Analysis of Prognostic Factors Influencing Survival and Recurrence in Breast Cancer: A Hybrid Machine Learning Approach

Jae Bin Lee
Kongju University: Kongju National University
Jihye Choi
The University of Texas Health Science Center at Houston School of Public Health
Mi Sun An
Kongju University: Kongju National University
Jong-Yeup Kim
Konyang University College of Medicine
Seong Uk Kwon
Konyang University Hospital
Jungeun Kim
Kongju National University
Seunghee Lee
Konyang University Hospital
Seongwoo Jeon
Konyang University Hospital
ChungChun Lee
Konyang University Hospital
Suehyun Lee
Konyang University College of Medicine https://orcid.org/0000-0003-0651-6481
Hyekyung Woo (hkwoo@kongju.ac.kr)
Kongju National University https://orcid.org/0000-0001-5489-3404

Research Article

Keywords: Breast cancer, precise prognosis, machine learning, electronic medical records

Posted Date: October 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-946765/v1

Abstract

Purpose: The present study sought to identify prognostic factors for breast cancer survival and recurrence using a machine learning approach and electronic medical record data.

Methods: We used a machine learning technique called feature selection to identify factors influencing breast cancer prognosis, and factors affecting survival and recurrence in a Cox regression model.

Results: History of relapse, type of surgery, diagnostic route, SEER stage, and hormone therapy all affected breast cancer survival. Recurrence of breast cancer was affected by age, history of diabetes, breast reconstruction, pain, breast lumps, nipple discharge, and the presence of other symptoms. According to the survival analysis based on feature selection, patients with diabetes had a significantly higher risk of early recurrence of breast cancer (hazard ratio, 4.8; 95% confidence interval, 2.04–11.2, $p < 0.05$).

Conclusions: The present study identified several factors associated with breast cancer prognosis. While survival was affected by the diagnostic route, recurrence was primarily influenced by breast cancer symptoms and other underlying health conditions. A more accurate and standardized model considering time-to-event data could be developed in the future to evaluate prognostic factors and predict prognoses, and for clinical validation

Introduction

According to Global Cancer Statistics 2020, breast cancer is the most commonly diagnosed cancer and leading cause of cancer mortality in women worldwide [1]. The National Cancer Institute recently reported an annual incidence rate for female breast cancer of 12.9% [2]. Breast cancer is a highly heterogeneous disease with multiple risk factors and clinical outcomes associated with biologically distinct subtypes of breast tumors [3, 4]. Accurate cancer prognosis, along with early diagnosis, has become a necessity in oncology for effective clinical management of patients [5]. As with most types of cancer, accurate prognosis for breast cancer is important, as it informs the patient about likely disease progression, enables patient-specific treatments to be implemented, and allows policymakers to compare mortality rates among hospitals [6].

Cancer prognosis has mainly been predicted using clinical data, such as age at diagnosis, histological type, hormone therapy, lymph node status, and tumor characteristics [7, 8]. However, few prognostic models consider the behavior, psychological health, and quality of life of patients, despite the critical role of these factors in predicting cancer prognosis [9, 10]. The inconsistency in the clinical and nonclinical factors included in prognostic models, and the limited accuracy of these models due to manual entry of variables, has been improved by using electronic medical record (EMR) data to develop automated prognostic models in oncology [11]. EMRs, which include a wide range of personal and clinical data recorded in real time, offer numerous benefits including easy transfer of patient information, support of

optimal decision-making by care providers, and insight into the relationship between healthcare provider practice patterns and prognosis [12].

In addition to greater use of EMR data, there has been a gradual shift toward leveraging artificial intelligence (AI) for cancer prognosis prediction. As a subfield of AI, machine learning (ML) methods have been successfully applied for cancer prognosis prediction [5]. Deep learning (DL) algorithms, as a subtype of ML, have also improved the accuracy of cancer prognosis relative to traditional analytic methods, such as Cox proportional hazard regression, the Kaplan-Meier estimator, and the log-rank test [13]. Although ML and DL are useful for discovering new patterns in high dimensional data, one of their disadvantages is that they tend to disregard time-to-event data due to their focus on predictions of death and recurrence [6]. To overcome this problem, a hybrid ML approach using time-to-event data is being devised for survival analysis, as a means of integrating traditional and novel methods [14, 15].

The present study sought to identify prognostic factors for breast cancer survival and recurrence using ML approaches and EMR data.

Materials And Methods

Data source

This study utilized secondary data collected from breast cancer patients at Konyang University Hospital (KUH), which participated in the construction of a nationwide big data platform. The Korea Cancer Big Data Platform is a multi-database system that collects and stores electronic cancer patient data on diagnoses, examinations, treatments, and surgeries. KUH currently has 1,900 breast cancer patients; the cancer registration data of 896 of these patients were analyzed in this study.

Data preprocessing

Variables with missing data were excluded from the analysis. Additional data were collected from the cancer register. Table 1 presents the variables in the dataset analyzed in this study.

Table 1
Description of dataset variables

Independent variables	
Demographics	Gender, age, marital status, number of childbirths
General characteristics	Medication status, menopausal status, HRT, current smoking status, breastfeeding, contraceptive use, insomnia
Past medical history	Diabetes mellitus, tuberculosis, hypertension, heart disease, smoking history, other health conditions
Family health history	Cancer, hypertension
Surgery characteristics	Surgery type (total mastectomy, breast-conserving surgery), ALND, SLNB, breast reconstruction
Preoperative screening	Asymptomaticity, pain, breast lump, axillary lump, nipple discharge, axillary pain
Cancer registration data	Foreign, laterality, cancer topography, “morphological diagnostic” method, diagnostic route, SEER stage, chemotherapy, radiotherapy, hormone therapy, immunotherapy
Dependent variables	
Survival	Status, time
Recurrence	

HRT hormone replacement therapy, ALND axillary lymph node dissection, SLNB sentinel lymph node biopsy, SEER stage The Surveillance, Epidemiology and End Results stage

Data splitting and oversampling

The dataset was divided into training and test sets. Imbalances were detected in the death and recurrence data. As imbalanced datasets are likely to result in overfitting in some ML models [16], we applied the synthetic minority oversampling technique (SMOTE) to the training dataset. In this technique, the minority class is oversampled by creating “synthetic” examples rather than replacing the data [17].

Feature selection

Feature selection was used to determine prognostic factors for breast cancer. Feature selection is an ML technique that reduces dimensionality by removing redundant and irrelevant features from an original dataset based on certain criteria [18, 19]. Feature selection reduces noise in the data and improves the performance of the model [20]. LASSO regression was used in this study as the feature selection algorithm. Formulated by Robert Tibshirani in 1996 [21], LASSO is a powerful method for regularization and feature selection [22]. LASSO performs feature selection by estimating the regression coefficients for

non-significant features, which are close to zero. The LASSO method achieves good predictive accuracy by reducing the variance caused by shrinking and removing coefficients, and enhances interpretability by eliminating variables not associated with the response variable [22]. To evaluate the feature selection results, we generated an ML-based prognostic model. Techniques to evaluate model performance include random forest (RF), support vector machine (SVM), and logistic regression (LR). 10-fold cross-validation was applied to the model to prevent overfitting. Finally, we analyzed the performance of the model using a receiver operating characteristic curve.

Statistical analysis

Hazard ratios (HRs) with 95% confidence intervals (CIs) were calculated to assess the associations of prognostic factors with breast cancer survival and recurrence using multivariate survival analysis. All statistical analyses were conducted using R software (version 4.0.0; R Foundation for Statistical Computing, Vienna, Austria).

Results

The prognostic impact of the features selected using the LASSO method is shown in Table 2. In the case of survival, binary deviance was lowest when five features were selected. This result was different from that of recurrence, where binary deviance was lowest when eight features were selected. Among the factors affecting the likelihood of death of breast cancer patients, recurrence had the highest coefficient. When recurrence was the intercept, a history of diabetes mellitus (DM) had the greatest impact. The LR recurrence prediction model demonstrated the best performance (area under the curve: 0.83).

Table 2
Statistics for the prognostic factors

Variable	Categories	Alive (n = 841)	Dead (n = 55)	*p*-value
Recurrence	No	812 (94.7)	45 (5.3)	**<0.001**
	Yes	29 (74.4)	10 (25.6)	
Type of surgery	No	366 (90.6)	38 (9.4)	**<0.001**
	Mastectomy	56 (98.2)	1 (1.8)	
	Breast-conserving surgery	419 (96.3)	16 (3.7)	
Diagnostic route	Health screening	383 (97.5)	10 (2.5)	**<0.001**
	Accidental discovery	13 (68.4)	6 (31.6)	
	Symptoms	308 (93.3)	22 (6.7)	
	Unknown	137 (89.0)	17 (1.9)	
SEER stage	In situ	131 (98.5)	2 (1.5)	**<0.001**
	Localized	325 (96.7)	11 (3.3)	
	Regional	233 (94.3)	14 (5.7)	
	Distant	26 (68.4)	12 (31.6)	
	Unspecified	126 (88.7)	16 (11.2)	
Hormone therapy	Negative	610 (92.4)	50 (7.6)	**<0.01**
	Positive	231 (97.9)	5 (2.1)	
		No recurrence (n = 857)	Recurrence (n = 39)	
Age (years)	≤ 40	119 (96.7)	4 (3.3)	**0.24**
	41–64	603 (96.0)	25 (4.0)	
	≥65	135 (93.1)	10 (6.9)	
History of DM	No	823 (96.4)	31 (3.6)	**<0.001**
	Yes	34 (81.0)	8 (19.0)	
Breast reconstruction	No	823 (96.0)	34 (4.0)	**<0.05**

Accidental discovery: Discovered incidentally during surgical examination

Variable	Categories	Alive (n = 841)	Dead (n = 55)	*p*-value
	Yes	34 (87.2)	5 (12.8)	
Pain	No	812 (96.1)	33 (3.9)	**<0.05**
	Yes	45 (88.2)	6 (11.8)	
Breast lump	No	542 (97.8)	12 (2.2)	**<0.001**
	Yes	315 (92.1)	27(7.9)	
Nipple discharge	No	834 (95.9)	36 (4.1)	**0.18**
	Yes	23 (88.5)	3 (11.5)	
Other symptoms	No	802 (96.3)	31 (3.7)	**<0.01**
	Yes	55 (87.3)	8 (12.7)	
SEER stage	In situ	133 (100.0)	0(0.0)	**<0.001**
	Localized	325 (96.7)	11 (3.3)	
	Regional	233 (94.3)	14 (5.7)	
	Distant	32 (84.2)	6 (15.8)	
	Unspecified	135 (94.4)	8 (5.6)	
Accidental discovery: Discovered incidentally during surgical examination				

Performance of the prognostic factors

The statistics for the prognostic factors derived by feature selection are shown in Table 3. In total, 74.4% (29/39) of the recurrence group survived, and recurrence was a significant prognostic factor for mortality ($p < 0.001$). Regarding type of surgery, which was also a significant prognostic factor for mortality ($p < 0.001$), the death rate was highest in the no surgery subgroup (38/404; 9.4%). SEER stage ($p < 0.001$), diagnostic route ($p < 0.001$), and hormone receptor therapy ($p < 0.01$) were also significant prognostic factors for mortality. In total, 6.9% (10/145) of patients aged ≥ 65 years experienced recurrence and the difference in recurrence among the age groups was not significant. In total, 19.0% (8/42) of the patients with a history of DM experienced breast cancer recurrence, and DM was a significant prognostic factors for mortality ($p < 0.001$).

Table 3
Multivariate analysis of prognostic factors for breast cancer

Prognosis	Variable	Categories	HR (95% CI)	*p*-value
Overall Survival	Recurrence	No	1	
		Yes	2.98 (1.44–6.07)	<0.01
	Type of surgery	None	1	
		Mastectomy	0.53 (0.07–4.14)	0.55
		Breast-conserving surgery	0.33 (0.16–0.68)	<0.01
	Diagnostic route	Health screening	1	
		Incidentally discovered	11.15 (2.38–21.5)	<0.001
		Symptoms	2.49 (1.11–5.60)	<0.05
		Unknown	1.29 (0.53–3.18)	0.58
	SEER stage	In situ	1	
		Localized	1.16 (0.24–5.67)	0.85
		Regional	1.26 (0.25–6.41)	0.79
		Distant	7.96 (1.52–41.8)	<0.05
		Unspecified	1.79 (0.34–9.32)	0.49
	Hormone therapy	No	1	
		Yes	2.12 (0.79–5.74)	0.14
Recurrence -free survival	Age (years)	≤ 40	1	
		41–64	1.3 (0.44–3.9)	0.63
		≥ 65	2.7 (0.78–9.2)	0.12
	History of DM	No	1	
		Yes	4.8 (2.04–11.2)	<0.001
	Breast reconstruction	No	1	
		Yes	2.1 (0.72–5.9)	0.18
	Pain	No	1	
		Yes	1.1 (0.40–3.3)	0.81
	Breast lump	No	1	
		Yes	2.8 (1.32–5.8)	<0.01

Prognosis	Variable	Categories	HR (95% CI)	*p*-value
	Nipple discharge	No	1	
		Yes	1.9 (0.54–6.4)	0.33
	Other symptoms	No	1	
		Yes	1.3 (0.51–3.5)	0.58
	SEER stage	In situ/localized	1	
		Regional	2.4 (1.03–5.4)	<0.05
		Distant	7.3 (2.35–22.9)	<0.001
		Unspecified	4.0 (1.55–10.1)	<0.001

In the multivariate survival analysis, SEER stage (HR, 7.96; 95% CI, 1.52–41.8; $p < 0.05$) and diagnostic route (HR, 11.15; 95% CI, 2.38–21.5, $p < 0.001$) were poor prognostic factors. In the recurrence analysis, patients who had diabetes had a significantly higher risk of early breast cancer recurrence (HR, 4.8; 95% CI, 2.04–11.2, $p < 0.001$). A history of diabetes mellitus had the highest coefficient, where univariate and multivariate analyses both showed that a history of diabetes was associated with a significantly higher risk of early recurrence. Table 4 presents the results of the multivariate analysis.

Discussion

In the present study, we analyzed EMR data to identify significant prognostic factors for breast cancer survival and recurrence using a combination of ML approaches and the Cox proportional hazard model. History of recurrence and SEER stage were related to breast cancer prognosis, and the type of surgery and diagnostic route were prognostic factors for breast cancer survival. A history of diabetes and the presence of a breast lump were prognostic factors for breast cancer recurrence.

Among the diagnostic routes, an incidental diagnosis of breast cancer (i.e., not during formal cancer screening or surveillance) in the absence of symptoms related to the tumor [23] was associated with a significantly higher risk of death than diagnoses via breast screening. Although it is difficult to draw definite conclusions regarding the effect of the diagnostic method on survival time due to a lack of relevant studies, breast cancer detected by regular health screening is an independent prognostic factor for breast cancer and associated with a more favorable survival rate [24, 25]. Future research should build on these findings and explore potential reasons for prognostic differences between screening-detected and incidentally detected breast cancer. Consistent with the present findings, previous research has reported that breast cancer patients who experience recurrence tend to report breast lumps [26]. Although the clinical significance of a breast lump for the risk of recurrence has already been established, our

findings based on big data and an ML approach provide confirmatory evidence that a breast lump is a key prognostic factor.

Notably, a history of diabetes increased the risk of early recurrence of breast cancer in this study. Previous prognostic analyses have emphasized the need to consider underlying diseases. For example, meta-analyses have revealed that poor overall and disease-free survival are associated with breast cancer among patients previously diagnosed with diabetes [27]. Furthermore, cooccurring cardiovascular and pulmonary diseases are likely to increase the risk of death in breast cancer patients [28]. Although the effect of diabetes on the early recurrence of breast cancer has been extensively investigated [27], our study provides additional evidence, based on ML and big data analysis, of the vital role that underlying health conditions play in breast cancer prognosis.

A strength of this study was the use of data from “scalable” EMRs of hospitals to identify key prognostic factors for breast cancer. Conventional prognostic predictions of breast cancer use data restricted in terms of populations or hospitals, or based on randomized controlled trials [6]. In contrast, EMR systems are comprehensive databases that help physicians to document patient care and contain detailed lists of patient symptoms. This expands the scope and accuracy of prognostic predictions. Despite these advantages of EMR systems, there are also some disadvantages, including redundant, missing or inaccurate data, and internally inconsistent progress notes [29], all of which pose a threat to patient safety. Thus, EMR systems require meticulous updating and maintenance.

Several limitations of this study should be discussed. First, we analyzed partial big data breast cancer datasets. Analysis of full datasets remains a challenge, as libraries of all breast cancer cases have not yet been compiled. Moreover, although hormone therapy, for example with estrogen or progesterone, is a key prognostic factor for breast cancer, this was not analyzed in the present study. In addition, while oversampling was done due to the relatively small number of patients available for the study, the elimination of variables with a large amount of missing data may have substantially affected the results. Nonetheless, the hybrid LASSO-Cox method used in this study was effective for identifying the most pertinent prognostic factors and improved the precision of the Cox proportional hazard model by omitting redundant features [14]. It is also notable that LASSO regression, which has strengths in terms of feature selection, was applied given that it has not been used extensively in breast cancer prognostic studies. Finally, our study demonstrated the immense potential of big data and ML techniques for obtaining new insight into breast cancer; in this manner, we have obtained a deeper understanding of the clinical course of the disease [30].

Conclusion

This study aimed to identify prognostic factors for breast cancer survival and recurrence through analysis of a comprehensive dataset (EMR system). By combining traditional statistical methods and ML, we confirmed key prognostic factors for breast cancer, such as a history of diabetes, which also showed an association with the risk of early recurrence. The most important task to achieve accurate prognostic

predictions using real-time health and medical data is to develop a standardized model that allows analysis of time-to-event data, to further evaluate prognostic factors and predict prognoses, and assess clinical validity.

Abbreviations

EMR electronic medical record

AI artificial intelligence

ML machine learning

DL deep learning

SMOTE synthetic minority oversampling technique

HR hazard ratios

CIs confidence intervals

HRT hormone replacement therapy

ALND axillary lymph node dissection

SLNB sentinel lymph node biopsy

SEER stage The Surveillance, Epidemiology and End Results stage

Declarations

Acknowledgements: We are grateful to Inmyung Song and Hyun Suk Lee for her valuable comment in writing the manuscript.

Funding: This study was supported by grants from the Big Data Center, National Cancer Center of Korea (grant number: 2020-data-we06), and National Research Foundation of Korea (NRF; grant number: 2020R1C1C009679).

Authors' contributions: HW, SL and JBL contributed to the design of the study and SL, JYK, SL, and CCL collected the data. JBL, HW, JC, MSA and JK carried out the statistical analyses, interpreted the results and drafted the manuscript. All the authors critically reviewed the manuscript and approved the final version.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability: The code that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Ethics approval: The study protocol was approved by the Clinical Trial Committee of Konyang University (KYUH2020-07-014-006)

Informed consent: This is an IRB-approved study, and the need for consent to written notification was waived.

The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see:

http://www.textcheck.com/certificate/ASCZ6M

References

1. Sung H et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA 71(3):209–249. https://doi.org/10.3322/caac.21660
2. Kohler BA et al., *Annual report to the nation on the status of cancer, 1975-2011, featuring*
3. *incidence of breast cancer subtypes by race/ethnicity, poverty, and state.* JNCI, 2015. **107**(6): p. djv048. https://doi.org/10.1093/jnci/djv048
4. Polyak K (2011) Heterogeneity in breast cancer. JCI 121(10):3786–3788. http://doi.org/10.1172/JCI60534
5. Dawood S et al (2011) Defining breast cancer prognosis based on molecular phenotypes: results from a large cohort study. Breast Cancer Res Treat 126(1):185–192. 10.1007/s10549-010-1113-7
6. Kourou K et al., *Machine learning applications in cancer prognosis and prediction.* CSBJ, 2015. **13**: p. 8-17. https://doi.org/10.1016/j.csbj.2014.11.005
7. Phung MT, Tin ST, Elwood JM (2019) Prognostic models for breast cancer: a systematic review. BMC Cancer 19(1):1–18. http://doi.org/10.1186/s12885-019-5442-6
8. Lee J-Y et al., Different prognostic significance of Bcl-2 based on Cancer molecular subtype. Journal of Breast Cancer, 2011. 14(Suppl 1): p. S10-S16. https://doi.org/10.4048/jbc.2011.14.S.S10
9. Sun Y-S et al (2017) Risk factors and preventions of breast cancer. International Journal of Biological Sciences 13(11):1387. DOI:10.7150/ijbs.21635
10. Lehto U-S et al (2019) Early quality-of-life and psychological predictors of disease-free time and survival in localized prostate cancer. Qual Life Res 28(3):677–686. DOI:10.1007/s11136-018-2069-z

11. Von Oetinger A, *Effect of lifestyle interventions over quality of life in colorectal cancer survivors.* Revista de Gastroenterologia del Peru: organo oficial de la Sociedad de Gastroenterologia del Peru, 2019. 39(2): pp 153–159
12. Gensheimer MF et al (2019) Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. JNCI 111(6):568–574. DOI:10.1093/jnci/djy178
13. Zhao C, Zhang L. *Research of information presentation for electronic medical record based on ontology*. in (2013) *6th International Conference on Information Management, Innovation Management and Industrial Engineering*. 2013. IEEE. **DOI**: 10.1109/ICIII.2013.6703628
14. Zhu W et al (2020) The application of deep learning in cancer prognosis prediction. Cancers 12(3):603. https://doi.org/10.3390/cancers12030603
15. Shahraki HR, Salehi A, Zare N, *Survival prognostic factors of male breast cancer in Southern Iran: a LASSO-Cox regression approach.* APJCP (2015) **16**(15): p. 6773-6777. http://doi.org/10.7314/APJCP.2015.16.15.6773
16. Du M et al (2020) Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database. Cancers 12(10):2802. DOI:https://doi.org/10.3390/cancers12102802
17. Santos MS et al., *Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier].* ieee ComputatioNal iNtelligeNCe magaziNe, 2018. 13(4): p. 59–76. **DOI**: 10.1109/MCI.2018.2866730
18. Chawla NV et al (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953, , :. **DOI**
19. Naseriparsa M, Bidgoli A-M, Varaee T, *A hybrid feature selection method to improve performance of a group of classification algorithms.* arXiv preprint arXiv:1403.2372, 2014. DOI: 10.5120/12065-8172
20. Cai J et al (2018) Feature selection in machine learning: A new perspective. Neurocomputing 300:70–79. https://doi.org/10.1016/j.neucom.2017.11.077
21. Khalid S, Khalil T, Nasreen S. *A survey of feature selection and feature extraction techniques in machine learning*. in *2014 science and information conference* (2014) IEEE. **DOI**: 10.1109/SAI.2014.6918213
22. Tibshirani R, *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267- 288. https://doi.org/10.1111/j.25176161.1996.tb02080.x
23. Fonti V, Belitser E (2017) Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics 30:1–25
24. Koo MM, Rubin G, McPhail S, Lyratzopoulos G (2019) Incidentally diagnosed cancer and commonly preceding clinical scenarios: a cross-sectional descriptive analysis of English audit data. BMJ Open 9(9):e028362. http://dx.doi.org/10.1136/bmjopen-2018-028362
25. Lehtimäki T, Lundin M, Linder N, Sihto H, Holli K, Turpeenniemi-Hujanen T, Lundin J (2011) Long-term prognosis of breast cancer detected by mammography screening or other methods. Breast Cancer

Res 13(6):1–11

26. Pham TM et al (2019) Diagnostic route is associated with care satisfaction independently of tumour stage: Evidence from linked English Cancer Patient Experience Survey and cancer registration data. Cancer Epidemiol 61:70–78. DOI:10.1016/j.canep.2019.04.011
27. Buist DSM et al (2010) Diagnosis of second breast cancer events after initial diagnosis of early stage breast cancer. Breast Cancer Res Treat 124(3):863–873. DOI:10.1007/s10549-010-1106-6
28. Zhao X-B, Ren G-S, *Diabetes mellitus and prognosis in women with breast cancer: A systematic review and meta-analysis*. Medicine, 2016**95**(49). DOI: 10.1097/MD.0000000000005602
29. Riihimäki M et al (2012) Death causes in breast cancer patients. Ann Oncol 23(3):604–610
30. Bowman S (2013) Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in Health Information Management*, *10*(Fall)
31. Lánczky A, Nagy Á, Bottai G, Munkácsy G, Szabó A, Santarpia L, Győrffy B (2016) miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. Breast Cancer Res Treat 160(3):439–446

Figures

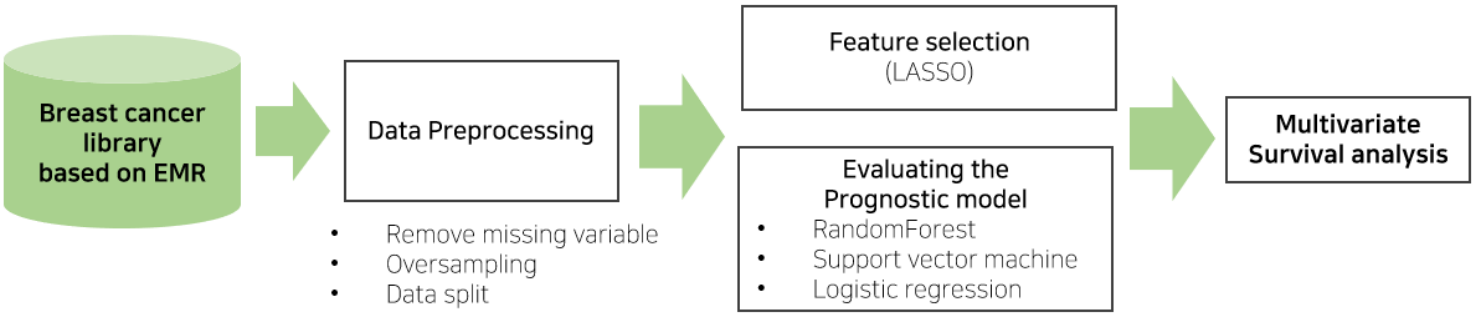
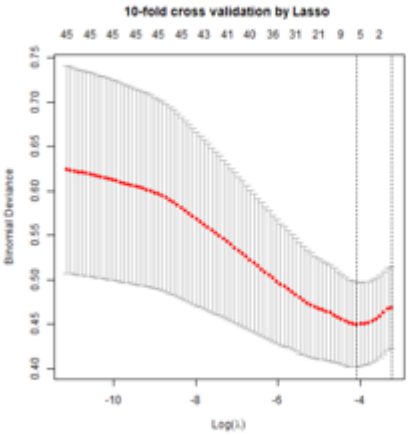
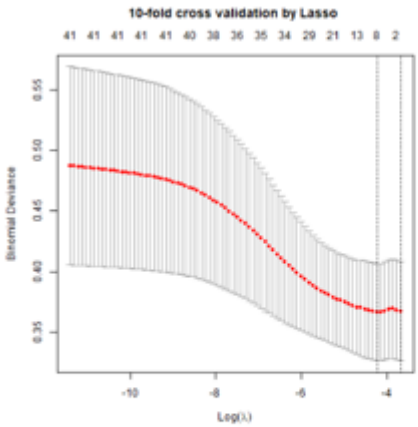


Figure 1

Study flow chart.



Survival data.



Recurrence data.

Dependent variable	Best prognostic factors			Coefficient
Death	(Intercept)			−3.50
	Type of surgery			0.01
	Recurrence			0.37
	Diagnostic route			0.21
	SEER stage			0.07
	Hormone therapy			−0.11
Performance evaluation	Accuracy	Sensitivity	Specificity	AUC
RF	0.80	0.80	0.44	0.70
SVM	0.80	0.82	0.50	0.71
LR	0.67	0.66	0.75	0.73
Recurrence	(Intercept)			−3.56
	Age			0.04
	History of diabetes mellitus			1.00
	Breast reconstruction			0.05
	Pain			0.24
	Breast lump			0.40
	Nipple discharge			0.54
	Other symptoms			0.08
	SEER stage			0.05
Performance evaluation	Accuracy	Sensitivity	Specificity	AUC
RF	0.78	0.78	0.82	0.80
SVM	0.76	0.77	0.73	0.79
LR	0.83	0.83	0.73	0.83

RF random forest, SVM support vector machine, LR logistic regression

Figure 2

Best prognostic factors based on LASSO

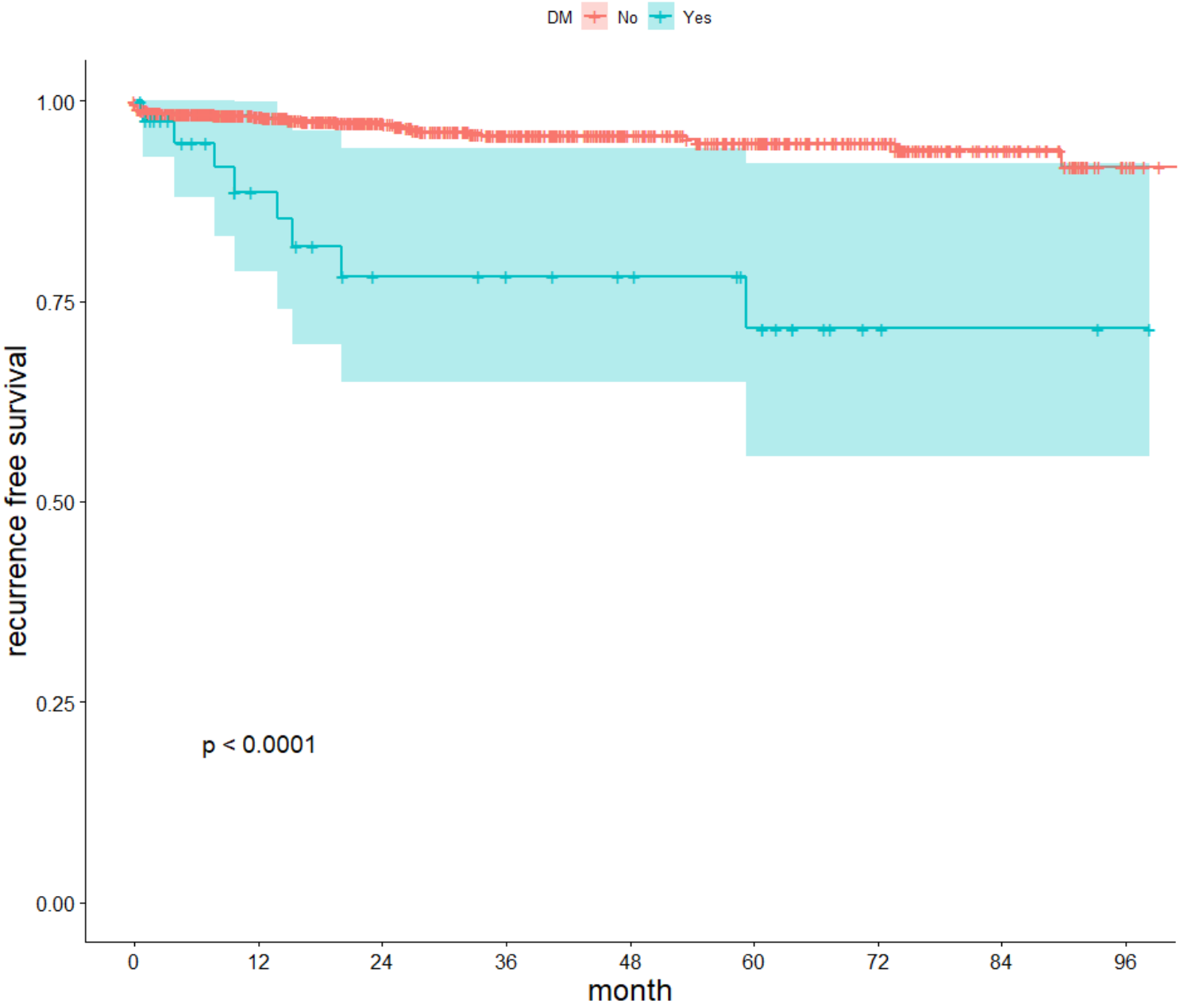


Figure 3

Recurrence-free survival time by history of diabetes mellitus.