Journal of Integrative Oncology
ISSN: 2329-6771

Chen et al., J Integr Oncol 2016, 5:3
DOI: 10.4172/2329-6771.1000174

Research Article Open Access

Validation of Breast Cancer Survival Prediction Model with SEER Database

Yu-Chieh Chen[1], Hung-Wen Lai[2], Wen-Ching Wang[3]* and Yao-Lung Kuo[4]*

[1]*Department of Gynecology and Obstetrics, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan*
[2]*School of Medicine, National Yang Ming University, Taipei, Taiwan*
[3]*Division of General Surgery, Department of Surgery, Chi-Mei Medical Center, Tainan, Taiwan*
[4]*Department of Surgery, National Cheng Kung University, College of Medicine, Tainan and Dou-Liou, Taiwan*

Abstract

Objective: The accurate estimation of outcome in postoperative breast cancer patients is an essential component of the individualized treatment, decision-making, and patient counseling processes. The disease outcome and prognosis of breast cancer patients may vary according to geographic and ethnic factors. To clarify this topic, we created a new prognostic and predictive model for breast cancer patients, based on clinical and pathological variables.

Study design and setting: Clinical and pathological data were collected from 1587 patients with breast cancer who underwent surgical intervention. A survival prediction model was used to allow the analysis of the optimal combination of variables. The area under the receiver operating characteristic (ROC) curve, as applied to an independent validation data set, was used as the measure of accuracy. Results were assessed by comparing the area under the ROC curve with the SEER database.

Results: Our predictive model of survival predicted disease outcome for individual patients with breast cancer. The comparison between our predictive model and SEER databases showed that our model underestimated outcome in the SEER cohort and that the SEER model overestimated outcome in our breast cancer patients.

Conclusion: Our model may present an alternative as personalized prognostic tool for breast cancer patients. Decision regarding the survival prediction should take every consideration about regional and racial factors into account.

Keywords: Breast cancer; SEER; Predictive model

Introduction

Accurate survival prediction is a crucial component in the decision-making process for postoperative breast cancer patients. Therefore, estimation of the disease risk may help patients and physicians reach a decision regarding further adjuvant treatment. Prediction of breast cancer survival remains an important issue for each patient [1,2]. Using the well-established, classical prognostic factors recommended by the St. Gallen consensus, age, tumor grade, tumor size, lymph node status, and hormone receptor status, we developed a prognostic model to predict the survival of breast cancer in Taiwanese women. To verify the accuracy of the formula, we compared our database with the Surveillance, Epidemiology, and End Results (SEER) data bank [3]. An accurate prognostic model should prevent any over- or underestimation of each patient. In a previous publication, we demonstrated that different race can vary in the prognosis and development of the disease [4]. Therefore, the comparison of different prognostic models or patient collectives is essential and necessary for the verification of the accuracy of the model [5,6].

Our model is a prognostic and predictive model based on postoperative breast cancer patients treated in a university hospital. This model is based on the prognostic factors recommended by the St Gallen consensus, age, tumor size and grade, lymph node status, and hormonal status [7]. This model was developed to predict survival in Taiwanese breast cancer patients. The validation of our model is conducted via comparison with the data set of the SEER program.

Therefore, the primary aim of this study was to develop a model of prognostication of the overall survival in a large cohort of Taiwanese breast cancer patients who were diagnosed from 2002 to 2009. The secondary aim was to validate our model as a prognostic and predictive model for postoperative breast cancer patients in Taiwan, using the SEER data set.

Patients and Method

The original data were collected from 2105 patients with breast cancer diagnosed and treated at the National Cheng Kung University Hospital (NCKUH), Tainan and Dou-Liou Branch, Taiwan. Patient databases were identified from the medical records of the cancer registry at NCKUH. The accuracy of the clinical and pathological information of each patient was reviewed and revised by clinicians and study nurses. As our objective was to study the prognostic factors of breast cancer and to develop more precise predictive survival models, patients who were followed for less than 1 year were excluded from our analyses. Patients with metastatic disease at diagnosis or ductal carcinoma *in situ* were also excluded. Ethical approval was provided by Human Experiment and Ethics committee of the National Cheng Kung University Hospital (NCKUH9901006).

***Corresponding authors:** Wen-Ching Wang, Division of General Surgery, Department of Surgery, Chi-Mei Medical Center, Tainan, Taiwan, Tel: 886-6-276-6689; Fax: 886-6-276-6189; E-mail: surgeoncage122@gmail.com

Yao-Lung Kuo, College of Medicine, Department of Surgery, National Cheng Kung University Hospital, Tainan and Dou-Liou Branch, Taiwan, 138 Sheng-Li Road, Tainan 704, Taiwan, Tel: 886-6-276-6689, Fax: 886-6-276-6189; E-mail: ylkuo@mail.ncku.edu.tw

Received July 03, 2016; **Accepted** July 15, 2016; **Published** July 22, 2016

Citation: Chen YC, Lai HW, Wang WC, Kuo YL (2016) Validation of Breast Cancer Survival Prediction Model with SEER Database. J Integr Oncol 5: 174. doi: 10.4172/2329-6771.1000174

A variety of potential breast cancer risk factors were constructed for each patient. The demographic data included age at the presence of cancer and the pathological findings included tumor size, axillary lymph node status, tumor histological grade, estrogen receptor (ER) status, and progesterone receptor (PR) status. All pathological specimens were reviewed by breast pathologists at NCKUH. Tumor size was determined based on pathological reports from NCKUH. The Bloom–Richardson system was used for tumor grading, which was based on the following morphological features: nuclear pleomorphism of tumor cells, degree of tumor tubule formation, and tumor mitotic activity. To determine the ER and PR status, immunohistochemistry (IHC) was performed on formalin-fixed, paraffin-embedded breast cancer tissue samples from the patients. Positive ER and PR status was defined as nuclear staining of >1%. IHC was performed using anti-ER (clone 6F11, Ventana Medical System, Strasbourg, France) and anti-PR (clone PR636, Dako, Carpinteria, CA) antibodies. Postoperative adjuvant chemotherapy was performed according to NCCN and St. Gallen guidelines.

The SEER database contained patient data obtained from the SEER web site. Both databases (SEER and NCKUH databases) collected only patients who had completed a follow-up period of >5 years. The collection period for both cohorts was from 2002 to 2009.

Statistical methods

The data were expressed as the mean ± SD for continuous prognostic factors. The χ^2 and two-sample independent t tests were used to compare variables between the SEER and NCKUH data banks (Table 1). A logistic regression model with computed odds ratio and *p*-value was used to assess the risk of 5-year mortality relative to the prognostic factors in breast cancer patients [8]. Significance was set at a *p*-value < 0.05. Models 1 (SEER model) and 2 (NCKUH model) were derived from multivariate logistic models and $\hat{P}$, which was used to run ROC curves, represented the predicted 5-year death probability [9].

The following formula was used for the SEER model.

PI=-6.115+0.031Age+0.549Grade(IIv.s.I)+1.162Grade(IIIv.s.I)+1.023Grade(IVv.s.I)+0.765Tumor(2v.s.0-1)+1.491Tumor(3-4v.s.0-1)+0.882Node(1v.s.0)+1.484Node(2v.s.0)+2.161Node(3v.s.0)+1.054Hormone Receptor(-v.s.+)

$$\hat{P} = \frac{e^{PI}}{1+e^{PI}}$$

The following formula was used for our model.

PI=-4.725-0.003Age+0.975Grade(IIv.s.I)+1.440Grade(IIIv.s.I)+0.228Tumor(2v.s.0-1)+0.853Tumor(3-4v.s.0-1)+0.921Node(1v.s.0)+1.257Node(2v.s.0)+1.882Node(3v.s.0)+1.001Hormone Receptor(-v.s.+)

Two methods were used for the evaluation of the fitness of the multivariate logistic regression model. First, the Hosmer–Lemeshow test, written as $H = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{N_g \pi_g (1-\pi_g)} \sim \chi^2_{n-2}$ for the tested statistic (where O_g was the observed event, E_g was the expected event, N_g was the observation, and π_g was the predicted risk for the gth risk decile group), was used to examine the fitness of the logistic model taking into consideration the difference between the predicted and observed death probabilities caused by breast cancer. The statistic H was well approximated by the χ^2 distribution with n–2 degrees of freedom, χ^2_{n-2}. The constructed logistic model would be considered reasonable if *p*> 0.05. The fit of the model improved with the increase of the *p*-value, as estimated using the Hosmer–Lemeshow test [10]. Second, a receiver operating characteristic (ROC) curve was drawn to show the sensitivity and specificity of the predictive model at each cut point. The area under the curve (AUC) was calculated to assess the discriminative power of the model. All statistical analyses were performed using the SPSS 22.0 software (SPSS Inc., Chicago, IL).

Results

The SEER data included 68,634 subjects and the NCKUH database included 1,587 patients. Among the other prognostic factors, histological grading, tumor size, lymph node status, and hormone receptor status also exhibited significant differences between the cohorts (Table 1).

The SEER and NCKUH data sets were used to develop the primary prognostic models of breast-cancer-specific mortality. ß coefficients and standard errors were also calculated in both models for each prognostic factor. Univariate and multivariate logistic analyses performed using the SEER data revealed that age, tumor size, lymph node status, tumor grade, and hormone receptor status were prognostic factors that were significantly associated with the overall survival of breast cancer patients (Tables 2 and 3).

The univariate and multivariate logistic analyses of the prognostic factors associated with overall survival showed that age was not a significant prognostic factor in the NCKUH cohort. Tumor size, lymph node status, tumor grade, and hormone receptor status were significantly associated with overall survival (Tables 4 and 5).

The SEER and our models

The fitness of both models was well validated. Therefore, these models were well calibrated. Figure 1 shows that the SEER database fitted well into the SEER model (Model 1), with a perfect prediction of the 5-year mortality probability. Model discrimination was also good—the calculated area under the ROC curve (AUC) for the overall model was 0.822 ($p < 0.00001$) (Figure 2). The same validation was performed for the NCKUH database, which showed that the NCKUH data fitted well into the NCKUH model (Model 2) (Figure 3). Model

	SEER (n = 68634)	NCKUH (n = 1587)	p-value
Age			
Mean	59.62	49.8	<0.0001
SD	13.33	10.54	
Grade			
I	12820 (18.7 %)	313 (30.5 %)	<0.0001
II	29262 (42.6 %)	473 (46 %)	
III	24586 (35.8 %)	241 (23.5 %)	
IV	1966 (2.9 %)		
Tumor Size			
T0 or T1	40874 (59.6 %)	408 (39.7 %)	<0.0001
T2	23567 (34.3 %)	468 (45.6 %)	
T3 or T4	4193 (6.1 %)	151 (14.7 %)	
Lymph Node			
N0	45307 (66.0 %)	563 (54.8 %)	<0.0001
N1	14979 (21.9 %)	274 (26.7 %)	
N2	5439 (7.9 %)	117 (11.4 %)	
N3	2909 (4.2 %)	73 (7.1 %)	
Hormone Receptor			
Positive	54708 (79.7 %)	737 (71.8 %)	<0.0001
Negative	13926 (20.3 %)	290 (28.2 %)	

Table 1: Descriptive statistics of each prognostic factor.

Prognostic factors	Exp (ß)	Asymptotic 95% CI of exp (ß)		p-value
		lower	upper	
Age	1.014	1.012	1.016	<0.0001
Grade				<0.0001
II vs. I	2.651	2.378	2.955	<0.0001
III vs. I	7.850	7.071	8.715	<0.0001
IV vs. I	6.434	5.515	7.506	<0.0001
Tumor Size				<0.0001
T2 vs. T0-T1	3.865	3.659	4.083	<0.0001
T3-T4 vs. T0-T1	11.599	10.754	12.511	<0.0001
Lymph Node				<0.0001
N1 vs. N0	2.989	2.817	3.171	<0.0001
N2 vs. N0	6.859	6.394	7.358	<0.0001
N3 vs. N0	14.849	13.670	16.13	<0.0001
Hormone Receptor				
Negative vs. Positive	3.653	3.478	3.837	<0.0001

Table 2: Prognostic factors for overall survival in univariate logistic regression analysis for SEER databank.

Prognostic factor	Exp (ß)	Asymptotic 95% CI of exp (ß)		p-value
		lower	upper	
Age	1.032	1.030	1.034	<0.0001
Grade				<0.0001
II vs. I	1.731	1.546	1.938	<0.0001
III vs. I	3.197	2.856	3.579	<0.0001
IV vs. I	2.782	2.349	3.293	<0.0001
Tumor Size				<0.0001
T2 vs. T0-T1	2.149	2.023	2.282	<0.0001
T3-T4 vs. T0-T1	4.443	4.074	4.847	<0.0001
Lymph Node				<0.0001
N1 vs. N0	2.416	2.266	2.575	<0.0001
N2 vs. N0	4.410	4.077	4.771	<0.0001
N3 vs. N0	8.682	7.910	9.530	<0.0001
Hormone Receptor				
Negative vs. Positive	2.868	2.704	3.042	<0.0001

Table 3: Prognostic factors for overall survival in multiple logistic regression analysis for SEER databank.

discrimination was also good—the calculated area under the ROC curve (AUC) for the overall model was 0.798 ($p < 0.00001$) (Figure 4).

Validation

Overall, the model was well calibrated and model discrimination was good. Fitting of the SEER data into Models 1 and 2 showed that the calculated area under the ROC curve (AUC) for the overall model was 0.822 and 0.792, respectively ($p < 0.00001$) (Figure 5). Similarly, the NCKUH data fitted into Models 1 and 2. The calculated area under the ROC curve (AUC) for the overall model was 0.78 and 0.80, respectively ($p < 0.00001$) (Figure 6).

Model 1 tended to overestimate the mortality of the NCKUH patients. Similarly, the NCKUH model (Model 2) exhibited a tendency to underestimate the mortality of the SEER cohort.

The 5-year survival probability for breast cancer in the SEER group was 88.4% and the 5-year survival probability for the breast cancer patients in the NCKUH cohort was 92.2%. NCKUH patients tended to have a younger mean age compared with the SEER patients, of about 10 years (49.8 vs 59.62 years, $p < 0.0001$).

Discussion

The individualized and precise prediction of survival, and its consequent benefits to treatment modalities, has become increasingly sophisticated and important in the management of postoperative breast cancer patients worldwide and in Taiwan [4,5,11].

Several prognostic tools, such as the web-based program Adjuvant! Online, are available for breast cancer patients. However, few studies have evaluated the accuracy of prediction models by comparing them with regional databases [2,12-14]. We developed a model of prognostication of postoperative breast cancer patients based on data collected from a large number of cases within a cancer registry in

Prognostic factor	Exp (ß)	Asymptotic 95% CI of exp (ß)		p-value
		lower	upper	
Age	1.005	0.984	1.027	0.626
Grade				<0.0001
II vs. I	3.710	1.632	8.431	0.002
III vs. I	7.677	3.352	17.583	<0.0001
Tumor Size				<0.0001
T2 vs. T0-T1	2.165	1.188	3.945	0.012
T3-T4 vs. T0-T1	5.096	2.648	9.806	<0.0001
Lymph Node				<0.0001
N1 vs. N0	3.002	1.630	5.526	<0.0001
N2 vs. N0	5.206	2.639	10.268	<0.0001
N3 vs. N0	8.692	4.274	17.675	<0.0001
Hormone Receptor				
Negative vs. Positive	3.483	2.190	5.538	<0.0001
Chemotherapy Type				0.01
Type I vs. no	0.743	0.416	1.326	0.315
Type II vs. no	1.857	1.055	3.267	0.032
Reject or interrupt vs. no	2.125	0.246	18.378	0.493

Table 4: Prognostic factors for overall survival in univariate logistic regression analysis for NCKUH databank.

Prognostic Factor	Exp (ß)	Asymptotic 95% CI of exp (ß)		p-value
		lower	upper	
Age	0.992	0.969	1.015	0.485
Grade				0.003
II vs. I	2.549	1.091	5.956	0.031
III vs. I	4.333	1.801	10.420	0.001
Tumor Size				0.036
T2 vs. T0-T1	1.320	0.693	2.513	0.398
T3-T4 vs. T0-T1	2.482	1.187	5.187	0.016
Lymph Node				<0.0001
N1 vs. N0	2.774	1.421	5.415	0.003
N2 vs. N0	3.777	1.673	8.527	0.001
N3 vs. N0	7.698	3.285	18.037	<0.0001
Hormone Receptor				
Negative vs. Positive	2.954	1.761	4.956	<0.0001
Chemotherapy Type				0.329
Anthracycline contained regimen vs. no	0.658	0.341	1.272	0.214
Taxane contained regimen vs. no	0.554	0.284	1.079	0.083
Reject or interrupt vs. no	0.347	0.027	4.444	0.416

Table 5: Prognostic factors for overall survival in multiple regression analysis for NCKUH databank.

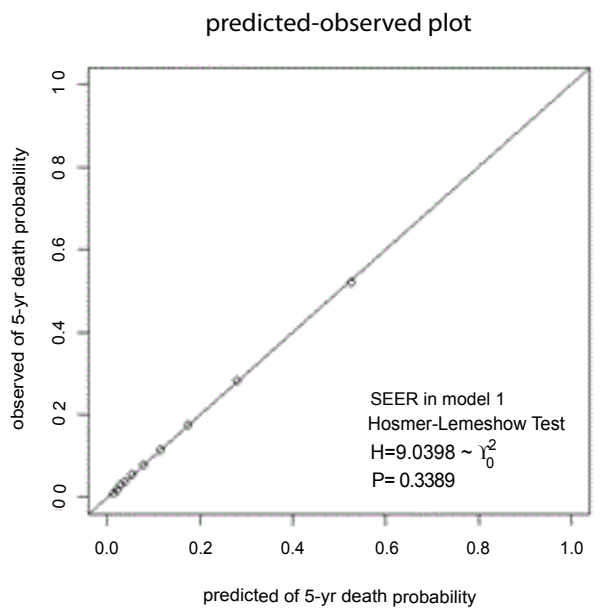

Figure 1: The fitness of SEER database in the model 1 (SEER model).

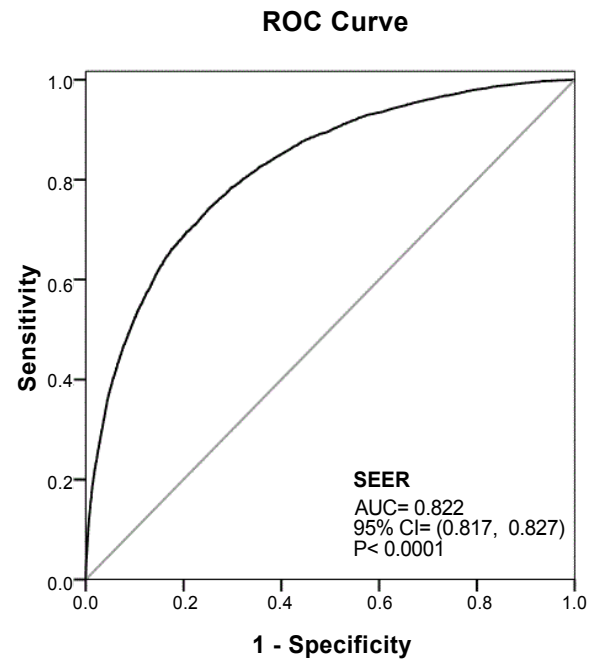

Figure 2: Receiver operator characteristic curves for breast cancer specific mortality in SEER collective.

Taiwan. Our model was well calibrated and provided a high degree of discrimination across different prognostic groups. Although this model provided survival estimation into different prognostic groups, this prognostic model provided also survival prognostication based on the survival in each patient. Furthermore, the model was based on database at a single institution, which can minimize treatment biases may influence the overall and cause data deviation. Therefore, we investigated whether the outcomes predicted using the SEER database and our database were accurate and independent of the method of detection.

The Surveillance, Epidemiology, and End Results (SEER) program provides a large data set of cancer statistics in the United States [3]. The database comprises cancer reports on ~28% of the US population. SEER also collects relevant information on cancer mortality and survival in different areas. The SEER system uses risk factors such as age, comorbidity, ER status, tumor grade, tumor size, and number of lymph nodes. These risk factors, which are assessed for each patient, are integrated into the formula, which is then used to calculate the survival probability of the individual. However, this formula was mainly constructed using American breast cancer data; thus, it may not be suitable for application to Taiwanese women. Our formula exhibited

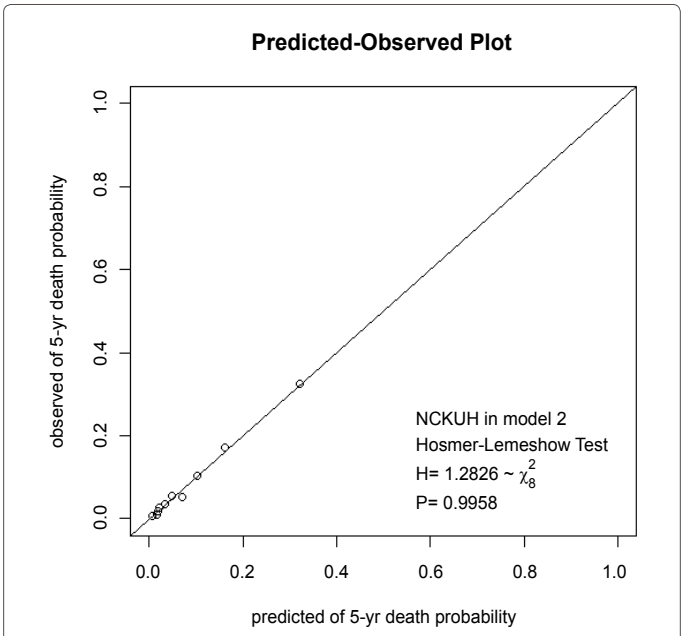

Figure 3: The fitness of NCKUH patients in the model 2 (NCKUH model).

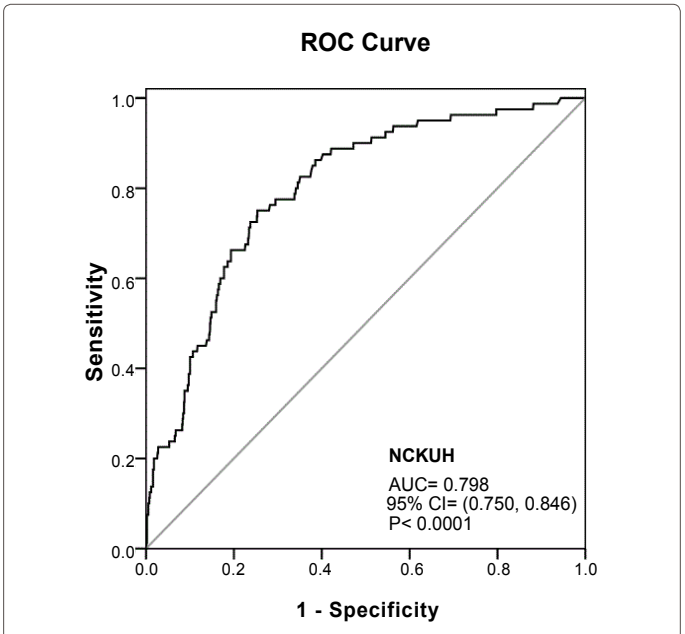

Figure 4: Receiver operator characteristic curves for breast cancer specific mortality in NCKUH breast cancer patients.

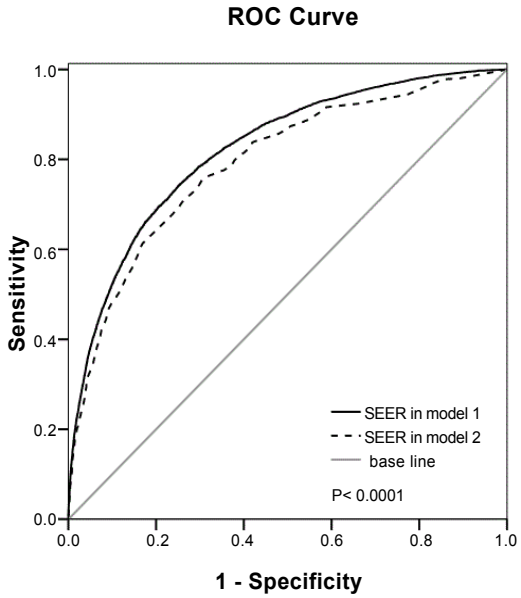

Figure 5: Receiver operator characteristic curves for breast cancer specific mortality in SEER cohort comparing between the different prognostication models.

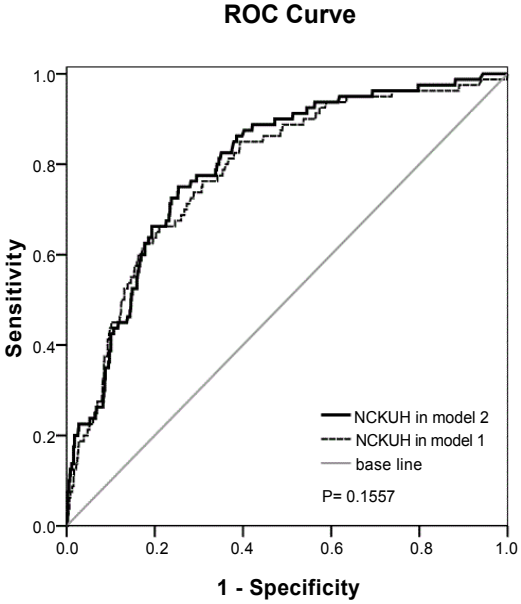

Figure 6: Receiver operator characteristic curves for breast cancer specific mortality in NCKUH cohort comparing between the different prognostication models.

an enhanced performance and was a better survival prediction tool in these women compared with the SEER database. Nevertheless, no objective and quantitative details of the relationships between the risk factors and survival probabilities was provided.

The development of a prognostication and treatment tool that benefited from the many attributes of the SEER database, but one that was specifically tailored to the Taiwanese population, was a key aim of the elaboration of this model. NCKUH cancer database include a single institution, prospective data on postoperative breast cancer patients, which included clinical and pathological information, details of adjuvant treatment, and complete follow up data.

The high predictive accuracy of our model may argument from several factors. First, current model used standard histopathological parameters for input data, which facilitate its application in the clinical setting. Second, the current study is the first to use these prognostic factors as a prognostic and predictive model in Asian breast cancer populations. Finally, this model was validated with an independent and reliable database, such SEER database.

However, some caution should be employed when introducing and interpreting data using our prognostic model. First, this model was assembled with the data from a single institution. The validity of this model should be verified and validated before its application. The variability in survival rates observed for breast cancer patients from different countries seems to support this argument [15,16]. A possible method for overcoming this limitation is development of independent prognostic model from each nation or population. Second, current application of this model needs high human resource cost and is time-consuming. Finally, this study was unable to include human epidermal growth factor 2 (Her-2) receptor status into our model, due to SEER database initiate the collection of HER-2 status since year 2010. Therefore, incorporation of HER-2 status into prognosis calculation will be essential in the future. Eventually, the development of a web-based and user-friendly application tool will be beneficial to facilitating and encouraging its use by physicians for the clinical decision making.

Conclusions

Our predictive model represents a novel method that may provide important information to breast cancer patients after surgical intervention. The SEER database, which is a powerful and independent data set that includes different ethnic groups, proved an ideal reference point for the design of a new prognostic and predictive model. We also emphasize that specific regional or national prognostic models are necessary to improve the choice of appropriate, effective, and individualized therapies for each breast cancer patient.

References

1. Whelan T, Sawka C, Levine M, Gafni A, Reyno L, et al. (2003) Helping patients make informed choices: a randomized trial of a decision aid for adjuvant chemotherapy in lymph node-negative breast cancer. J Natl Cancer Inst 95: 581-587.
2. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, et al. (2001) Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 19: 980-991.
3. Surveillance, Epidemiology, and End Results (SEER)
4. Chang TW, Kuo YL (2010) A model building exercise of mortality risk for Taiwanese women with breast cancer. BMC Med Inform Decis Mak 10: 43.
5. Blamey RW, Pinder SE, Ball GR, Ellis IO, Elston CW, et al. (2007) Reading the prognosis of the individual with breast cancer. Eur J Cancer 43: 1545-1547.
6. De Laurentiis M, De Placido S, Bianco AR, Clark GM, Ravdin PM (1999) A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. Clin Cancer Res 5: 4133-4139.
7. Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thürlimann B (2007) Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. Ann Oncol 18: 1133-1144.
8. Agresti A (2002) Categorical data analysis.
9. Metz CE, Herman BA, Shen JH (1998) Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 17: 1033-1053.
10. Hosmer D, Lemeshow S (1989) Applied logistic regression.

11. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkänen L, et al. (1999) Artificial neural networks applied to survival prediction in breast cancer. Oncology 57: 281-286.

12. Michaelson JS, Chen LL, Bush D, Fong A, Smith B, et al. (2011) Improved web-based calculators for predicting breast carcinoma outcomes. Breast Cancer Res Treat 128: 827-835.

13. Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, et al. (2005) Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol 23: 2716-2725.

14. Ozanne EM, Braithwaite D, Sepucha K, Moore D, Esserman L, et al. (2009) Sensitivity to input variability of the Adjuvant! Online breast cancer prognostic model. J Clin Oncol 27: 214-219.

15. Ghafoor A, Jemal A, Ward E, Cokkinides V, Smith R, et al. (2003) Trends in breast cancer by race and ethnicity. CA Cancer J Clin 53: 342-355.

16. Hausauer AK, Keegan TH, Chang ET, Clarke CA (2007) Recent breast cancer trends among Asian/Pacific Islander, Hispanic, and African-American women in the US: changes by tumor subtype. Breast Cancer Res 9: R90.