:27–43, 2003.

[62] American Cancer Society. Endometrial cancer risk factors. `https://www.cancer.org/cancer/endometrial-cancer/causes-risks-prevention/risk-factors.html`. Accessed: 2016-02-10.

[63] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[64] National Cancer Institute Surveillance Research Program, Cancer Statistics Branch. Surveillance, epidemiology, and end results program public use data (1973–1999). `http://seer.cancer.gov/canques`, 2002. Accessed: 2005-06-01.

[65] Alvarez T, Miller E, Duska L, and Oliva E. Molecular profile of grade 3 endometrioid endometrial carcinoma: is it a type i or type ii endometrial carcinoma? *Am J Surg Pathol.*, 36(5):753–61, 2012 May.

[66] Ingvild L Tangen, Reidun K Kopperud, Nicole CM Visser, Anne C Staff, Solveig Tingulstad, Janusz Marcickiewicz, Frédéric Amant, Line Bjørge, Johanna MA Pijnenborg, Helga B Salvesen, et al. Expression of l1cam in curettage or high l1cam level in preoperative blood samples predicts lymph node metastases and poor outcome in endometrial cancer patients. *British journal of cancer*, 117(6):840–847, 2017.

[67] Jone Trovik, Elisabeth Wik, Henrica MJ Werner, Camilla Krakstad, Harald Helland, Ingrid Vandenput, Tormund S Njolstad, Ingunn M Stefansson, Janusz Marcickiewicz, Solveig Tingulstad, et al. Hormone receptor loss in endometrial carcinoma curettage predicts lymph node metastasis and poor outcome in prospective multicentre trial. *European journal of cancer*, 49(16):3431–3441, 2013.

[68] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[69] Serdar Uckun, Kai Goebel, and Peter JF Lucas. Standardizing research methods for prognostics. In *2008 International Conference on Prognostics and Health Management*, pages 1–10. IEEE, 2008.

[70] TS Vermal J udea Pearl. Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence*, pages 220–227, 1991.

[71] Laura Uusitalo. Advantages and challenges of bayesian networks in environmental modelling. *Ecological modelling*, 203:312–318, 2007.

[72] Louis JM van der Putten, Nicole CM Visser, Koen van de Vijver, Maria Santacana, Peter Bronsert, Johan Bulten, Marc Hirschfeld, Eva Colas, Antonio Gil-Moreno, Angel Garcia, et al. L1cam expression in endometrial carcinomas: an enitec collaboration study. *British journal of cancer*, 115(6):716, 2016.

[73] Marcel AJ Van Gerven, Babs G Taal, and Peter JF Lucas. Dynamic bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics*, 41(4):515–529, 2008.

[74] Inge C Van Gool, Ellen Stelloo, Remi A Nout, Hans W Nijman, Richard J Edmondson, David N Church, Helen J MacKay, Alexandra Leary, Melanie E Powell, Linda Mileshkin, et al. Prognostic significance of l1cam expression and its association with mutant p53 expression in high-risk endometrial cancer. *Modern pathology*, 29(2):174, 2016.

[75] Marion Verduijn, Niels Peek, Peter MJ Rosseel, Evert de Jonge, and Bas AJM de Mol. Prognostic bayesian networks: I: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, 40(6):609–618, 2007.

[76] Wikipedia. Endometrium. `https://en.wikipedia.org/wiki/Endometrium`. Accessed: 2018-04-02.

[77] AC AD WYATT. Prognostic models: clinically useful or quickly forgotten? *Br Med J*, 311:539–541, 1995.

[78] Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In *Data mining, fifth IEEE international conference on*, pages 4–pp. IEEE, 2005.

[79] Geels YP, van der Putten LJ, van der Steen-Banasik EM, Snijders MP, Massuger LF, and Pijnenborg JM. The opinion of gynecologists on the management of early-stage, high-grade endometrioid endometrial cancer. *European Journal of Gynaecological Oncology*, 36(4):402–405, Jan 2015.

[80] Alain G Zeimet, Daniel Reimer, Monica Huszar, Boris Winterhoff, Ulla Puistola, Samira Abdel Azim, Elisabeth Müller-Holzner, Alon Ben-Arie, Léon C Van Kempen, Edgar Petru, et al. L1cam in early-stage type i endometrial cancer: results of a large multicenter evaluation. *Journal of the National Cancer Institute*, 105(15):1142–1150, 2013.

[81] Kilian Zwirglmaier and Daniel Straub. A discretization procedure for rare events in bayesian networks. *Reliability Engineering & System Safety*, 153:96–109, 2016.