Modeling the Factors Associated with Mortality in Patients with Breast Cancer: A Machine Learning Approach

Mohammad Asghari Jafarabadi
Tabriz University of Medical Sciences
Zeynab Iraji
Tabriz University of Medical Sciences https://orcid.org/0000-0002-3844-6599
Roya Dolatkhah
Tabriz University of Medical Sciences
Tohid Jafari Koshki (tjkoshki@gmail.com)
Tabriz University of Medical Sciences https://orcid.org/0000-0002-6928-1387

Research article

Keywords: Machine learning; IPCW; Breast Cancer; Survival; GAM

Posted Date: August 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-57685/v1

Abstract

Background: Breast cancer (BC) was the fifth leading cause of death worldwide in 2015 and the second leading cause of death in Iran in 2012. This study aimed to model the factors associated with mortality in patients with BC utilizing the machine learning approach.

Methods: We used data of patients with primary BC during 2007-2016 in Tabriz, Iran. The data were analyzed using decision tree (DT), boosted tree (BT), random forest (RF), k-nearest neighbors (KNN) and generalized additive model (GAM) with inverse probability of censoring weighting (IPCW) technique to assess the risk factors of mortality. The models were compared by using diagnostic accuracy measures.

Results: Accuracy of the models ranged from 76.0 to 93.0%, with sensitivity of 82.5-98.8% and specificity of 72.2-99.4%. The GAM fit the data best with accuracy of 93.0% (95% CI: [90.5, 95.0]), sensitivity of 98.8% (95% CI: [96.9, 99.7]) and specificity of 84.3% (95% CI: [78.8, 88.9]) where non-linear effect of age (p-value = 0.006), grade (p-value = 0.024) and time to event (p-value < 0.001) on mortality were significant.

Conclusion: The GAM seems to be an optimal model for classifying the mortality in patients with BC. Considering the time to event, age and grade, as the prognostic factors obtained by GAM, more accurate prevention planning may be designed.

Background

Breast cancer (BC) is a prevalent cancer among women and the common cause of mortality from cancers.[1] The prevalence of BC is highest in the USA and Western Europe and lowest in East Asia.[1] Although the prevalence and mortality rates of BC are decreasing in many countries, the global prevalence is increasing with annual increase of 12% in age standardized incidence rates.[1–4] The mortality and prevalence of BC are increasing in Iran and in East Azerbaijan province as well. Although, the BC incidence rate in East Azerbaijan is lower than Western Europe and the USA, the increment in BC incidence rates needs further study.[5,6]

The BC incidence and survival are affected by different factors such as environment, hormones, nutrition, heredity, number of pregnancies, tumor size, morphology, age, and tumor grade and stage.[7–10] Identification of main risk factors and high risk groups could help proper planning to reduce the BC incidence and mortality.

Various analyzes are used for investigating the relationships between the survival of patients with BC and its risk factors, like log-rank test, Kaplan-Meier approach, cox regression model, and parametric survival models.[7–10] These techniques need certain assumptions (like proportional hazard and linear relationship) to be satisfied, where in many practical situations may not be the case. According to the importance of more accurate risk prediction, different machine learning techniques are expanded for predicting with

approaches that can be used with these

techniques to provide more accurate predictions. The IPCW is an approach that reduces the role of censored data and bias of risk estimation by giving less weight to censored participants.[11] There are many machine learning techniques that are used for prediction and classification [12–17] but it is necessary to determine which technique has more precision and accuracy. According to an extensive research in the literature, there was no study to assess the machine learning approach for classifying the mortality in patients with BC. In this study, we used several machine learning methods with IPCW approach including decision tree (DT), boosted tree (BT), random forest (RF), k-nearest neighbors (KNN) and generalized additive model (GAM) to the best model the relationships of mortality with common prognostic factors in patients with BC.

Material And Methods

Study population

This study included the data of 1154 patients diagnosed with primary BC (C50.0 in the International Classification of Oncology Diseases-9) from March 2007 to March 2016 in population-based cancer registry in East Azerbaijan.[18]

Data collection

All clinic-pathologic data including age, morphology (containing ductal carcinoma [DC]; lobular carcinoma [LC]; and other types), sex, and Grade (I–IV) were recorded regularly using 11-digit personal identification number. Other information about outcome and survival time was achieved by referring to the information system in hospital and contacting patients or their relatives. We considered the survival time as the time between the diagnosis date and the death from BC date.

Statistical analyses

Data were expressed as mean (SD) and frequency (percentages) for numeric and categorical variables, respectively. Mortality from BC, the primary outcome of this study, was equal to 1 where the patient had experienced death, otherwise it was zero.

For calculating IPCWs, the value of τ should have been defined.[11] The IPCWs were computed for censored event time at $\tau = 9$ years. The dataset was split randomly into training and validation dataset, and five machine learning techniques including DT (using CART algorithm), BT, RF, KNN and GAM were fitted in them. These techniques were used for classifying the patients with BC using sex, age, grade, morphology and survival time. The fitted models were compared using diagnostic criteria containing accuracy, sensitivity, specificity, ROC area, likelihood ratio +, likelihood ratio -, odds ratio, positive predicted value, and negative predicted value. All analyses were performed using STATISTICA 12 (Statistica, StatSoft, Texas, USA).

Results

The mean age of participants was 50.4 (SD 12.5) years and 1132 individuals were female (98.1%). The details of demographic and background characteristics of the study participants were provided elsewhere.[8]

Table 1 shows diagnostic statistics for comparing DT, BT, RF, KNN and GAM. The results indicate that most of the diagnostic statistics values are approximately close to each other in the investigated models, however GAM has the highest sensitivity, accuracy and negative predicted value.

Table 1
Diagnostic statistics for DT, BT, RF, KNN and GAM in patients with BC

	CART	Boosted tree	Random forest	GAM	KNN
Sensitivity	82.5 (76.8,87.3)	87.9 (81.4,92.8)	84.8 (78.5,89.9)	98.8 (96.9,99.7)	87.7 (78.5,93.9)
Specificity	99.4 (97.8,99.9)	95.9 (92.4,98.1)	98.6 (96.1,99.7)	84.3 (78.8,88.9)	72.2 (66.4,77.5)
Accuracy	92.6 (90.1,94.7)	92.8 (89.6,95.2)	92.8 (89.7,95.1)	93.0 (90.5,95.0)	76.0 (71.0,80.0)
Roc area	0.909 (0.884,0.935)	0.919 (0.889,0.949)	0.917 (0.889,0.946)	0.915 (0.891,0.940)	0.800 (0.750,0.840)
Likelihood ratio +	134 (33.5,533.0)	21.6 (11.4,41.1)	62.8 (20.4,194.0)	6.3 (4.6,8.6)	3.15 (2.5,3.9)
Likelihood ratio -	0.176 (0.132,0.235)	0.126 (0.080,0.196)	0.154 (0.107,0.220)	0.015 (0.005,0.039)	0.170 (0.090,0.310)
Odds ratio	758 (198,NE)	172 (74.9,393)	409 (127,1297)	431 (156,1183)	18.40 (9.1,37.2)
Positive predicted value	98.9 (96.1,99.9)	93.2 (87.5,96.9)	97.9 (94.0,99.6)	90.4 (86.8,93.3)	49.00 (40.6,57.4)
Negative predicted value	89.4 (85.8,92.4)	92.6 (88.4,95.6)	89.8 (85.2,93.3)	97.9 (94.6,99.4)	95.00 (91.1,97.6)

The results of DT showed that the most important cause of mortality from BC was the survival time. Other independent variables were in the very low importance level (Fig. 1).

Figure 1. Importance of the independent variables in relation to mortality from breast cancer using CART

The results of BT indicated that the survival time was most important predictor for mortality from BC. [illegible] of importance (Fig. 2).

Figure 2. Importance of the independent variables in relation to mortality from breast cancer using BT

The results of RF showed that the survival time was in the high level of importance for mortality from BC whereas other independent variables were in the low level of importance (Fig. 3).

Figure 3. Importance of the independent variables in relation to mortality from breast cancer using RF

Figure 4 show cross validation accuracy against number of nearest neighbors in KNN method. According to this figure, k = 34 chosen as the optimal number of nearest neighbors for predicting mortality status in patients with BC.

Figure 4. Cross validation accuracy against number of nearest neighbors

The results of GAM are presented in Table 2. The results indicated that age, grade and survival time have significant nonlinear effect on mortality from BC.

Table 2
GAM result using τ = 9 in patients with BC

Variable	GAM coef.	Degree of freedom	Non-linear p-value
Age	-0.019	4.276	0.006*
Sex	Ref	1	-
Male	-2.045	1	
Female			
Grade	-0.375	3.004	0.024*
Morphology	Ref	1	-
DC	-0.675	1	
LC	0.923	1	
Other			
Time	0.107	4.032	< 0.001*
Note: *p < 0.05			

Discussion

In this study, some machine learning techniques were used and compared to classify the patients with BC. In this regard, five machine learning methods including DT, BT, RF, KNN and GAM were applied in the

prediction of mortality status.

The IPCW approach was used to reduce the role of censored data in classification of mortality in patients with BC. The IPCW is an approach that reduces the bias of risk estimation by giving less weight to censored participants. This approach can be used with many machine learning methods.[11]

Interestingly, in our data, the GAM using IPCW approach had the best performance which had the highest accuracy, sensitivity and negative predictive value among the investigated machine learning methods. So we can use this model to achieve the accurate classifications. In the line with our finding, GAM has outperformed other machine learning techniques in various fields such as classification of functional setting, detection of diplodia sapinea (which inflict severe damage upon pine trees) and forecast of postoperative complications.[19–21] While Garosi et al and Goetz et al concluded that random forest method is a better one.[22,23]

Using the GAM, time to event, age and grade had significant nonlinear effect on the survival of patients with BC so that patients with lower time to event, higher age and higher grade had more mortality. Similarly, previous studies have revealed that the survival of patients with BC is directly related by time and inverse related by age and grade.[8,9,12,24–28]

Limitations and strengths

Clearly, this study has some limitations. The GAM does not provide statistical significance for each of the predictive variables, however examination of nonlinear relationships compensates for this problem. Some important clinical information like tumor size, progesterone receptor, estrogen receptor, and tumor-node-metastasis stage were not available to consider in the analyses. Furthermore, in Iran, the National Cancer Registry of Iran and EA-PBCR have not yet expanded. To avoid missing data; all of the patients which diagnosed with primary BC were registered from all over the province and data was collected via a combined active and passive protocol of follow-up in the EA-PBCR. However as an advantage, non-linear relationships between predictive variables and response variable are investigated in GAM.

Conclusion

Interestingly, the GAM had the highest accuracy and precision in classification of patients with BC. Using this model we investigated the non-linear relationships between the mortality status and its predictors. Considering the time to event, age and grade, as the prognostic factors obtained by GAM, more accurate prevention planning may be designed.

List Of Abbreviations

BC: Breast cancer

DT: Decision tree

BT: Boosted tree

RF: Random forest

KNN: K-nearest neighbors

GAM: Generalized additive model

IPCW: Inverse probability of censoring weighting

DC: Ductal carcinoma

LC: Lobular carcinoma

Declarations

Ethics approval and consent to participate

The study protocol was approved by the institutional review board of Tabriz University of Medical Sciences (IRB no.: IR.TBZMED.REC.1397.986).

Availability of data and materials

The data that support the findings of this study are available from MAJ but restrictions are applied to the availability of these data, which were used under license for the current study, and are not publicly available. Data are, however, available from the authors upon reasonable request by MAJ.

Consent for publication

All authors have given consent for publication.

Conflict of interest statement

This research has been conducted in the absence of any potential conflict of interest, in the other words in the absence of any commercial or financial relationships.

Funding

No funding to declare.

Author contributions

Conception and design: M. Asghari jafarabadi, Z. Iraji

raji

Acquisition of data: R. Dolatkhah

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Asghari jafarabadi, Z. Iraji

Writing, review, and/or revision of the manuscript: M. Asghari jafarabadi, T. Jafari koshki, Z. Iraji

Study supervision: M. Asghari jafarabadi, Z. Iraji

Acknowledgment

Data of patients diagnosed with primary Breast cancer and registered in the East Azerbaijan population-based cancer registry were included in our analysis. As the ethics rules of East Azerbaijan population-based cancer registry, all patients' information, and records are confidential. The study protocol was approved by the Institutional Review Board of Tabriz University of Medical Sciences (IRB no.: IR.TBZMED.REC.1397.986).

References

1. Fitzmaurice C, Allen C, Barber RM, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA oncology. 2017;3(4):524-548.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.
3. Fitzmaurice C, Akinyemiju TF, Al Lami FH, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. JAMA oncology. 2018;4(11):1553-1568.
4. Somi MH, Dolatkhah R, Sepahi S, et al. Cancer incidence in the East Azerbaijan province of Iran in 2015–2016: results of a population-based cancer registry. BMC public health. 2018;18(1):1266.
5. Jafari-Koshki T, Schmid VJ, Mahaki B. Trends of breast cancer incidence in Iran during 2004-2008: a Bayesian space-time model. Asian Pac J Cancer Prev. 2014;15(4):1557-1561.
6. Sharifian A, Pourhoseingholi MA, Emadedin M, et al. Burden of breast cancer in Iranian women is increasing. Asian Pac J Cancer Prev. 2015;16(12):5049-5052.
7. Carlo JT, Grant MD, Knox SM, et al. Survival analysis following sentinel lymph node biopsy: a validation trial demonstrating its accuracy in staging early breast cancer. Paper presented at: Baylor University Medical Center Proceedings2005.
8. Iraji Z, Koshki TJ, Dolatkhah R, Jafarabadi MA. Parametric survival model to identify the predictors of pproach. Journal of Research in Medical

Sciences. 2020;25(1):38.

9. Møller H, Henson K, Lüchtenborg M, et al. Short-term breast cancer survival in relation to ethnicity, stage, grade and receptor status: national cohort study in England. British journal of cancer. 2016;115(11):1408.
10. Yang MT, Rong T-H, Huang ZF, et al. Clinical analysis of resectable breast cancer: a report of 6 263 cases. Ai zheng= Aizheng= Chinese journal of cancer. 2005;24(3):327-331.
11. Vock DM, Wolfson J, Bandyopadhyay S, et al. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. Journal of biomedical informatics. 2016;61:119-131.
12. Iraji Z, Jafari Koshki T, Dolatkhah R, Asghari Jafarabadi M. A conditional probability model to predict the mortality in patients with breast cancer: a Bayesian network analysis. In press 2020.
13. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbors. Paper presented at: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)2017.
14. Khademi M, Nedialkov NS. Probabilistic graphical models and deep belief networks for prognosis of breast cancer. Paper presented at: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)2015.
15. Sivakami K, Saraswathi N. Mining big data: breast cancer prediction using DT-SVM hybrid model. International Journal of Scientific Engineering and Applied Science (IJSEAS). 2015;1(5):418-429.
16. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology. 2015;8(29):1-8.
17. Pawlovsky AP, Matsuhashi H. The use of a novel genetic algorithm in component selection for a kNN method for breast cancer prognosis. Paper presented at: 2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)2017.
18. International Classification of Diseases for Oncology. In: Fritz A, Percy C, KJack A, al. E, eds. 3rd Edition ed.: WHO Library Cataloguing-in-Publication Data; 2013. Accessed 2016/05/04.
19. Cuesta-Albertos JA, Febrero-Bande M, de la Fuente MO. The

$$\mathrm{DD}^G$$

-classifier in the functional setting. Test. 2017;26(1):119-142.
20. Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A. Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. arXiv preprint arXiv:180311266. 2018.
21. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PloS one. 2016;11(5).
22. Goetz J, Brenning A, Petschko H, Leopold P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & geosciences. 2015;81:1-11.

23. Garosi Y, Sheklabadi M, Conoscenti C, Pourghasemi HR, Van Oost K. Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. Science of the Total Environment. 2019;664:1117-1132.
24. Afshar HL, Ahmadi M, Roudbari M, Sadoughi F. Prediction of breast cancer survival through knowledge discovery in databases. Global journal of health science. 2015;7(4):392.
25. Elsheikh SE, Green AR, Rakha EA, et al. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. Cancer research. 2009;69(9):3802-3809.
26. Fallahzadeh H, Momayyezi M, Akhundzardeini R, Zarezardeini S. Five year survival of women with breast cancer in Yazd. Asian Pac J Cancer Prev. 2014;15(16):6597-6601.
27. Fouladi N, Amani F, SHarghi A, Nayebyazdi N. Five year survival of women with breast cancer in Ardabil, north-west of Iran. Asian Pacific journal of cancer prevention: APJCP. 2011;12(7):1799-1801.
28. Haghighat S. Survival rate and its correlated factors in breast cancer patients referred to Breast Cancer Research Center. Iranian Quarterly Journal of Breast Disease. 2013;6(3):28-36.

Figures

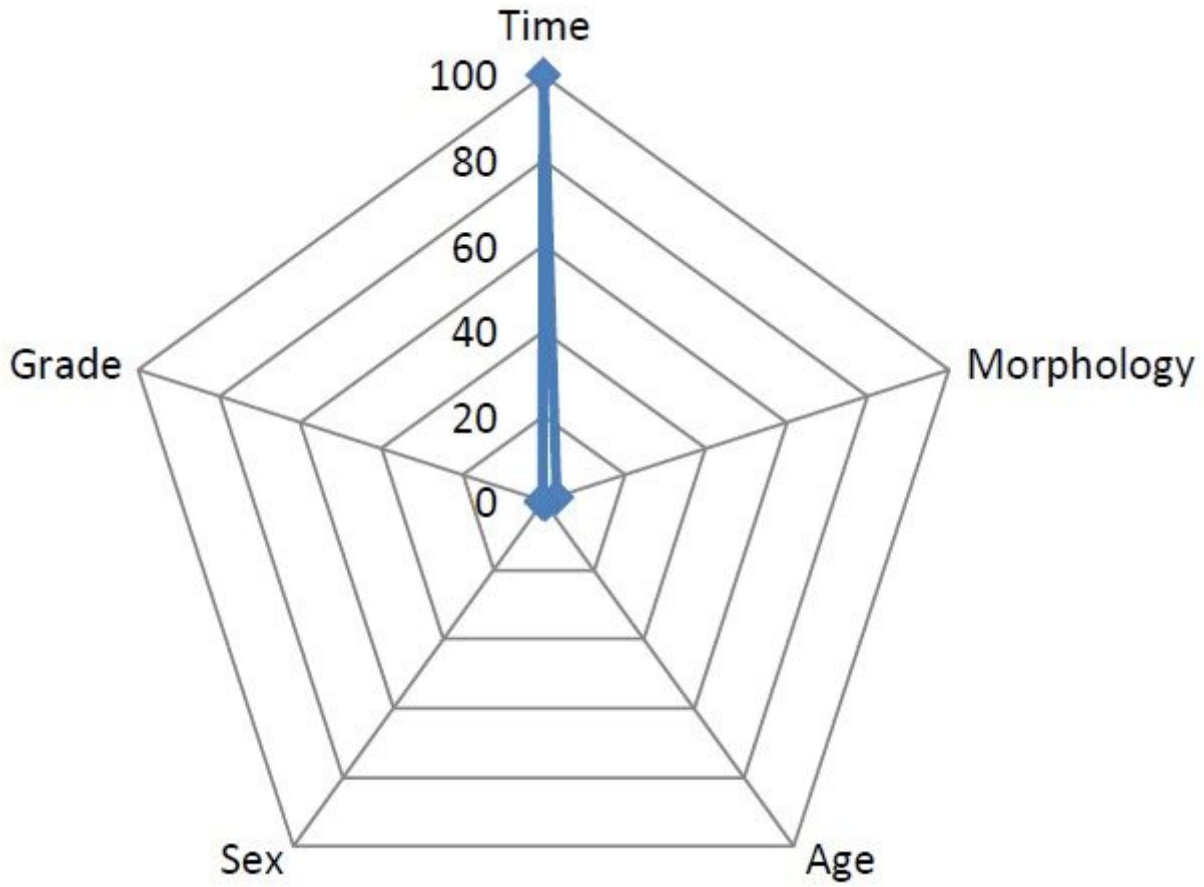

Figure 1

rtality from breast cancer using CART

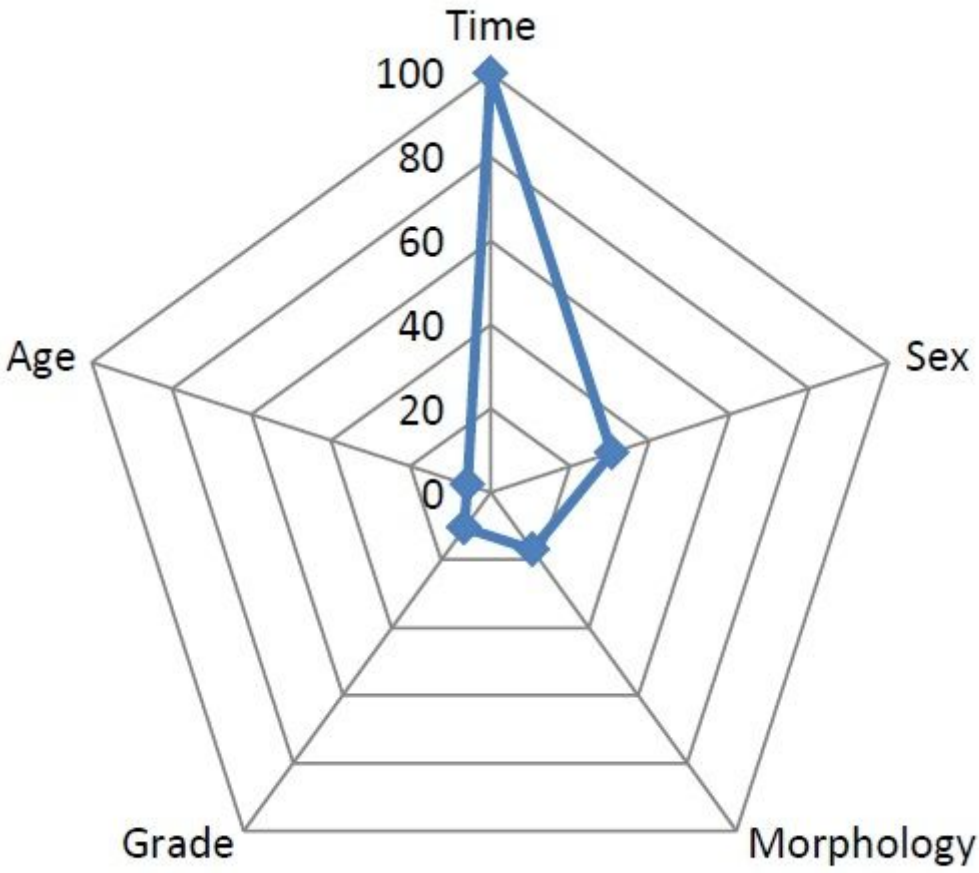

Figure 2

Importance of the independent variables in relation to mortality from breast cancer using BT

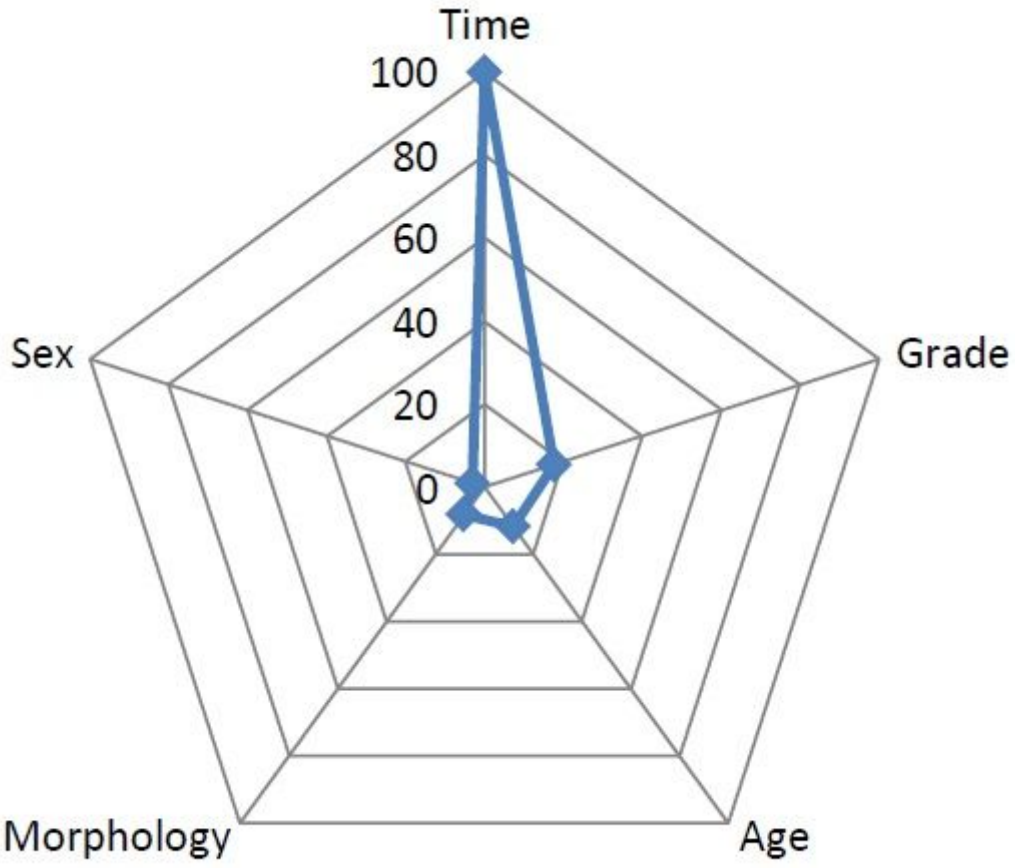

Figure 3

Importance of the independent variables in relation to mortality from breast cancer using RF

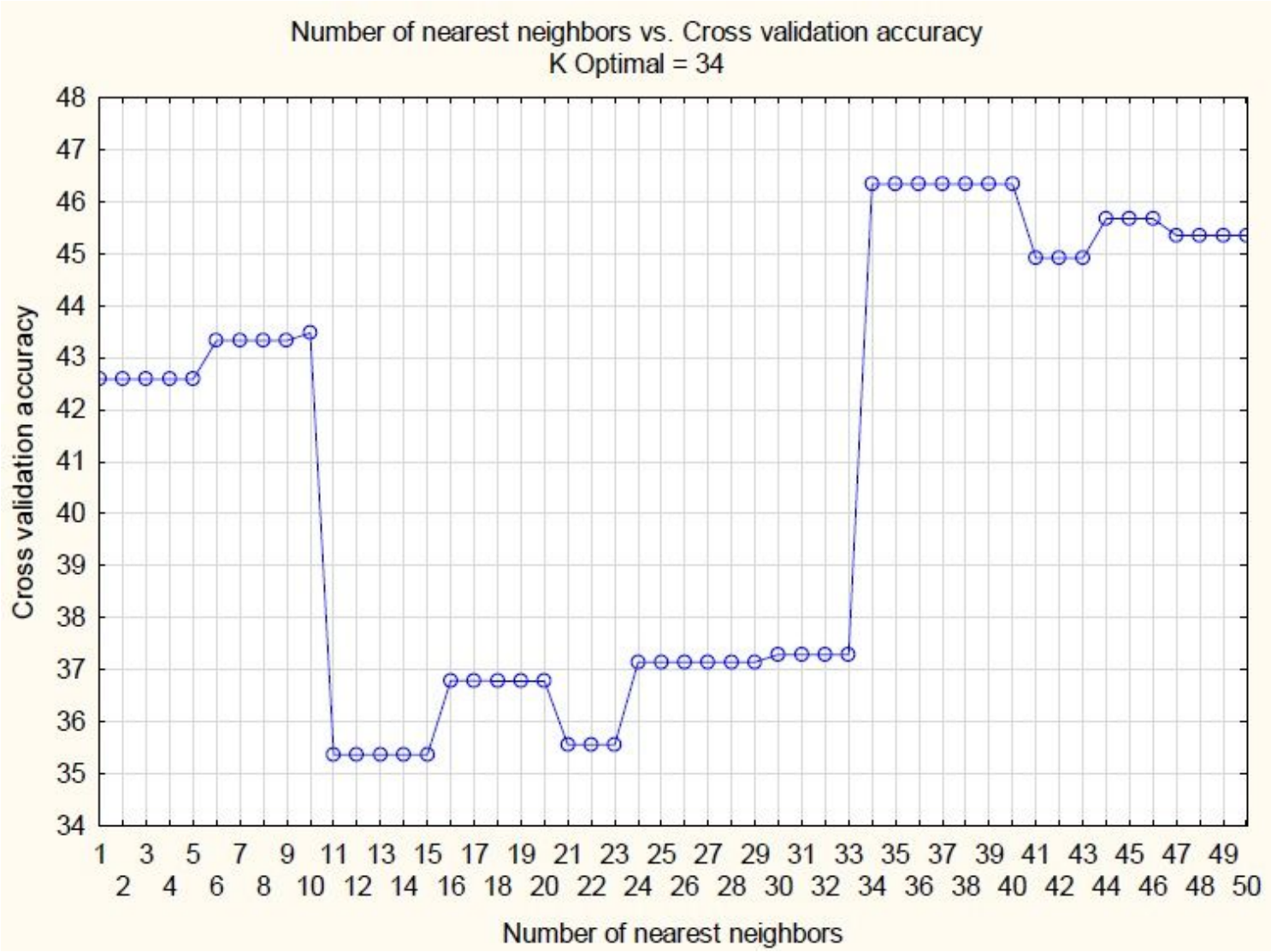

Figure 4

Cross validation accuracy against number of nearest neighbors