Heliyon 9 (2023) e20360

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Survival analysis of recurrent breast cancer patients using mix Bayesian network

Parviz Shahmirzalou [a], Majid Jafari Khaledi [b], Maryam Khayamzadeh [c], Aliakbar Rasekhi [a,*]

[a] *Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran*
[b] *Department of Statistics, Tarbiat Modares University, Tehran, Iran*
[c] *Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

ARTICLE INFO

Keywords:
Mixed Bayesian network
Directed acyclic graph
R package deal
Breast cancer
Recurrence

ABSTRACT

Introduction: Breast cancer (BC) is the most common cancer among women. Iranians have an 11% BC recurrence rate, which lowers their survival rates. Few studies have investigated cancer recurrence survival rates. This study's major purpose is to use a mixed Bayesian network (BN) to analyze recurrent patients' survival.
Material and methods: This study aimed to evaluate the pathobiological features, age, gender, final status, and survival time of the patients. Bayesian imputation was used for missing data. The performance of BN was optimized through the utilization of a blacklist and prior probability. After structural and parametric learning, posterior conditional probabilities and mean survival periods for the node arcs were predicted. The hold-out technique based on the posterior classification error was used to investigate the model's validation.
Results: The study included 220 cancer recurrence patients. These patients averaged 47 years old. The BN with a blacklist and prior probability has a higher network score than other networks. The hold-out technique verified structural learning. The Directed Acyclic Graph showed a statistically significant relationship between cancer biomarkers (ER, PR, and HER2 receptors), cancer stage, and tumor grade and patient survival duration. Patient death was also significantly associated with education, ER, PR, HER2, and tumor grade. The BN reports that HER2 negative, ER positive, and PR positive patients had a higher survival rate.
Conclusion: Survival and death of relapsed patients depend on biomarkers. Based on the findings, patient survival can be predicted with their features.

1. Introduction

Cancer is one of the most common diseases in humans. In 2020, in the whole world (Iran), cancer has led to the death of 10 million (79,136) people, which includes one case out of every 10 deaths [1,2]. Although there are many types of cancer, some of them have high prevalence. In 2020, in the whole world (Iran), about 2.3 million (16,967) new cases and 685 thousand (4,810) deaths from Breast Cancer (BC) were reported. BC is very prevalent among women, as evidenced by multiple sources [2,3]. According to a national study

* Corresponding author.
E-mail addresses: p.shahmirzalou@modares.ac.ir (P. Shahmirzalou), jafari-m@modares.ac.ir (M.J. Khaledi), khayamzadeh@yahoo.com (M. Khayamzadeh), rasekhi@modares.ac.ir (A. Rasekhi).

https://doi.org/10.1016/j.heliyon.2023.e20360
Received 11 July 2023; Received in revised form 6 September 2023; Accepted 20 September 2023
Available online 21 September 2023

in Iran, most of the patients are under 50 years old [4]. The follow-up of patients with BC shows that about 11% of them experience disease recurrence [5], which reduces the amount of time they survive (and may result in death or censoring). As a result, identifying the causes of recurrence and conducting an investigation can aid in preventing BC from occurring again. The assessment of the literature reveals that only a small number of research have been carried out only on recurrent patients, with the majority of studies concentrating their analysis on all patients [6–8]. This is why the current study's focus is on the survival analysis of individuals who have experienced recurrence. The missing data is a constant issue in survival analysis and this feature at survival time is defined as censoring. For example, in right censoring, the survival time is equal or more than the recorded time [9]. Missing data imputation is done to maximize the use of available information. So far, different methods have been proposed for imputation (average, median, single and multiple imputation), each of which has its strengths and weaknesses [10–12]. However, the Bayesian approach has some advantages for imputing incomplete datasets, including high efficiency in low sample sizes, the ability to handle heavy censoring, and the ability to assign prior distributions to parameters [13,14]. After analyzing the specifics of various methodologies, we finally used a Bayesian strategy to perform imputation of censored data in the current investigation, after which we examined the patient survival rates.

There exist multiple methodologies for evaluating patient survival, such as the Kaplan-Meier, Cox regression, parametric [9] and non-parametric [15] techniques. In the majority of the aforementioned methodologies, every variable is incorporated into the model either as a predictor or as a dependent variable. In reality, it is common for one variable to have an impact on other variables and to be influenced by them as well. This implies that certain components have the potential to fulfill dual roles concurrently. Then, for each role of that variable, a separate model should be created or methods should be used that a variable can play both roles at the same time. The selection of a suitable methodology represents a further challenge that is addressed within the scope of this investigation. There are two excellent approaches if we want a cause-and-effect perspective on relationships. In causal discussions, Structural Equation Models (SEM) and Bayesian network (BN) are always mentioned. Both methods are used to identify causal relationships between study variables, although they have differences in application that are mentioned. The SEM methodology is founded on the concepts of cause and effect, but the BN framework incorporates probabilistic causality, wherein the presence of a cause enhances the likelihood of the corresponding effect. Furthermore, BNs are utilized for both prediction and diagnosis, whereas SEM is predominantly employed for prediction purposes. Unlike the SEM, the relationships identified by the BN are valid for new data with the same variables as well [16]. Based on the mentioned topics, the BN is the focus of this study.

BNs are composed of variables, also known as vertices or nodes, which exhibit probable (causal) relationships with one another. These relationships are represented by directed arcs, sometimes referred to as edges. In other words, a BN consists of two parts: quantitative and qualitative. The qualitative component pertains to the Acyclic Directed Graph (DAG), whereas the quantitative component relates to potential relationships. In a DAG, all the arcs (edges) have directions, and a circle is not allowed [17]. BNs are classified into discrete, continuous and mixed types. Continuous BN uses variables with a quantitative scale, while discrete BN uses variables with a qualitative scale [18]. Despite the potential loss of information, discrete BNs can effectively handle continuous variables by categorizing them. However, employing a mixed Bayesian network is considered more advantageous. The mixed BN analyzes the relationships between continuous and discrete variables simultaneously. The current investigation encompasses both continuous factors, such as age and survival time, and discrete variables, such as pathobiological characteristics. The majority of studies primarily focus on quantifying the impact of different variables on survival time. However, a distinctive aspect of the present studies is their examination of the interrelationships among many variables. Based on the aforementioned documentation, the utilization of mixed Bayesian networks is deemed valuable and preferred. So far, numerous R packages have been published under R for data analysis with BN. The utilization of R packages, such as catnet for discrete nodes [19], pcalg for discrete and continuous nodes [20], gRbase for continuous and discrete nodes [21], rbmn for continuous nodes [22], has been documented. A comprehensive comparison of the subjects is detailed in the publication authored by Scutari et al. [23] namely in the book of BNs. This study centers its attention on two comprehensive R packages. The review of available literature has revealed that two specific programs, namely bnlearn and deal, have been identified as suitable tools for the purpose of constructing Bayesian Networks (BN) that incorporate both continuous and discrete data. The deal package possesses the capability to conduct both structural and parametric learning. It stands out among other R packages due to its unique ability to handle mixed Bayesian networks within the context of conditional Gaussian distribution. The hill-climbing greedy search algorithm is utilized in this package for conducting structural learning [23].

A further advantage is the ability to set up a black list and prior probability distributions. As an illustration, the prior probability distribution may be configured such that the probability of mortality is greater for those with an advanced cancer stage compared to those with an early cancer stage. Furthermore, the utilization of a black list allows for the establishment of a structural learning process wherein the age of the patient consistently serves as a vertex of the root. This implies that no arc is connected to the age of the patient [24]. The aforementioned settings enhance the accuracy of learning. In order to investigate and confirm this matter, four configurations of the BN will be fitted, including both the presence and absence of the blacklist and prior probability. The resulting change in the score index will be documented. The aim of this study is to establish a network of connections among the pathobiological characteristics of patients experiencing recurrence. This will be accomplished using a Mixed BN and the deal R packages. In order to enhance the accuracy of predictions, the utilization of settings pertaining to the prior probability distribution and the black list is employed. Consequently, predictions are made regarding the probability of mortality and the average duration of survival for the pathobiological attributes of patients.

2. Material and Methods

2.1. Study dataset

The current investigation is a comprehensive analysis of the BC data received from the Cancer Research Center of Shahid Beheshti University of Medical Sciences, spanning from 2015 to 2022. The study's inclusion criteria pertain to the presence of cancer recurrence in the patient. In this study, the features of the patients were investigated, which encompassed survival or censoring time, patient age at recurrence, cancer stage, tumor grade, Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal growth factor Receptor-2 (HER2) receptors [25].

2.2. Imputation missing data

The time to recurrence of the patients is a significant variable that is being investigated, with approximately 40% of the data being censored. In order to address this issue, a Bayesian strategy was employed for imputing the censored values. Imputation has been performed in accordance with the stated purpose [13]. A total of 11,000 simulations were conducted for each censored time, with

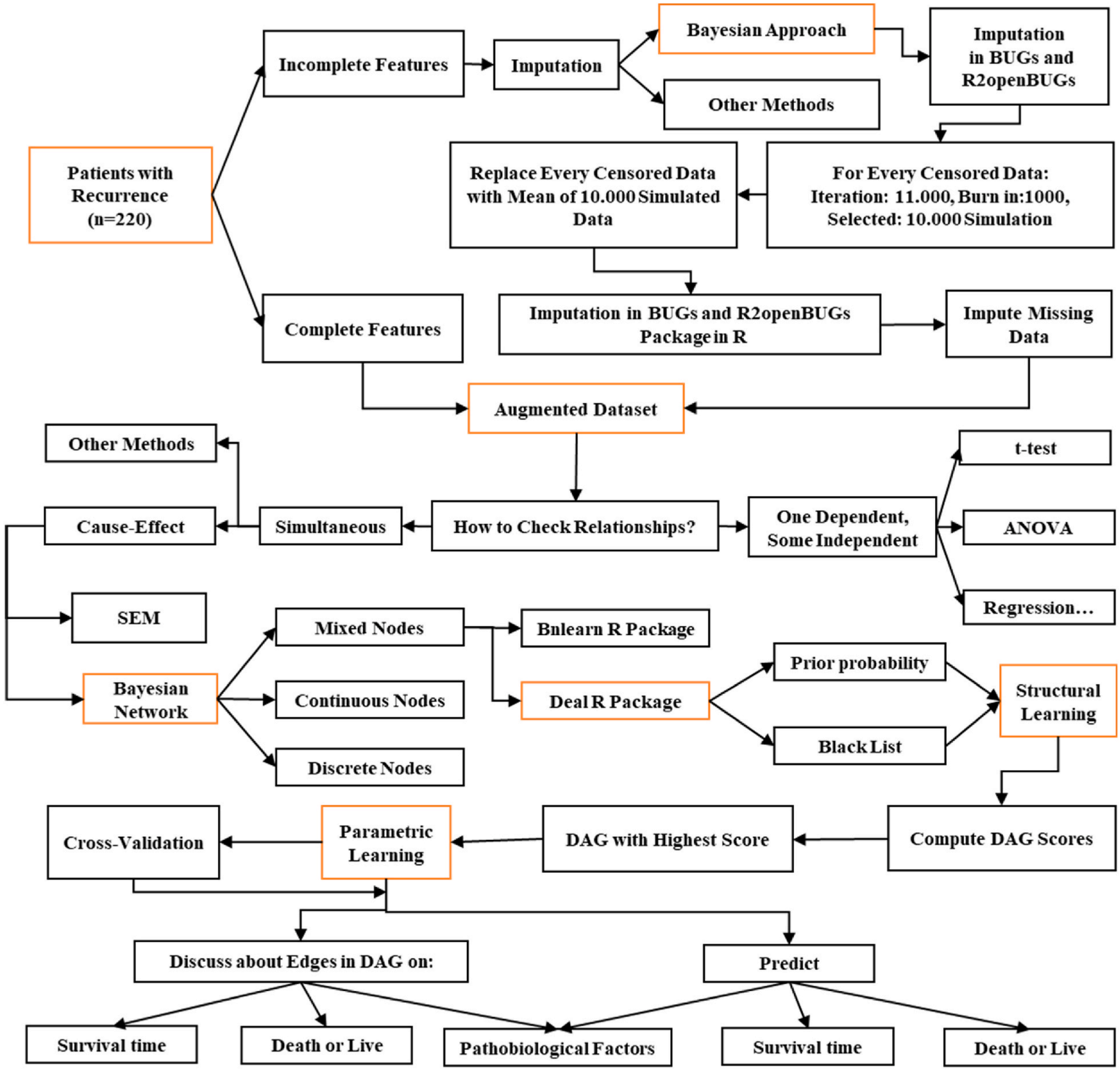

Fig. 1. Steps to prepare dataset and run BN on BC dataset.

exactly 1000 simulations used for burn-in and the remaining 10,000 simulations accessible for analysis. The censored times are replaced with the average of 10,000 simulated times. By following these procedures, the mean simulated times are substituted for the censored times, resulting in the acquisition of a complete dataset.

2.3. Bayesian network

Once the incomplete data has been inputted, the BN can be fitted to the Breast Cancer (BC) data. A BN is represented as a directed acyclic graph, denoted as $D = (V,A)$, where V represents the set of vertices and A represents the set of arcs. The vertices of the graph are composed of both continuous (Γ) and discrete (Δ) variables, and can be represented as $V = \Delta \cup \Gamma$. The mixed Bayesian network is fitted using the deal software package, given that the variables under study are both continuous ($\gamma \in \Gamma$) and discrete ($\delta \in \Delta$). Each level of the discrete variable is denoted by the index i, while each value of the continuous variable is represented by the symbol y. In Bayesian networks, the probability distribution of a vertex is computed based on the presence of its parents. The distribution of the conditional probability for an arbitrary discrete vertex in subgroup i and a continuous variable y, given its parents (pa), can be denoted as $p(i_\delta \big| i_{pa(\delta)}))$ and $p(y_\gamma \Big| i_{pa(\gamma)}, y_{pa(\gamma)}))$ respectively. The joint probability distribution of a BN can be expressed as the multiplication of the probability distributions of its continuous and discrete vertices, conditioned on their respective parents [24].

2.4. Structure and parametric learning

In this study, it is assumed that the conditional probability distributions of discrete and continuous vertices are, respectively, uniform and conditional Gaussian. In the field of BN, it is important to include both structural learning, which involves figuring out of the DAG, as well as parametric learning. Parametric learning is conducted within the framework of two specific circumstances. The parameters (θ) associated with each vertex are mutually independent and also independent of the parameters associated with other vertices [26]. Consequently, it is imperative that we acquire knowledge in the following manner:

$$p(\theta) = \left[\prod_{\delta \in \Delta} \prod_{i_{pa(\delta)} \in I_{pa(\delta)}} p\left(\theta_{\delta | i_{pa(\delta)}}\right) \right] \left[\prod_{\gamma \in \Gamma} \prod_{i_{pa(\gamma)} \in I_{pa(\gamma)}} p\left(\theta_{\gamma | i_{pa(\gamma)}}\right) \right] \tag{1}$$

Two components make up equation [1]. The initial section presents the probability distribution of discrete vertices, whereas the subsequent section illustrates the probability distribution of continuous vertices. In the BN, the prior distribution of the hyper-parameters is frequently thought to be conjugate, making it simple to estimate the parameters from the posterior distribution [24,27].

By employing a deal, it is possible to fit many Bayesian networks to a given data set. The main criteria for selecting the best possible BN is the evaluation of network scores. The scoring mechanism in the mixed BN is determined by multiplying the scores assigned to each discrete and continuous vertex. This scoring approach is elucidated in the deal package's manual [26]. The ultimate Bayesian network will possess the maximum score. By utilizing the final BN, it becomes possible to investigate, infer, and even predict the connections among the vertices. Fig. 1 illustrates the sequential process of implementing BN using Breast Cancer (BC) data.

Before the structural and parametric learning, the blacklist was set for the BN (Table 1). The survival time and the final status (death or live (censor)) of the patients are the terminal record of the study. Accordingly, no arc comes out of them, but they receive arcs from other nodes (a leaf). On the other hand, the demographic variable such as age is not influenced by the pathobiological characteristics and education of the patients, therefore, only the arc comes out of them.

The biomarkers ER, PR, and HER2 have been identified as prognostic and predictive factors [28,29] In BN model, there are no incoming arcs directed towards these biomarkers. Some variables, including the stage of disease and the grade of the tumor, have an open input and output arc and are not on any blacklists. Table 1 displays additional things. The prior probability distribution is an important component in both structural and parametric learning. The default configuration of the deal package entails the assignment of equal prior probabilities to subgroups of qualitative variables. These options lack validity, especially when taking into account the qualitative traits related to cancer. The prior probabilities for this analysis were determined by referencing the thorough studies on a cohort of 3010 Iranian breast cancer patients [5] (Table 2).

Table 1
The Black list of BN. The ✓ symbol is used to represent edges that are considered invalid or should not be present.

To From	Age	Education	ER	PR	HER2	Grade	Stage	Status	Time to event
Time to event	✓	✓	✓	✓	✓	✓	✓	✓	✓
Status	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age	✓		✓	✓	✓				
Education	✓	✓	✓	✓	✓	✓			
ER	✓	✓	✓	✓	✓				
PR	✓	✓	✓	✓	✓				
HER2	✓	✓	✓	✓	✓				
Grade	✓	✓	✓	✓	✓	✓			
Stage	✓	✓	✓	✓	✓		✓		

ER: Estrogen Receptor, PR: Progesterone Receptor, HER2: Human Epidermal Growth factor Receptor 2.

Table 2
Prior probabilities for subgroups of categorical variables.

Node	Probability	Node	Probability
ER		**PR**	
Negative	0.55	Negative	0.55
Positive	0.45	Positive	0.45
Stage of Cancer		**HER2**	
1	0.20	Negative	0.45
2	0.23	Positive	0.55
3	0.27	**Education level**	
4	0.30	Illiterate	0.22
Grade of tumor		Elementary school	0.22
Low	0.25	Middle school	0.20
Moderate	0.30	High school diploma	0.20
High	0.45	University	0.16

ER: Estrogen Receptor, PR: Progesterone Receptor.

2.5. *Cross-validation*

The validation of structural learning was conducted using the hold-out technique inside the Bnlearn R package. The dataset was partitioned into five distinct subsamples, denoted as K = 5. Each subsample underwent a validation process that was iterated twenty times, referred to as runs = 20. The allocation of training data to test data was established at a ratio of 70%–30%. The loss function was defined as the posterior classification error (PCE) using likelihood weighting [30]. In PCE, predictions are generated using likelihood weighting from a random collection of nodes to produce Bayesian posterior estimates. Hybrid networks use the pred-lw-cg classification criterion. Lower values are preferable. The expected loss function was computed separately for every discrete nodes through the utilization of PCE [23,31].

3. Results

This study incorporated the data of 220 female patients who experienced recurrent breast cancer. The study observed a 40% censoring rate in the time to death of patients, which was subsequently addressed through imputation using the Bayesian approach. Table 3 presents an in-depth description of the demographic and pathobiological attributes of the patients.

The degree of collinearity among the variables was assessed using the VIF. The maximum level of collinearity observed was 1.95, indicating a relatively low degree of severity. In order to assess survival durations, it is customary to construct the Kaplan-Meier curve in relation to variables. Fig. 2 depicts the overall survival rates of patients who had relapse, while Fig. 3 showcases the Kaplan-Meier curve based on patient characteristics. The log-rank test has yielded statistically significant findings on the disparity in patient survival based on factors such as cancer stage, tumor grade, and the presence of ER and PR receptors.

To conduct an initial analysis of the factors impacting the time to patients' death, we employed the Cox regression model to assess all independent variables. The findings are presented in Table 4. There exists a statistically significant correlation between the time to death of patients and both their age and cancer stage. Furthermore, a statistically significant association ($\alpha = 0.10$) has been observed between the variables under investigation, namely tumor grade and progesterone receptor.

In order to ascertain the precise associations, structural learning was conducted for the variables under study using the deal R package. To elucidate the impact of employing the blacklist and configuring the prior probabilities on enhancing the score of the BN, the BN was executed in four distinct modes [1]: without the inclusion of the prior probability and the blacklist [2], without the blacklist

Table 3
Description of patient features in the study.

Node	Frequency (%)	Node	Frequency (%)
HER2		**Grade of tumor**	
No	93 (42.3%)	1	7 (3.2%)
Yes	127 (57.7%)	2	105 (47.2
Stage of Cancer		3	108 (49.1%)
1	10 (4.5%)	**ER**	
2	69 (31.4%)	No	86 (39.1%)
3	106 (48.2%)	Yes	134 (60.9%)
4	35 (15.9%)	**Education level**	
PR		Illiterate	56 (25.5%)
No	93 (42.3%)	Elementary school	71 (32.3%)
Yes	127 (57.7%)	Middle school	31 (14.1%)
Final status		High school diploma	38 (17.3%)
Live	88 (40%)	University	24 (10.9%)
Death	132 (60%)	**Age** (mean (SD))	47 (11.73)
Survival time (mean (SD))	137.96 (125.15)		

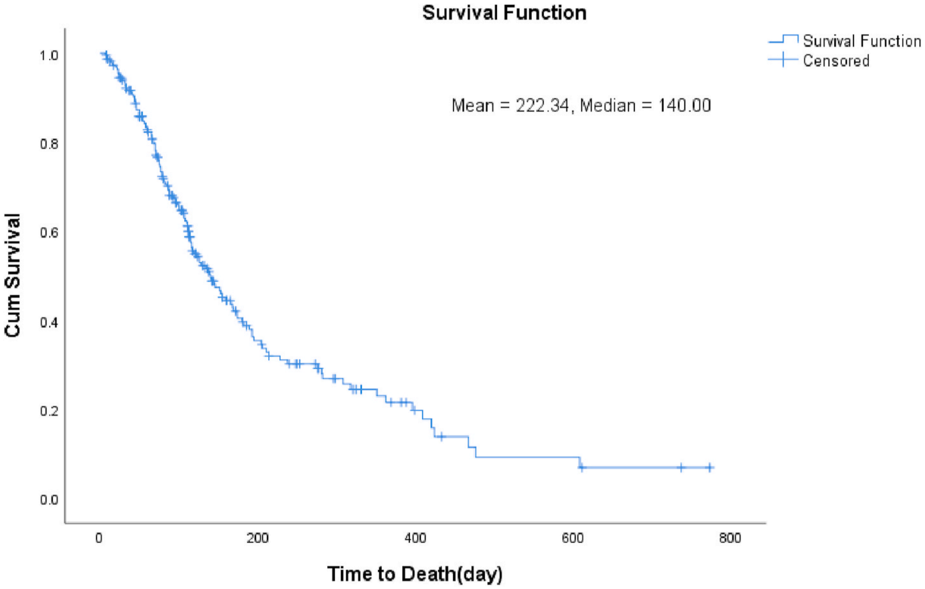

Fig. 2. Kaplan-Meier Survival plot for Patients with Recurrence.

but with the incorporation of the prior probability [3], with the inclusion of the blacklist but without the prior probability, and [4] with both the blacklist and the prior probability. The incremental change in the score index, from the BN without arcs (empty graph) to the Bayesian network with the highest score, was computed at each stage. The aforementioned conditions yielded the following calculations: 98.95, 450.19, 57.55, and 704.93. The data indicates that the most significant enhancement in the score index has been observed when both the blacklist and prior probabilities are simultaneously configured. Accordingly, the BN was fitted using the black list and prior probability values. The score of the initial and final (DAG A) DAG for the aforementioned BN is 4803.8 and 4098.8, respectively. Fig. 4 presents the DAG that have the greatest scores.

The hold-out technique was employed to validate the process of structural learning. The evaluation of the predictive capability of structure learning in determining the required vertex in the final graph (Fig. 4) was conducted using the Posterior Classification Error (PCE) criterion. Table 5 presents the average (standard deviation) values of the loss function across several categorical nodes.

The calculated loss function for the classification target node using the deal, HC, and Tabu graphs exhibits a remarkably low value. The concept of structural learning is well-supported. Following the structural and parametric learning process, predictions about the final status and survival time for recurrent patients can be made in light of their unique characteristics (see Tables 1–4 in the Supplementary Material). The research appendix, namely Tables 1–4, presents a comprehensive analysis of the conditional probabilities associated with the patients' ultimate state and average survival duration. These probabilities are further classified based on various patient characteristics.

4. Discussion

This research investigates the connection between pathological characteristics, such as the stage of cancer and grade of tumor, biomarkers including ER, PR, and HER2, as well as age and education, with the recurrence time to death and final outcome (death or survival) via BN. Multiple investigations have demonstrated that cancer recurrence significantly influences patient survival [6,32]. Nevertheless, the existing literature on the subject of recurrence and their impact on survival rates is quite limited. Given these circumstances, it is crucial to undertake a thorough evaluation of the extended outlook for individuals experiencing recurrence of cancer utilizing a specific methodology. This study investigated a cohort of 220 patients who experienced cancer recurrence. The researchers utilized the Bayesian network analysis method to explore the factors that influence the survival of these patients. The BN improperly applies a uniform prior distribution to subgroups of qualitative vertices, which is not appropriate for analyzing cancer data. For example, there is a discrepancy in the probability of cancer recurrence between the initial and terminal stages. In order to achieve this objective, the prior probability for the subgroups of qualitative vertices was determined by consulting source and recording the corresponding data in Table 2 [5]. The BN was fitted using the deal R package, as outlined in the materials and methods section.

Multiple studies have repeatedly demonstrated that the age of patients is a significant factor in predicting their survival results. Nevertheless, there is a dearth of studies investigating the significant influence of age specifically on the survival rates of persons who undergo cancer recurrence. The current study aims to explore the relationship between the age of patients at the time of cancer recurrence and the duration of survival among patients who experience recurrence. The analysis reveals a correlation coefficient of −0.13 between the variables, with a corresponding P-value of 0.056. Moreover, the statistical significance of this link is supported by the results of regression analysis, as indicated by a P-value of 0.028. The average survival time of patients who experience recurrence is 170 days, with a standard variation of 132 days. To clarify, it may be inferred that the recurrence of the disease significantly

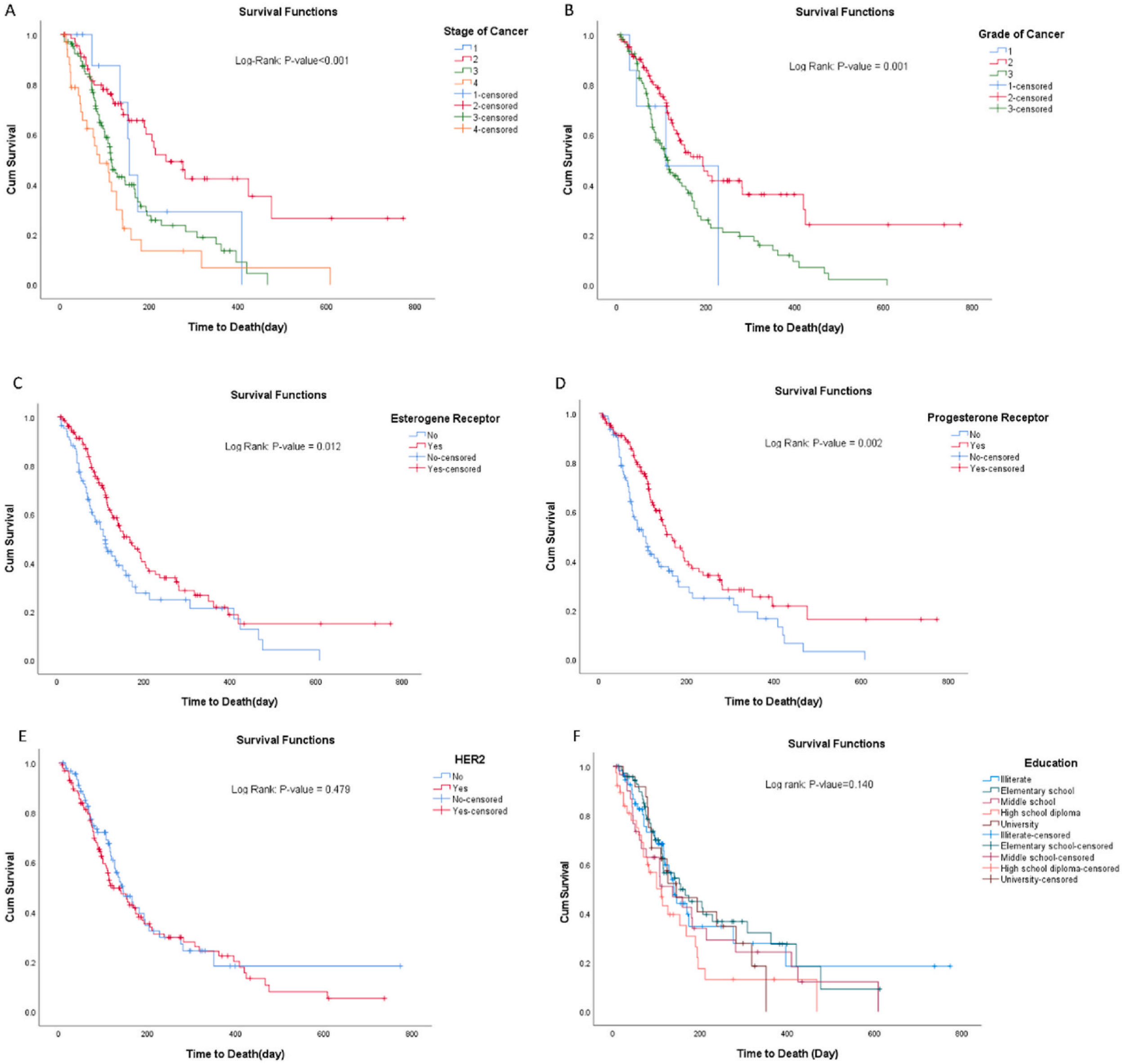

Fig. 3. Kaplan-Meier survival plots adjusted to study Covariates (A: Stage of Cancer, B: Grade of tumor, C: Estrogen receptor, D: Progesterone receptor, E: HER2, F: Education level).

Table 4

Results of Cox regression model for all of study covariates based on Backward: LR variable selection method.

Covariate	B(SE)	*P*-value	Hazard Ratio
Age at recurrence	0.02 (0.009)	.028	1.019
Stage of Cancer	0.34 (0.126)	.007	1.409
Grade of Cancer	0.37 (0.187)	.050	1.444
Progesterone Receptor	−0.35 (0.185)	.061	0.707

diminished the patients' survival rates, potentially leading to the lack of statistical significance in the association identified in the Bayesian network (DAG in Fig. 4).

According to a study conducted by Marti et al. the age of patients has been identified as a risk factor for recurrence among individuals who have survived for more than 10 years [33]. There exists a lack of information regarding the survival rates of persons who undergo recurrence across different age groups. In previous research [34], researchers observed a statistically significant relationship

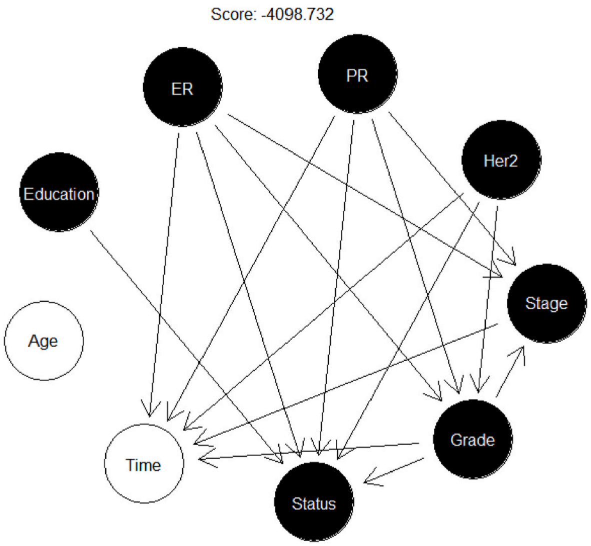

Fig. 4. The last DAG with high scores (Relative score = 1) extracted from mix BN using package deal. Discrete vertices are represented by black circles.

Table 5
Average (SD) loss function over the 20-runs in 10-folded cross-validation.

Node	Bayesian Network		
	deal	HC	Tabu
Status	0.40 (0.02)	0.32 (0.01)	0.32 (0.02)
Stage of Cancer	0.52 (0.02)	0.52 (0.02)	0.52 (0.02)
Grade of Tumor	0.43 (0.02)	0.46 (0.03)	0.46 (0.03)
HER2	0.48 (0.02)	0.46 (0.02)	0.46 (0.02)
ER	0.23 (0.01)	0.17 (0.01)	0.17 (0.01)
PR	0.23 (0.01)	0.16 (0.01)	0.17 (0.01)

between age and the survival probability of patients experiencing recurrent illnesses. However, this link was only observed among individuals aged 65 years and older, with a Hazard Ratio of 1.44. It seemed that additional research is necessary to examine the potential association between age and the survival outcomes of individuals who have had recurrence. This is particularly important due to the observed restricted period of survival following relapse in the present study.

Other characteristics that have been looked into in patients include the disease stage and tumor grade. Given that the study's target population includes recurrent patients, the proportion of patients with early-stage cancer and low-grade tumors is rather low. Because odds of recurrence in patient with initial stage and lower tumor grade is very low. The previously mentioned matter is also observable in the Kaplan-Meier curve depicted in Fig. 3. In the BN, the survival time and final condition of recurrent patients are significantly influenced by the tumor grade of the disease, while only the survival time is impacted by the stage of cancer. A closer look at the Supplementary Table 1 indicates that patients with cancer in stages three and four are more likely to die than other patients. Furthermore, taking into account the impact of the HER2 receptor on prognosis and treatment approaches, we conducted a comparative analysis between the predicted probability estimation of the vertices described in Supplementary Table 2 and the survival outcomes of relapsed patients in relation to the HER2 variable. As the tumor grade and cancer stage escalate, there is a notable decrease in the average survival duration of the disease, as depicted in Supplementary Table 2. An examination of Supplementary Table 2 reveals that patients with a positive HER2 receptor consistently exhibit a longer mean survival duration, irrespective of tumor grade and disease stage. The Supplementary Table 3 depicts the average survival time, which includes comparisons based on the stage of cancer and various tumor grades, in adition to the previously discussed comparisons. The authors of the current study have indicated that there is either no observable synergistic relationship or a minimal one between the average survival time and the stage of cancer, as well as tumor grade. The identical matter was examined in comparable research efforts.

Sopik et al. conducted a study using a sample of 1675 women diagnosed with breast cancer, specifically focusing on cases of local recurrence and early stages of the disease, including Ductal Carcinoma In-Situ (DCIS) as well as the first and second stages of cancer. The incidence of disease recurrence among female patients diagnosed with DCIS, as well as those in the first and second stages of cancer, was documented at 16%, 15%, and 16%, respectively, over a 15-year period of observation. Furthermore, previous studies have documented mortality rates of 16%, 32%, and 59% in the aforementioned groups following relapse, which indicates a relationship between patients' cancer stage and their survival outcomes [35]. Kawaguchi et al. conducted an additional investigation to examine the relationship between the survival rates of relapsed patients and certain biomarkers, as well as tumor grade. This study aimed to analyze and evaluate the survival rates of patients with grade 2 and 3, in comparison to those with grade one, utilizing a sample size of 598 patients. The results of the multivariate analysis revealed a statistically significant disparity in patient survival rates based on the severity of the grade, with patients classified as grade three to one exhibiting a significantly lower survival rate (P-value$<$0.001, Hazard Ratio $=$ 1.87). Furthermore, it has been observed that patients who have a lower grade of tumor or a lesser degree of cell abnormality tend to have a more favorable prognosis and increased survival rates [36].

Furthermore, this study aimed to examine the impact of biomarkers on patient mortality, survival rates, and average duration of illness. All three biomarkers have an impact on the duration and final status of patient survival. According to the analysis of Supplementary Table 1 and it is observed that the probability of mortality in individuals with ER negative and HER2 positive is somewhat higher compared to the control group. In this study, we investigated the combined impact of three biomarkers on the mean survival time of patients by analyzing their pairwise interactions (Supplementary Table 1). The provided reports indicate that patients who are HER2-negative, ER-positive, and PR-positive exhibit a comparatively extended average projected survival duration. On the contrary, patients diagnosed with triple-negative breast cancer (TNBC) have exhibited a significantly reduced overall survival rate. Additionally, it can be observed from Supplementary Table 4 that the presence of both estrogen and progesterone receptors corresponds with a higher level of patient survival. It is evident that when all three biomarkers are present, the impact of biomarkers on survival may exhibit variations compared to the associations observed in individual and combined contexts. In order to examine the impact of three factors on the survival of relapsed patients, a total of eight combinations of three biomarkers were retrieved and afterwards depicted in Supplementary Table 1. According to the findings of this study, patients who exhibit triple negative status demonstrate a decreased mean survival rate and exhibit aggressive disease characteristics. Conversely, patients who possess ER-positive, PR-positive, and Her2-negative exhibit the greatest recorded mean survival rate. The researchers hypothesized that patients with ER-positive, PR-positive, and HER2-negative receptors would exhibit the longest projected survival time in this investigation. On the contrary, patients diagnosed with triple-negative breast cancer (TNBC) exhibit the shortest overall survival duration.

In 2021, Lu et al. did a study that investigated the involvement of biomarkers in breast cancer cases, both with and without recurrence. A total of 156 individuals who had experienced relapse were included as participants in this study. This study presents findings indicating a statistically significant disparity in patient survival rates based on the presence or absence of estrogen and progesterone receptors. To clarify, it was shown that patients who had recurring instances and possessed estrogen and progesterone receptors exhibited improved and extended survival rates [37]. In the same investigation, the use of Cox regression was employed for analyzing pathobiological elements that impact the survival of recurrent patients. The present study involved a sample size of 442 patients and its findings indicate that, according to multivariate Cox regression analysis, only the influence of the number of positive lymph nodes and recurrence tissue on the survival of recurrent patients was found to be statistically significant. Conversely, factors such as age, disease stage, tumor size, surgical method (MRM compared to BCS), and the presence of HER2 receptor, TNBC, and luminal B subtype (compared to luminal A subtype) were not found to have a statistically significant effect on patient survival. The survival rate of patients experiencing visceral recurrence is comparatively lower in comparison to those with local and bone recurrence [38].

The level of education among patients is a factor that potentially influences patient mortality rates. This study demonstrates that the educational attainment of patients exerts a notable impact on the survival outcomes of individuals who have experienced a relapse. Patients who have a history of recurring visits and possess a lower level of education have a diminished probability of mortality, as reported in Supplementary Table 1. Conversely, patients with a higher level of education exhibit an increased risk of mortality. Given the comprehensive care and ongoing monitoring provided to all patients within the cancer research center, it is important to acknowledge that the aforementioned observation regarding a greater chance of mortality among individuals with higher levels of education may be influenced by various factors, including the cancer stage, patient age, and other individual characteristics.

A study conducted by Hjorth et al. (2021) examined a sample of 286 recurring cases, including a total of 2616 patients. The findings of this study indicated that individuals with lower levels of education had a significantly increased risk of mortality. A review of the graph pertaining to education in the present study reveals a positive relationship between the duration of time and the mortality risk for patients across all levels of education. Despite the aforementioned conclusion being given in the analysis conducted within this study, an inspection of the education-related graph shown in the study (Fig. 1) reveals that there is no statistically significant disparity in the mortality risk among patients with varying levels of education [39]. Based on this, it seems that additional studies are important to identify the effect of education level on recurrence, survival and mortality of relapsed patients.

Based on this, the level of education, grade of tumor, ER, PR, and HER2 have been shown to be effective variables on the death of recurring patients. Furthermore, the average survival of recurrent patients is significantly influenced by the presence of all three biomarkers (ER, PR, and HER2), as well as the disease stage. From the reports in this study, we can draw an essential link about HER2. This suggests that possessing the HER2 receptor worsens the patient's prognosis. Finally, the effect of education level and patient's age at the time of cancer recurrence on the survival of recurring patients' needs additional exploration.

5. Conclusion

The objective of this study is to examine and ascertain the associations among pathobiological factors, age of cancer diagnosis, and educational attainment with regards to patient mortality or survival, as well as the mean survival duration of patients. Up to this moment, numerous studies has investigated and elucidated the aforementioned associations using various statistical methodologies, such as simple description, linear regression, logistic regression, and Cox regression. The study population comprises patients with and without recurrence of cancer, with only a limited number of studies focusing specifically on recurring patients. Hence, the significance of this work lies in its originality, which can be viewed from two distinct perspectives. The focus of the present investigation pertains to those who have experienced recurrence of cancer. The investigation of factors influencing mortality and survival rates, as well as the estimation of average survival duration, holds significant moral and humanitarian importance for this cohort of patients. The second innovation pertains to the approach employed for surveying relationships. The relationships identified in the BN exhibit a cause-effect nature, where the presence of an arc from one node to another indicates the probability of causation. To enhance the accuracy of learning, BNs have incorporated rules such as the black list and prior probabilities to ascertain the connections among variables. This aids in the identification of links between variables.

Author contribution statement

Parviz Shahmirzalou: Aliakbar Rasekhi: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Majid Jafari Khaledi: Maryam Khayamzadeh: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Data availability statement

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e20360.

References

[1] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, et al., Global Cancer Observatory: Cancer Today, International Agency for Research on Cancer, Lyon, 2020 [Available from: https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] 2020, Cancer in Islamic Republic of Iran, Globocan, 2020 [Available from: https://gco.iarc.fr/today/data/factsheets/populations/364-iran-islamic-republic-of-fact-sheets.pdf.

[3] Breast Cancer: World Health Organization, 2021 [Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[4] M.E. Akbari, A. Akbari, M. Khayamzadeh, R. Salmanian, M. Akbari, Ten-year survival of breast cancer in Iran: a national study (retrospective cohort study), Breast Care 18 (1) (2023) 12–21.

[5] M.E. Akbari, S. Sayad, S. Sayad, M. Khayamzadeh, L. Shojaee, Z. Shormeji, et al., Breast cancer status in Iran: statistical analysis of 3010 cases between 1998 and 2014, Int. J. Breast Cancer 2017 (2017), 2481021.

[6] A.R. Baghestani, P. Shahmirzalou, F. Zayeri, M.E. Akbari, M. Hadizadeh, Prognostic factors for survival in patients with breast cancer referred to omitted cancer research center in Iran, Asian Pac. J. Cancer Prev. APJCP 16 (12) (2015) 5081–5084.

[7] Z. Li, S. Yin, L. Zhang, W. Liu, B. Chen, Prognostic value of reduced E-cadherin expression in breast cancer: a meta-analysis, Oncotarget 8 (10) (2017) 16445–16455.

[8] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, A. Stanisławek, Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review, Cancers 13 (17) (2021).

[9] D.G. Kleinbaum, M. Klein, Survival Analysis: A Self-Learning Text, third ed., Springer, 2012.

[10] M. Geraci, A. McLain, Multiple imputation for bounded variables, Psychometrika 83 (4) (2018) 919–940.

[11] N. Erler, Multiple Imputation of Missing Data in Simple and More Complex Settings, Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Kiel, 2019.

[12] M. Jalali, Z. Bagheri, N. Zare, S.M.T. Ayatollahi, A new method for imputing censored values in crossover designs with time-to-event outcomes using median residual life, BioMed Res. Int. 2020 (2020), 8475154.

[13] S. Moghaddam, J. Newell, J. Hinde, A bayesian approach for imputation of censored survival data, Stats 5 (1) (2022) 89–107.

[14] J.G. Ibrahim, M.-H. Chen, D. Sinha, Bayesian Survival Analysis, Springer series in statistics, New York, 2001, pp. 26–27.

[15] E.T. Lee, J.W W. Statistical Methods for Survival Data Analysis, third ed., John Wiley & Sons, Canada, 2003.

[16] Mittla ankush, K. Ashraf, Bayesian Network Technologies: Applications and Graphical Models, IGI PuBlIshING, New York, 2007, pp. 103–106.

[17] R. Nagarajan, M. Scutari, S. Lèbre, Bayesian Networks in R with Applications in Systems Biology, 1 ed., Springer-Verlag New York, New York, 2013, pp. 1–4.

[18] M. Scutari, J.-B. Denis, Bayesian Networks: with Examples in R, 1 ed, Chapman and Hall, 2022, pp. 155–168.

[19] Balov Nikolay, Salzman Peter, Catnet: Categorical Bayesian Network Inference, 1.16.1, 2013.

[20] Kalisch Markus, Alain Hauser, Maechler Martin, Entner Doris, Hoyer Patrik, Hyttinen Antti, et al., Pcalg: Methods for Graphical Models and Causal Inference, 2.7-8 ed, 2022.

[21] H. Søren, gRbase: A Package for Graphical Modelling in R, 1.8.9 ed, 2022.

[22] J.B. Denis, Macro S. Rbmn: Handling Linear Gaussian Bayesian Networks, 0.9-5 ed, 2022.

[23] Scutari Macro, Silander Tomi, N. Robert, Bnlearn: bayesian network structure learning, parameter learning and inference, 4.8.1, 2022. Available from: https://cran.r-project.org/web/packages/bnlearn/index.html.

[24] S.G. Boettcher, C. Dethlefsen, Deal: a package for learning bayesian networks, J. Stat. Software 8 (20) (2003) 1–40.

[25] M.H. Kutner, C.J. Nachtsheim, J. Neter, W. Li, Applied Linear Statsitcal Models, McGrow -Hill Irwin, New York, 2005, pp. 268–289.

[26] S.G. Bottcher, C. Dethlefsen, Deal: A Package for Learning Bayesian Networks, 2003.

[27] D. Heckerman, D. Geiger, D.M. Chickering, Learning bayesian networks: the combination of knowledge and statistical data, Mach. Learn. 20 (3) (1995) 197–243.

[28] Y. Hou, Y. Peng, Z. Li, Update on prognostic and predictive biomarkers of breast cancer, Semin. Diagn. Pathol. 39 (5) (2022) 322–332.

[29] F. Fei, G.P. Siegal, S. Wei, Characterizing clinicopathologic features of estrogen receptor-positive/progesterone receptor-negative breast cancers, Clin. Breast Cancer 22 (7) (2022) e788–e797.

[30] J.-H. Kim, Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap.", Computational Statistics & Data Analysis. Comput. Stat. Data Anal. 53 (11) (2009) 3735–3745, 53:3735-45

[31] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009, p. 717.

[32] C. Siotos, A. Naska, R.J. Bello, A. Uzosike, P. Orfanos, D.M. Euhus, et al., Survival and disease recurrence rates among breast cancer patients following mastectomy with or without breast reconstruction, Plast. Reconstr. Surg. 144 (2) (2019), 169e-77e.

[33] J.L. Gomez Marti, A. Brufsky, A. Wells, X. Jiang, Machine learning to discern interactive clusters of risk factors for late recurrence of metastatic breast cancer, Cancers 14 (1) (2022).

[34] D. Courtney, M.G. Davey, B.M. Moloney, M.K. Barry, K. Sweeney, R.P. McLaughlin, et al., Breast cancer recurrence: factors impacting occurrence and survival, 1971, Ir. J. Med. Sci. 191 (6) (2022) 2501–2510.

[35] V. Sopik, S. Nofech-Mozes, P. Sun, S.A. Narod, The relationship between local recurrence and death in early-stage breast cancer, Breast Cancer Res. Treat. 155 (1) (2016) 175–185.

[36] H. Kawaguchi, Y. Yamamoto, S. Saji, N. Masuda, T. Nakayama, K. Aogi, et al., Factors associated with overall survival after recurrence in patients with ER-positive/HER2-negative postmenopausal breast cancer: an ad hoc analysis of the JBCRG-C06 Safari study, Jpn. J. Clin. Oncol. 52 (6) (2022) 545–553.

[37] Y. Lu, Y. Tong, X. Chen, K. Shen, Association of biomarker discrepancy and treatment decision, disease outcome in recurrent/metastatic breast cancer patients, Front. Oncol. 11 (2021).

[38] M.E. Akbari, M. Rohani Rasaf, N. Nafissi, L. Shojaee, The effect of patho-biological factors on the survival of recurrent breast cancer patients, Multidisciplinary Cancer Investigation 1 (0) (2017) 0.

[39] C.F. Hjorth, P. Damkier, B. Ejlertsen, T. Lash, H.T. Sørensen, D. Cronin-Fenton, Socioeconomic position and prognosis in premenopausal breast cancer: a population-based cohort study in Denmark, BMC Med. 19 (1) (2021) 235.