Original Article

Eur J Breast Health 2025; 21(3): 255-264
DOI: 10.4274/ejbh.galenos.2025.2025-2-14

Bayesian Model Prediction for Breast Cancer Survival: A Retrospective Analysis

Islam Bani Mohammad[1], Muayyad M. Ahmad[2]

[1]Department of Nursing, Al-Balqa Applied University, Faculty of Nursing, Al-Salt, Jordan
[2]Adult Health Nursing Clinical Nursing Department, School of Nursing, University of Jordan, Amman, Jordan

ABSTRACT

Objective: Over the recent years, machine learning (ML) models have been increasingly used in predicting breast cancer survival because of improvements in ML algorithms. However, cancer researchers still face a significant challenge in accurately predicting breast cancer patients' survival rates. The purpose was to predict breast cancer survival using a Bayesian network.

Materials and Methods: This retrospective study included 2,995 patients diagnosed with breast cancer and subsequently hospitalized between January 1, 2012, and December 30, 2024. SPSS Modeler version 18.0 was used to build prediction models. The data were randomly split into a training set (2,097 cases, 70%) and a test set (898 cases, 30%) for developing the Bayesian network model and predicting the overall survival of patients diagnosed with breast cancer. The model included demographic variables (age, marital status, and governorate), laboratory/clinical variables (hemoglobin level, white blood cell count, presence of hypertension, and diabetes mellitus) and the outcome variable, patient survival status (binary value: survived/died). The discriminative ability of models was evaluated by accuracy and the area under the curve (AUC) in terms of superior predictive performance for breast cancer outcomes.

Results: The Bayesian model exhibited the best discriminatory performance among the nine models, with an AUC of 0.859 and the highest accuracy of 96.661%. In the context of feature importance, white blood cell value at the time of diagnosis was the most important feature for predicting the survival of breast cancer. Patients who had below-normal hemoglobin and above-normal white blood count values had a higher death probability than patients who had normal white blood count and hemoglobin values. The presence of hypertension and diabetes mellitus in patients with breast cancer led to a reduced survival probability.

Conclusion: The Bayesian model outperformed the other models in predicting the survival probability of breast cancer. Routine laboratory testing and demographic data can be included in a ML model to predict breast cancer survival. Accurate prediction of breast cancer survival is vital for clinical decision-making.

Keywords: Bayesian model; breast cancer; machine learning; survival; prediction models

Cite this article as: Bani Mohammad I, Ahmad MM. Bayesian model prediction for breast cancer survival: a retrospective analysis. Eur J Breast Health. 2025; 21(3): 255-264

Key Points

- Breast cancer is widely acknowledged as a serious global health problem.
- The study developed a Bayesian network/machine learning model using demographic and clinical variables to predict breast cancer survival with good discrimination (area under the curve 0.859) and accuracy (96.7%).
- The important variables for survival prediction were white blood cell count, presence of diabetes, age, hemoglobin concentration, presence of hypertension, and the governorate of residence.

Introduction

Breast cancer is a common and serious health concern on a worldwide scale (1). Statistics show that it is the most commonly diagnosed cancer among women, with millions of cases recorded each year (2). In 2020, almost 2.3 million new instances of breast cancer were reported, with 685,000 deaths (1). Breast cancer is expected to cause more than 3 million new cases and 1 million deaths per year by 2040 (1). It follows that it is essential to address the broad effects of breast cancer and to reduce breast cancer-related death rates (3).

Survival is defined as the duration of time a patient survives after the disease is diagnosed. Breast cancer is a complex disease with varying survival rates across individuals, despite gradual improvements in recent years (4). Predicting breast cancer survival effectively may help healthcare providers make better decisions regarding medical

Corresponding Author:
Muayyad Ahmad, PhD; mma4@ju.edu.jo

Received: 27.02.2025
Accepted: 10.05.2025
Epub: 27.05.2025
Available Online Date: 20.06.2025

treatment intervention planning, preventing excessive therapy, and so develop the optimal clinical management (5). However, accurate survival prediction is also important for further research. The outcome variable of the present study was the prediction of survival at the time of data collection (6, 7).

Breast cancer is considered a multifactorial disease. Research has identified several etiological risk factors that will change a woman's likelihood of getting breast cancer, such as lifestyle, social-psychological factors, genetic factors, and environmental factors (8). Thus, the effective prediction of breast cancer risk should include these different factors, including demographics, such as age and gender (9). Age increases the likelihood of developing breast cancer. The most common age group for developing breast cancer are women aged between 50 and 69 years (10). Furthermore, women are more vulnerable than men to develop breast cancer, because of exposure to estrogen and progesterone (11).

Clinical biomarkers such as white blood count (WBC), and hemoglobin (Hb) concentration are important in breast cancer, as high WBC and low Hb concentrations are associated with breast cancer (12). The presence of comorbidities, such as hypertension (HTN) and diabetes mellitus (DM) are also associated with worse outcomes in breast cancer (13, 14). Petrelli et al. (13) reported that HTN is characterized was associated with an increase in the probability of death among patients with breast cancer. Furthermore, DM, which is also a globally prevalent disease, is positively associated with breast cancer (15). DM has been associated with a higher incidence and a lower survival rate in breast cancer (16). Therefore, the presence of HTN and DM should be included in models that are attempting to predict breast cancer survival.

Machine learning (ML) models can play a significant role in predicting breast cancer. ML has numerous benefits, including survival prediction, earlier detection, and enhanced model accuracy. Furthermore, ML models can examine numerous risk variables, including genetics, lifestyle, and medical history, resulting in individualized risk estimations for patients with breast cancer (9). A Bayesian model is a type of ML algorithm, which includes a more advanced and sophisticated approach for parameter adjustment. The algorithm creates a probabilistic model that maps parameter values to the objective and evaluates it using a validation set (17, 18). Using this model, the algorithm selects the most promising parameters to assess in the objective function. This method is more efficient than grid and random search, particularly in high-dimensional parameter fields (17, 18). A Bayesian model has the potential to learn data models automatically, without any implicit assumptions, and it is able to handle multiple and non-linear relationships between variables (18).

Investigating breast cancer survival rates is one method for identifying risk factors for mortality and will thus address a major public health issue. The aim of this study was to predict breast cancer survival using a Bayesian network ML algorithm. The findings of the present study could be used to raise public awareness of the factors that contribute to breast cancer deaths. Furthermore, the results may be shared with the Jordanian Ministry of Health to help policymakers enhance public awareness about the factors that increase the risk of breast cancer-related death, which may allow disease avoidance, in some cases, and earlier detection and more successful, appropriate treatment once detected in future cases of breast cancer in Jordan.

Materials and Methods

This retrospective study was approved at Jordan University Hospital Ethics Committee (IRB approval no: 10/2024/1503, date: 16.01.2024). The obtained health records were from between January 1, 2012, and December 30, 2024. The complexity of the organizational processes and the electronic health system necessitated a four-month extraction period. The inclusion criteria specified that the patients should be female, with breast cancer, and aged above 18 years. We excluded patients with any cancer other than breast cancer.

An earlier retrospective study attempted to predict breast cancer using machine-learning approaches by applying demographic, mammographic, and laboratory data. It found that the random forest model resulted in an accuracy of 80% and an area under the curve (AUC) of 0.56, while a gradient boosting trees model showed an AUC of 0.59, a stronger performance compared to the neural network (9).

The dataset for the present study contained 2,995 records and eight variables. The initial variables that were requested included the stage of breast cancer, age, Hb concentration and WBC at the time of diagnosis, governorate, marital status, family history of cancer, patient outcome (survival versus death), presence of HTN, and DM at the time of diagnosis. However, the data concerning family history, the stage, and grade of breast cancer were excluded because they were not available in the electronic health records.

Before the study initiated, the institutional ethics committee authorized the procedure. The study was conducted retrospectively, hence, informed consent was waived. The patients' information was managed in confidence. Each record was given an anonymized ID, allowing for the secure unidentifiable processing of patient data. The retrieved data were saved to a password-protected file on a secure computer in the researcher's office. The study data were only available to the researchers. There was no funding for this study.

Statistical Analysis

Data Preparation

To build the Bayesian model, the following steps were performed as data preparation: Checking for missing data, cleaning, and removing duplicated and inconsistent data. All laboratory data were standardized using one international unit for analysis. All data were originally stored on an Excel sheet and exported to SPSS. Then the files were merged into one data sheet by matching the cases with the ID numbers. Descriptive statistics were conducted using Statistical Package for Social Sciences, version 29.1 (19).

Preprocess of Missing Data

Since the data were created and collected in a real medical setting, there were several observations with missing features. We excluded features with too many missing values to lessen the influence of missing variables on the training process for prediction models, such as family history, smoking, and the stage and grade of breast cancer (Table 1).

Data Mining

To apply the data mining step, SPSS Modeler version 18.0 was used (20) to generate multiple predictive models based on the available data. The database records were split 70/30% for training the network and testing data, respectively. Training data are used to develop a predictive model, whereas the testing data are used to evaluate the model's performance (21). The primary criteria for selecting the most effective

Table 1. Proportion of missing data for initial variables selected for inclusion

Variable	Missing (%)
Age	0 (0.0%)
Family history	2.995 (100%)
Smoking	2.993 (100%)
WBC count	0 (0.0%)
Hb concentration	0 (0.0%)
Governorates	0 (0.0%)
Breast cancer stage	2.995 (100%)
Breast cancer grade	2.995 (100%)
Marital status	0 (0.0%)
DM	0 (0.0%)
HTN	0 (0.0%)

Hb: Hemoglobin; WBC: White blood cell; DM: Diabetes mellitus; HTN: Hypertension

AI model are the overall accuracy and AUC (22). The accuracy is the percentage of all the used datasets that are properly predicted out of all the instances (23). The AUC represents the performance metrics that determine the predictive ability of ML models and it measures the overall performance of the model (24). Furthermore, the AUC assesses a model's discriminative ability by comparing projected probabilities to actual binary survival status and estimating the probability of death for censored data at a particular point (25). The AUC ranges from 0 to 1, and a value of 0.5 is comparable to random guessing, and a value of 1 represents perfect discrimination (25). The best model was selected using an iterative approach to select the superior model for accurate breast cancer mortality prediction. The Bayesian network was the most effective of the nine models in our study and achieved highest overall accuracy score (96.661%) and a greater discriminatory measure, AUC score (0.859).

A Bayesian network model is a probabilistic graphical model that illustrates variables and their interactions using an acyclic graph with a directed structure (26). Gashu and Aguade (27) demonstrated that the Bayesian network model is especially valuable in medical applications, like predicting the survival time of breast cancer, because a Bayesian network can model intricate correlations between risk variables and symptoms while successfully incorporating uncertainty and prior information. Furthermore, this model has edges that reflect the conditional probability and nodes that represent random variables for the related factors (26). In addition, a Bayesian network was chosen as it was likely to achieve the highest performance, the likelihood of survival versus death is estimated using only the available variables, successfully resolving the difficulty of risk assessment with partial knowledge, and their conditional likelihood correlations (18).

A Bayesian network model uses a probabilistic framework to generate predictions and interpret outcomes in terms of probabilities and uses expert knowledge to determine the conditional independence of predictors. Furthermore, they provide an intuitive visual representation of the correlations between survival and mortality parameters (28). As the model is based on available variables from daily clinical practice, it can be used as a predictive tool for particular breast cancer patients and for assisting doctors in the decision-making process (27).

Results

A total of 2,995 patients diagnosed with breast cancer were included (Figure 1). Age groups were categorized as young adults (19–45 years, 24.6%) and older adults (46–99 years, 75.4%). Most of the patients were married (73.3%). Furthermore, laboratory assessments included Hb concentration and WBC count. Below-normal Hb and above-normal WBC levels were observed in 28.0% and 45.2% of patients, respectively. Approximately 15% of the patients had a recorded history of HTN and 19% of the patients had DM. Geographically, 94.4% were from Middle Governorates, followed by North (3.1%) and South Governorates (2.5%). During the 12-year follow-up period from the start of EHR data storage in 2012 through the end of 2024, 96.6% of patients belonged to the survived category. Comparison of the study sample based on survive (n = 2.892) versus dead (n = 103) are presented in Table 2. The comparison revealed statistically significant differences in survival based on marital status, Hb levels, WBC counts, HTN, and DM (p<0.01), with lower survival rates associated with being married, having below-normal Hb, high white blood cell counts, and having diabetes.

Figure 2 illustrates Kaplan-Meier survival plots, depicting hazard functions for key variables. Low Hb and high WBC levels, as well as the presence of HTN, and DM, were associated with an increased cumulative hazard compared to normal levels or the absence of these conditions, suggesting a poorer prognosis. Marital status also showed differences, with potentially distinct hazard functions depending on whether patients were single, married, or divorced.

Structure of the Study Model

Among the nine generated models, the Bayesian network was the most effective of the nine models in our study, achieving the highest overall

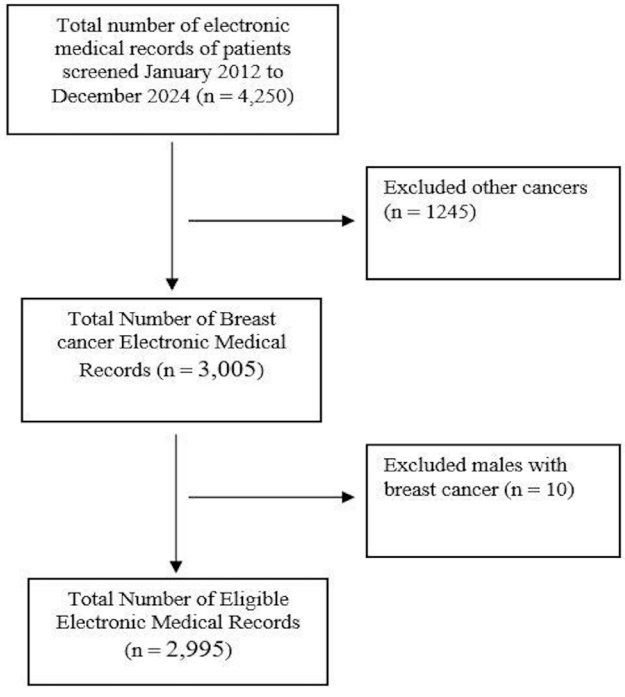

Figure 1. Patient flow chart

Table 2. Comparison of the study sample based on survival (*n* = 2.892) versus death (*n* = 103)

Characteristics	Dead n (%)	Survive n (%)	Chi-square
Age group			
Young adult (19–45)	24 (23.3)	712 (24.6)	0.093
Old adult (46–99)	79 (76.7)	2.180 (75.4)	
Marital status			
Single	7 (6.8)	362 (12.5)	
Married	72 (69.9)	2.123 (73.4)	8.694**
Divorced	24 (23.3)	407 (14.1)	
Hb			
Normal (12–15 g/dL for female)	45 (43.7)	2.110 (73.0)	42.225**
Below normal	58 (56.3)	782 (27.0)	
WBC			
Normal (WBC 4–11x10^3/microliter)	12 (11.7)	1.629 (56.3)	80.146**
Above normal (>11x10^3/microliter)	91 (88.3)	1.263 (43.7)	
HTN			
Yes	3 (2.9)	442 (15.3)	12.032**
No	100 (97.1)	2.450 (84.7)	
DM			
Yes	3 (2.9)	554 (19.2)	17.335**
No	100 (97.1)	2.338 (80.8)	
Governorates			
North	2 (1.9)	91 (3.1)	
Middle	99 (96.1)	2.729 (94.4)	0.618
South	2 (1.9)	74 (2.5)	

Hb: Hemoglobin; RBC: Red blood cell; WBC: White blood cell; DM: Diabetes mellitus; HTN: Hypertension; **p<0.01

accuracy score (96.661%) and the highest AUC score (0.859) (Table 3). This was followed by logistic regression (96.594%, AUC = 0.848), CHAID model (96.561, AUC = 0.826), neural network model (96.561, AUC = 0.688). Then, the C5 model, the Quest model, and the C&R Tree model had the same performance (96.561, AUC = 0.5). Followed by the decision list model (63.940, AUC= 0.788). The discriminant model (55.426, AUC = 0.532) had the lowest performance.

This study created a Bayesian network for binary classification to distinguish between patients who died and those who survived. The patient outcome node (survival versus death) in the graph represented a random variable, with 0 representing death and 1 representing survival. This encoding allowed for the discovery of probabilistic correlations between discrete variables, which made it easier to analyze linkages within the dataset. The Bayesian network model consisted of eight nodes, including the parent node. It comprised 13 edges, which indicate the factors that govern the interactions between these nodes (Figure 3). Every node in the network represents a random variable of interest. For the outcome (survival versus death) prediction of breast cancer, the predictor variables were age, marital status, governorate, Hb, WBC, HTN, and DM values. The parent node has direct edges that go to one or more child nodes (26). Directed edges between nodes reflect the probability link among the variables in the network (28).

Evaluation of Feature Importance

The Bayesian network found seven important predictors for survival outcome in breast cancer. The importance of WBC was 0.19, which was the most important predictor in our model, followed by DM and age predictors' importance (both 0.16), marital status (0.14), low Hb (0.13), presence of HTN (0.12), while the governorates predictor's importance was lowest at (0.10) (Figure 4).

Table 4 presents the conditional probabilities of survival versus death based on the Bayesian network model's analysis of key predictor variables. This table illustrates how the interplay of demographic factors (age, marital status, and governorate), laboratory/clinical variables (HB, WBC count, HTN and DM) influenced patient survival probabilities. For patients who had below-normal Hb and above normal WBC values, the conditional probability of death was 53%, while for patients who had normal WBC and Hb values, it was 0.17%. Survival probabilities are higher among individuals with normal WBC and Hb values (0.79%). Furthermore, survival probabilities among patients without DM and who had a normal WBC value (0.58%) were slightly higher than those of patients with an above normal WBC value and DM (0.51%). Survival probabilities among old patients who live in Middle Governorates (0.95%) were lower than among patients who live in South Governorates (0.02%).

The older adults who live in Middle Governorates had the highest probability of death (0.96%), while older adults residing in the South Governorates demonstrated the lowest recorded probability of death, at just 0.01%. Survival probabilities among married young adults (0.26%) were lower than that among single patients (0.40%). In addition, the survival probability among old-age patients who had HTN (0.17%) was lower than among young adults who did not have HTN (0.90%). However, the survival probabilities among young

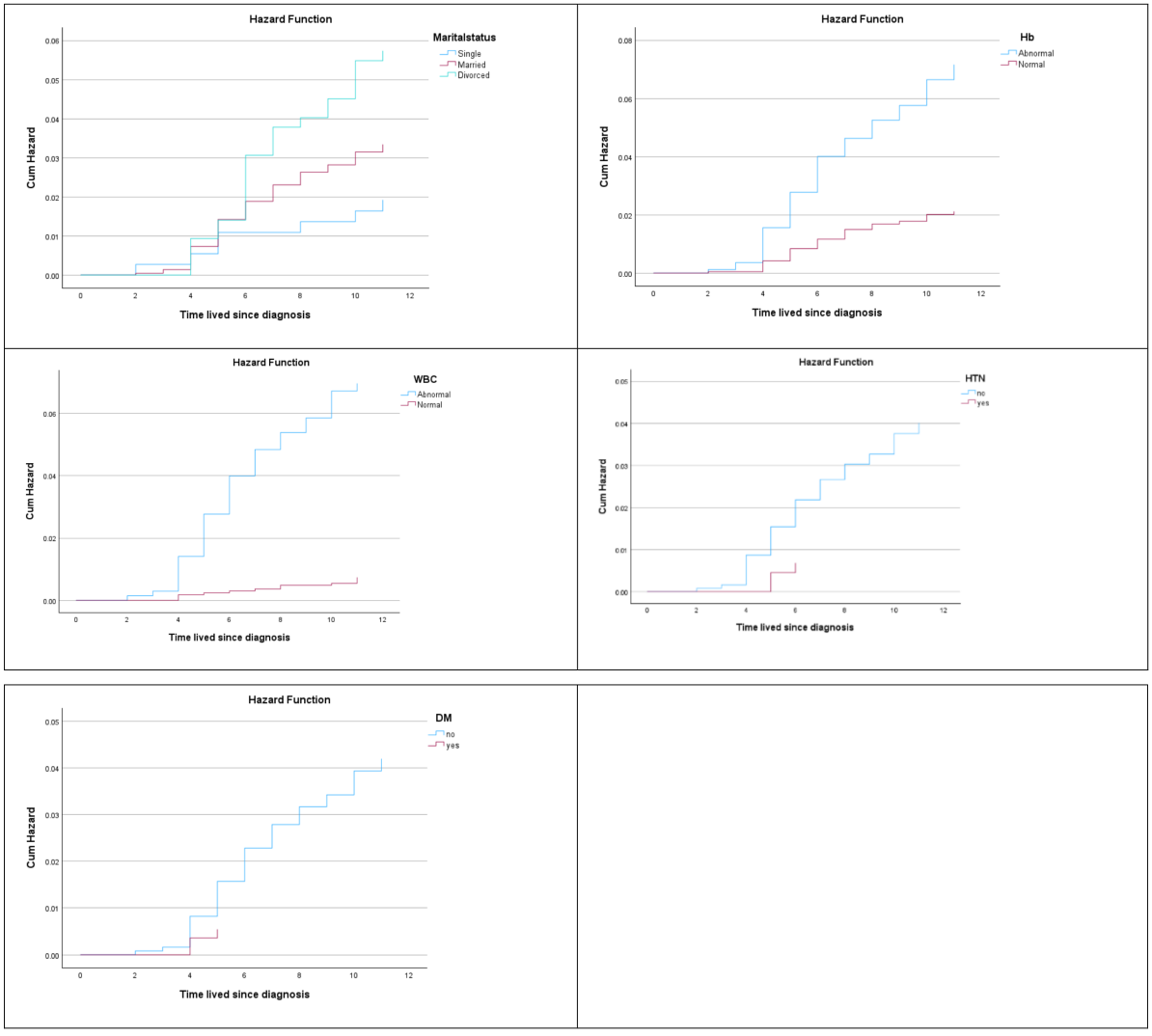

Figure 2. Kaplan-Meier survival plots and Hazard ratios

Table 3. The nine generated models in the study

Model	Overall accuracy (%)	Area under curve
Bayesian network	96.661	0.859
Logistic regression	96.594	0.848
CHAID	96.561	0.826
Neural net	96.561	0.688
C5	96.561	0.5
Quest	96.561	0.5
C&R tree	96.561	0.5
Decision list	63.940	0.788
Discriminant	55.426	0.532

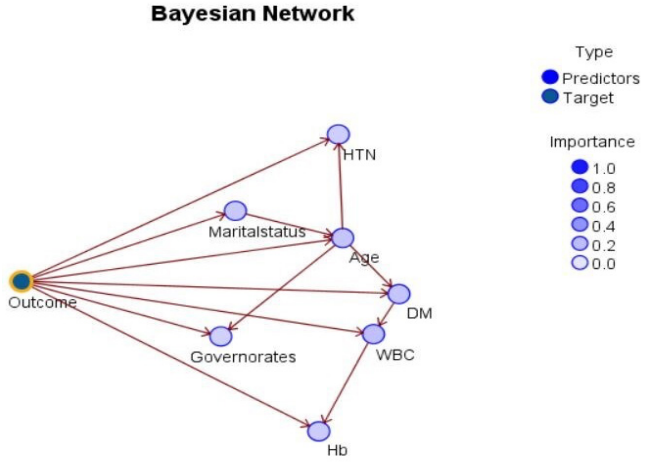

Figure 3. The structure of Bayesian network model

adult (0.24%) and older adult (0.17%) with DM had a lower survival probability compared to those young adults (0.76%) and older adult (0.83%) who did not have DM.

Discussion and Conclusion

This study investigated the connection between demographic characteristics, laboratory tests and presence of two key comorbidities, with the outcome (death or survival) via Bayesian network model in adult women diagnosed with breast cancer in Jordan over a 12-year period. Breast cancer incidence has increased over the past 30 years, while the death rate has decreased (9). The remarkably high survival rate in our 2012–2024 study cohort (96.6%) likely reflects a combination of factors: Increased breast cancer awareness leading to earlier detection through proactive screening programs; advancements in targeted therapies, chemotherapy, and surgical techniques; and limitations in data including exclusion of family history or cancer grade. The EHR data is a limitation, it does not account for socioeconomic variables. Further investigation into the relative contributions of these elements is important in improving future outcomes for the Jordanian population.

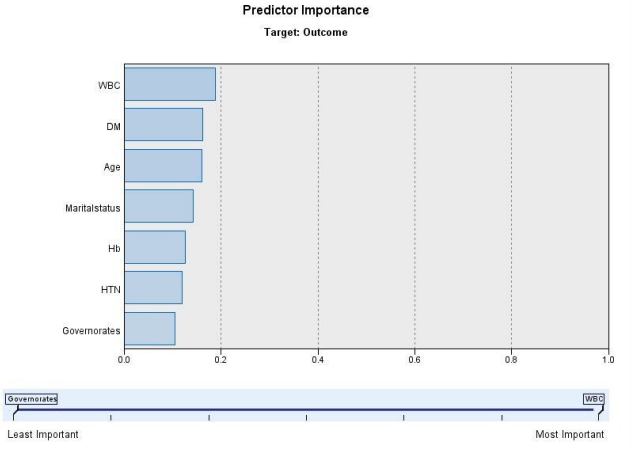

Figure 4. Predictors' importance for the outcome of breast cancer

Currently, ML is one of the most popular methods to create prediction models (29). It has been widely employed in medical science to assist healthcare providers with prognosis analysis. To analyze massive amounts of data, ML is important to create prediction models for predicting risk factors and can deal with real-world uncertainties and even missing data in training and test data sets. Using ML algorithms to analyze data can improve patient outcomes, specify needs, and improve quality of life (29, 30).

In this paper, a Bayesian network was used to predict survival versus death of adult female patients with breast cancer, based on a number of factors. Many researchers have assessed the usefulness of ML algorithms in predicting the risk of cancer, but few of them used a Bayesian network model to predict survival in breast cancer (31-36). The Bayesian network is a robust tool for predicting breast cancer survival. Its ability to integrate prior data makes it highly beneficial for medical decision support systems (34). Moreover, a Bayesian model is a type of probabilistic graphical model that predicts information about an uncertain area (27). In this paper, we predicted the survival of breast cancer using a Bayesian model. When comparing the performance of the Bayesian model to other models that are used for predicting the survival of breast cancer, the Bayesian model had the best performance. The results in this study were similar to those reported by some previous studies. For example, the Bayesian network model achieved the highest AUC value of 0.935 and a prediction accuracy of 87.2% for predicting breast cancer prognosis (32). Furthermore, previous research used the XGBoost method to predict breast cancer survival with a sample size of 4,575 patients (37). The results showed that the XGBoost model achieved a performance with an AUC of 0.8385. The possible reason for Bayesian model achieving better performance may be that Bayesian model is able to detect and account for higher-order interactions and non-linear relationships. However, the findings of this study provide insight into the efficacy of ML algorithms for predicting

Table 4. The Bayesian networks model's determination of the probabilities of survival versus death and predictors

Parents nodes	Conditional probability of HTN		Conditional probability of DM		Conditional probabilities of governorate			Conditional probabilities of marital status		
Dead or survive/age	**Yes**	**No**	**Yes**	**No**	**North**	**Middle**	**South**	**Single**	**Married**	**Divorced**
Survive/young adult	0.10	0.90	0.24	0.76	0.05	0.91	0.04	0.40	0.26	0.03
Death/young adult	0.04	0.96	0.04	0.96	0.00	0.96	0.04	0.00	0.33	0.00
Survive/old adult	0.17	0.83	0.17	0.83	0.03	0.95	0.02	0.60	0.74	0.97
Death/old adult	0.03	0.97	0.03	0.97	0.03	0.96	0.01	1.00	0.67	1.00

	Conditional probability of DM		Conditional probability of Hb	
Dead or survive/WBC	**Yes**	**No**	**Normal**	**Below normal**
Survive/normal	0.49	0.58	0.79	0.21
Death/normal	0.33	0.11	0.17	0.83
Survive/above normal	0.51	0.42	0.65	0.35
Death/above normal	0.67	0.89	0.47	0.53

Hb: Hemoglobin; RBC: Red blood cell; WBC: White blood cell; HTN: Hypertension; DM: Diabetes mellitus

survival probabilities among patients with breast cancer. The study's model illustrated the relationships between the outcome variable (survived or died) and the seven predictors. A Bayesian network model has several advantages over traditional survival models, including the elimination of the proportional hazard assumption, the imputation of missing data throughout the modeling process, and the ease with which results can be interpreted using graphical representations of variable interactions (31).

A growing number of studies using ML have been conducted on breast cancer diagnosis (5, 38). Furthermore, while the number of survival predictions grows gradually, the database set, modeling procedure, performance measures, methodological quality, and modeling of associated predictors vary significantly (5). Previous studies, that predicted breast cancer survivability using ML, identified predictors such as patient demographics, medical history, treatment information, and clinicopathological features of malignancies at various stages (36, 39-41).

Regarding factors influencing the survival of breast cancer, researchers have found many factors associated with breast cancer prognosis and survival. The most commonly used predictors are age, marital status, gender, laboratory tests, race, disease stage, grade, tumor size, number of nodes, histology, and primary site code, which have been entered into many predictive models as predictors (5, 31). Identifying the most significant predictors of survival in breast cancer can help healthcare providers in selecting effective treatment options and reducing data collection and treatment costs (5, 42). In the present study, there were seven important predictors of outcome in breast cancer, including WBC count, presence of DM, age, marital status, Hb value, presence of HTN, and governorate of residence. However, an earlier study found that the interpretation and identification of the important predictors was a key problem, and it was difficult to determine which variables had the greatest influence on survival (43).

Multiple studies have demonstrated that the age of patients with breast cancer is a significant factor in predicting their survival probability. In previous research, the age of patients has been considered a significant predictor for cancer among patients who have survived for more than 10 years (44). Moreover, researchers observed a significant relationship between age and the survival probability of patients experiencing cancer. In the present study, most of the patients were older adults. A previous study found that the age at the time of diagnosis of women with breast cancer was most commonly between 48 and 52 years old (45). However, Courtney and his colleagues only observed survival probability among patients aged 65 years and older (46). The evaluation of mortality among different ages that are vulnerable to breast cancer appears essential. In our study, patients' outcomes were predicted among young and older adults.

Laboratory tests can be used to help predict the survival and death probability of breast cancer (12). The most important predictor for determining the survival of breast cancer patients in the present study was WBC count. Below normal Hb and high WBC levels are considered as important predictors for low survival probability among patients with breast cancer. A recent study found that the overall mean difference for WBC between normal individuals and breast cancer patients was 8.554 (7.724) with a $p = 0.001$. Similarly, for Hb value in a breast cancer patient, the overall mean difference was 11.95 (12.19) compared to normal with a $p<0.05$ (12).

Other characteristics that have been investigated in patients include HTN or DM. Our results as shown in Table 4 indicated a lower survival probability among older patients with HTN. Several observational studies have established the relation between HTN in older women and breast cancer (13, 14). When analyzed according to cancer diagnosis, breast cancer was associated with increased mortality in patients with HTN (13, 47). Furthermore, the prevalence of HTN and breast cancer among women rises with age and could be caused by postmenopausal estrogen withdrawal (48, 49). In the present study, most of the patients had DM, while about 8.3% of patients with breast cancer had DM (50). Given that the existing DM with a breast cancer diagnosis was connected with decreased survival rates (51).

The marital status of the patients is fourth factor that influences patient survival and mortality rates. This study demonstrated that the marital status of patients exerts a notable impact on the survival outcomes of individuals. The survival probability was higher among young patients who were single than among those who were married. Conversely, the survival probability was slightly higher among older adults who were married than those who were unmarried. However, unmarried patients have been reported to have a worse overall survival (52). Other studies have also reported this association; married patients with breast cancer had a better survival rate than unmarried patients (53). Zhai and his colleague indicated that the mortality rate for unmarried patients was 24% higher than for married patients (54). The observed disparity in survival between married and unmarried patients may be influenced by the relatively small number of deceased patients in our sample. This limitation stems from the nature of the EHR used, which may not fully capture the range of survival outcomes within the studied population, nor allow for analysis of socioeconomic considerations which might influence these outcomes.

In the present study, the death probability was high in middle governorates (0.96), such as Amman, which is consistent with a Jordanian study showing that Amman, the capital of Jordan, had the highest incidence rates of breast cancer (45).

The Bayesian network provides a significant description of the correlations and effects of a number of variables on patient outcomes (55). Furthermore, the graphical presentation of Bayesian networks makes it easier to understand and communicate variable interactions than more sophisticated ML models. One of our study's strengths was that the Bayesian network can handle complex relationships efficiently, such as those having an effect in medical data. As leading tools in health informatics, ML has significant promise for use in normal healthcare. This study was especially unique in that it examined all patients with breast cancer, including young adults, rather than only the elderly. A Bayesian network model can overcome the issues of missing data in predicting patients' outcomes while retaining high accuracy in prediction. Furthermore, it was used in the study to analyze a big dataset.

While our study achieved robust performance using demographic and laboratory variables, we recognize that key prognostic factors, such as tumor stage, family history, and treatment details, were unavailable due to constraints in the electronic health records. These omissions may limit direct comparability to models incorporating full clinical staging data. However, our findings align with evidence that routine variables, such as WBC count and the presence of comorbidities at diagnosis are independently prognostic (12, 13), supporting their utility in settings where detailed pathological data are inaccessible.

Study Limitations

Our research has a few drawbacks. While it was intended to include critical characteristics relevant to predicting patient outcomes, such as breast cancer stage, grade, family history of cancer, and medical imaging, these were excluded due to their unavailability and high missing values in the electronic healthcare system. While the ideal dataset would include comprehensive data on tumor stage and grade, these variables were inconsistently documented within the available electronic health records in our study. Faced with this limitation, we focused on the most consistently available clinical and demographic variables to develop a predictive model based on real-world data, acknowledging that its performance is conditional on these constraints. Thus, the model's predictions are conditional on the available data and should be interpreted alongside standard clinical staging. In this study, a single 70–30 data split was used due to initial computational limitations, acknowledging this method's potential limitations compared to techniques like k-fold cross-validation. However, we mitigated bias through randomization and careful overfitting analysis, with plans to implement more robust validation methods in future research for improved model generalizability assessment. Furthermore, while our model does not replace comprehensive clinical staging, it demonstrates that readily available data can still offer valuable prognostic insights, particularly in settings with incomplete records.

Recommendations

Promoting awareness and global collaboration among medical professionals and researchers is essential in treating breast cancer. To fight breast cancer and reduce its impact on individuals and society globally, a comprehensive approach combining ML modeling of big data, research ideally including large inclusive prospective randomized trials, and accessible healthcare services is necessary. Future research should focus on finding additional risk factors, improving prediction approaches, and developing targeted treatment to reduce mortality associated with breast cancer. The level of anxiety and depression factors should be considered in the prediction. However, there is still tremendous room for improvement and development of ML modeling in breast cancer. Prospective research is recommended to verify the use of the Bayesian network in future research.

Implications for Practice

The Bayesian network can be used by healthcare providers to assess survival versus death probabilities and to guide hospital-based breast cancer treatment decisions, promoting tailored treatment options based on routine demographic and laboratory data. The Bayesian network identified the most influential determinants of breast cancer survival, including age, Hb concentration, WBC count at diagnosis, governorate of residence and the presence of important comorbidities, like HTN, and DM. This improved model interpretability and demonstrated its practical value. Furthermore, practice implications include using predictive models to deliver precise risk predictions, improve information systems, facilitate clinical decisions, enhance documentation, and estimate survival probabilities. Given Bayesian network model's simplicity and interpretability compared to other ML methods, the Bayesian network is becoming increasingly popular in healthcare and may be readily integrated into the practice of healthcare.

In summary, breast cancer remains a critical global health concern, affecting millions of people annually. This study has described the use of an ML approach for breast cancer survival prediction, highlighting various risk factors critical in survival prediction, using a Bayesian model. The Bayesian model outperformed the other ML models for discriminative ability, revealing the potential of the Bayesian method to be used as an effective approach to build prognostic prediction models in the context of survival analysis. Our future work will focus on additional predictors of the model using more complete data. Incorporating demographic data as well as routine laboratory tests improved the model's ability to predict survival outcomes, resulting in better clinical decision-making for breast cancer treatment.

Ethics

Ethics Committee Approval: This retrospective study was approved at Jordan University Hospital Ethics Committee (IRB approval no: 10/2024/1503, date: 16.01.2024).

Informed Consent: Retrospective study.

Footnotes

Authorship Contributions

Surgical and Medical Practices: I.B.M., M.M.A.; Concept: I.B.M., M.M.A.; Design: I.B.M., M.M.A.; Data Collection or Processing: I.B.M., M.M.A.; Analysis or Interpretation: I.B.M., M.M.A.; Literature Search: I.B.M., M.M.A.; Writing: I.B.M., M.M.A.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast. 2022; 66: 15-23. (PMID: 36084384) [Crossref]
2. Chhikara BS, Parang K. Global cancer statistics 2022: the trends projection analysis. Chem Biol Lett. 2023; 10: 451. [Crossref]
3. Kumari D, Naidu MVSS, Panda S, Christopher J. Predicting breast cancer recurrence using deep learning. Discover Applied Sciences. 2025; 7: 113. [Crossref]
4. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005; 34: 113-127. (PMID: 15894176) [Crossref]
5. Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. Predicting breast cancer 5-year survival using machine learning: a systematic review. PLoS One. 2021; 16: e0250370. (PMID: 33861809) [Crossref]
6. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak. 2019; 19: 48. (PMID: 30902088) [Crossref]
7. Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, et al. The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. JMIR Med Inform. 2022; 10: e33440. (PMID: 35179504) [Crossref]
8. Ahmad M, Bani Mohammad E, Tayyem E, Al Gamal E, Atout M. Pain and anxiety in patients with breast cancer treated with morphine versus tramal with virtual reality. Health Care Women Int. 2024; 45: 782-795. (PMID: 37703384) [Crossref]
9. Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of breast cancer using machine learning approaches. J Biomed Phys Eng. 2022; 12: 297-308. (PMID: 35698545) [Crossref]

10. Heer E, Ruan Y, Mealey N, Quan ML, Brenner DR. The incidence of breast cancer in Canada 1971-2015: trends in screening-eligible and young-onset age groups. Can J Public Health. 2020; 111: 787-793. (PMID: 32144720) [Crossref]
11. Obeagu EI, Obeagu GU. Breast cancer: a review of risk factors and diagnosis. Medicine (Baltimore). 2024; 103: e36905. (PMID: 38241592) [Crossref]
12. Botlagunta M, Botlagunta MD, Myneni MB, Lakshmi D, Nayyar A, Gullapalli JS, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Sci Rep. 2023; 13: 485. (PMID: 36627367) [Crossref]
13. Petrelli F, Ghidini A, Cabiddu M, Perego G, Lonati V, Ghidini M, et al. Effects of hypertension on cancer survival: a meta-analysis. Eur J Clin Invest. 2021; 51: e13493. (PMID: 33470426) [Crossref]
14. Pinder MC, Duan Z, Goodwin JS, Hortobagyi GN, Giordano SH. Congestive heart failure in older women treated with adjuvant anthracycline chemotherapy for breast cancer. J Clin Oncol. 2007; 25: 3808-3815. (PMID: 17664460) [Crossref]
15. Xiong F, Dai Q, Zhang S, Bent S, Tahir P, Van Blarigan EL, et al. Diabetes and incidence of breast cancer and its molecular subtypes: a systematic review and meta-analysis. Diabetes Metab Res Rev. 2024; 40: e3709. (PMID: 37545374) [Crossref]
16. Garczorz W, Kosowska A, Francuz T. Antidiabetic drugs in breast cancer patients. Cancers (Basel). 2024; 16: 299. (PMID: 38254789) [Crossref]
17. Ceylan Z. Diagnosis of breast cancer using improved machine learning algorithms based on bayesian optimization. IJISAE. 2020; 8: 121-30. [Crossref]
18. Alsabry A, Algabri M. Iterative tuning of tree-ensemble-based models' parameters using Bayesian optimization for breast cancer prediction. Informatics and Automation. 2024; 23: 129-168. [Crossref]
19. IBM. IBM SPSS Statistics for Windows. Armonk; 2023. [Crossref]
20. Giri A, Paul P. Applied marketing analytics using SPSS: modeler, statistics and amos graphics: PHI learning Pvt. Ltd.; 2020. [Crossref]
21. Witten I, Frank E, A Hall M, J Pal C. Data mining practical machine learning tools and techniques. Elsevier Inc.; 2017. [Crossref]
22. Lee KH, Choi GH, Yun J, Choi J, Goh MJ, Sinn DH, et al. Machine learning-based clinical decision support system for treatment recommendation and overall survival prediction of hepatocellular carcinoma: a multi-center study. NPJ Digit Med. 2024; 7: 2. (PMID: 38182886) [Crossref]
23. Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Informatics in Medicine Unlocked. 2019; 17: 100255. [Crossref]
24. Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric. J Classif. 2020; 37: 696-708. (PMID: 33250548) [Crossref]
25. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000; 56: 337-344. (PMID: 10877287) [Crossref]
26. Kim A, Lee D. Dynamic Bayesian network-based situational awareness and course of action decision-making support model. Expert Systems with Applications. 2024; 252: 124093. [Crossref]
27. Gashu C, Aguade AE. Assessing the survival time of women with breast cancer in Northwestern Ethiopia: using the Bayesian approach. BMC Womens Health. 2024; 24: 120. Erratum in: BMC Womens Health. 2024; 24: 162. (PMID: 38360619) [Crossref]
28. Ji J, Yang C, Liu J, Liu J, Yin B. A comparative study on swarm intelligence for structure learning of Bayesian networks. Soft Computing. 2017; 21: 6713-6738. [Crossref]
29. Mohammad EB, Ahmad M. A systematic evaluation of big data-driven colorectal cancer studies. Med Glas (Zenica). 2024; 21: 63-77. (PMID: 38341673) [Crossref]
30. Shamoun S, Ahmad M. Enhancing quality of life: the effect of complete decongestive therapy on Jordanian women with breast cancer after axillary lymph node dissection. Eur J Breast Health. 2025; 21: 122-131. (PMID: 40028896) [Crossref]
31. Teng J, Zhang H, Liu W, Shu XO, Ye F. A dynamic Bayesian model for breast cancer survival prediction. IEEE J Biomed Health Inform. 2022; 26: 5716-5727. (PMID: 36040947) [Crossref]
32. Choi JP, Han TH, Park RW. A hybrid Bayesian network model for predicting breast cancer prognosis. Journal of Korean Society of Medical Informatics. 2009; 15: 49-57. [Crossref]
33. Endo A, Shibata T, Tanaka H. Comparison of seven algorithms to predict breast cancer survival (< special issue> contribution to 21 century intelligent technologies and bioinformatics). International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association. 2008; 13: 11-16. [Crossref]
34. Thongkam J, Xu G, Zhang Y, Huang F, editors. Support vector machine for outlier detection in breast cancer survivability prediction. Advanced Web and Network Technologies, and Applications: APWeb 2008 International Workshops: BIDM, IWHDM, and DeWeb Shenyang, China, April 26-28, 2008 Revised Selected Papers 10; 2008: Springer. [Crossref]
35. Lotfnezhad Afshar H, Ahmadi M, Roudbari M, Sadoughi F. Prediction of breast cancer survival through knowledge discovery in databases. Glob J Health Sci. 2015; 7: 392-398. (PMID: 25946945) [Crossref]
36. Shouket T, Mahmood S, Hassan MT, Iftikhar A, editors. Overall and disease-free survival prediction of postoperative breast cancer patients using machine learning techniques. 2019 22nd International Multitopic Conference (INMIC); 2019: IEEE. [Crossref]
37. Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Trans Biomed Eng. 2021; 68: 148-160. (PMID: 32406821) [Crossref]
38. Crowley RJ, Tan YJ, Ioannidis JPA. Empirical assessment of bias in machine learning diagnostic test accuracy studies. J Am Med Inform Assoc. 2020; 27: 1092-1101. (PMID: 32548642) [Crossref]
39. Salehi M, Razmara J, Lotfi S. A novel data mining on breast cancer survivability using MLP ensemble learners. The Computer Journal. 2020; 63: 435-447. [Crossref]
40. Simsek S, Kursuncu U, Kibis E, AnisAbdellatif M, Dag A. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. Expert Systems with Applications. 2020; 139: 112863. [Crossref]
41. Shamoun S, Ahmad M. Literature review on breast cancer-related lymphedema and related factors. Arch Oncol. 2023; 29: 22-27. [Crossref]
42. Shamoun S, Ahmad M. Complete decongestive therapy effect on breast cancer related to lymphedema: a systemic review and meta-analysis of randomized controlled trials. Asian Pac J Cancer Prev. 2023; 24: 2225-2238. (PMID: 37505751) [Crossref]
43. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018; 51: 1-42. [Crossref]
44. Gomez Marti JL, Brufsky A, Wells A, Jiang X. Machine learning to discern interactive clusters of risk factors for late recurrence of metastatic breast cancer. Cancers (Basel). 2022; 14: 253. (PMID: 35008417) [Crossref]
45. Khatib S, Nimri O. General oncology care in Jordan. Cancer in the arab world: springer Singapore. Singapore; 2022. p. 83-98. [Crossref]

46. Courtney D, Davey MG, Moloney BM, Barry MK, Sweeney K, McLaughlin RP, et al. Breast cancer recurrence: factors impacting occurrence and survival. Ir J Med Sci. 2022; 191: 2501-2510. (PMID: 35076871) [Crossref]

47. Fan Y, Khan NH, Farhan Ali Khan M, Ahammad MDF, Zulfiqar T, Virk R, et al. Association of hypertension and breast cancer: antihypertensive drugs as an effective adjunctive in breast cancer therapy. Cancer Manag Res. 2022; 14: 1323-1329. (PMID: 35392356) [Crossref]

48. Potmešil P, Szotkowská R. Drug-induced liver injury after switching from tamoxifen to anastrozole in a patient with a history of breast cancer being treated for hypertension and diabetes. Ther Adv Chronic Dis. 2020; 11: 2040622320964152. Erratum in: Ther Adv Chronic Dis. 2021; 12: 20406223211002790. (PMID: 33240477) [Crossref]

49. Zhao Y, Wang Q, Zhao X, Meng H, Yu J. Effect of antihypertensive drugs on breast cancer risk in female hypertensive patients: evidence from observational studies. Clin Exp Hypertens. 2018; 40: 22-27. (PMID: 29115847) [Crossref]

50. Zhang Y, Wu J, Chen W, Liang X. Pretreatment system inflammation response index (SIRI) is a valuable marker for evaluating the efficacy of neoadjuvant therapy in breast cancer patients. Int J Gen Med. 2024; 17: 4359-4368. (PMID: 39346633) [Crossref]

51. Harborg S, Kjærgaard KA, Thomsen RW, Borgquist S, Cronin-Fenton D, Hjorth CF. New horizons: epidemiology of obesity, diabetes mellitus, and cancer prognosis. J Clin Endocrinol Metab. 2024; 109: 924-935. (PMID: 37552777) [Crossref]

52. Krajc K, Miroševič Š, Sajovic J, Klemenc Ketiš Z, Spiegel D, Drevenšek G, et al. Marital status and survival in cancer patients: a systematic review and meta-analysis. Cancer Med. 2023; 12: 1685-1708. (PMID: 35789072) [Crossref]

53. Dag AZ, Akcam Z, Kibis E, Simsek S, Delen D. A probabilistic data analytics methodology based on Bayesian belief network for predicting and understanding breast cancer survival. Knowledge-Based Systems. 2022; 242: 108407. [Crossref]

54. Zhai Z, Zhang F, Zheng Y, Zhou L, Tian T, Lin S, et al. Effects of marital status on breast cancer survival by age, race, and hormone receptor status: a population-based Study. Cancer Med. 2019; 8: 4906-4917. (PMID: 31267686) [Crossref]

55. Zhong L, Yang F, Sun S, Wang L, Yu H, Nie X, et al. Predicting lung cancer survival prognosis based on the conditional survival bayesian network. BMC Med Res Methodol. 2024; 24: 16. (PMID: 38254038) [Crossref]