Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_______________________________ _______________

Yiran Zhang Date

Statistical Analysis for validating and improving the staging system for breast cancer

By

Yiran Zhang

MSPH

Emory University

Rollins School of Public Health

Department of Biostatistics

____ [Chair’s signature]

Limin Peng

Committee Chair

____ [Member’s signature]

Xiaoxian Li

Committee Member

Statistical Analysis for validating and improving the staging system for breast cancer

By

Yiran Zhang

B.S, South China University of Technology

2016

MSPH, Emory University

Rollins School of Public Health

2018

Thesis Committee Chair: Limin Peng, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science of Public Health

in Department of Biostatistics

2018

Abstract

Statistical Analysis for validating and improving the staging system for breast cancer

By Yiran Zhang

This thesis project is aimed to utilize the National Cancer database (NCDB) to validate and improve the new breast cancer staging system proposed in the 8th edition of the American Joint Committee on Cancer (AJCC) Cancer Staging Manual published in 2017. This staging system incorporates breast cancer biomarkers and will be widely used to determine the breast cancer prognosis worldwide. Our analyses were based on 420,520 breast cancer (BC) cases that were diagnosed from 2010 to 2014 and received the standard treatments. With the primary time-to-event outcome specified as time from diagnosis to all cause death, our univariate and multivariate survival analyses show that age, tumor grade, presence of lymph vascular invasion (LVI), hormonal receptor (HR) and HER2 status, and being triple negative breast cancer (TNBC) status, were significantly associated with the overall survival (all log rank test p-value<0.0001). We further identified that TNBC patients had worse overall survival times than non-TNBC , which included HR+/HER2+, HR+/HER2-, HR-/HER2+ in all stages and sub-stages (all p-value <0.0001). We constructed 4 different staging systems: stage + HR and HER2 status + age group + grade + LVI; stage + TNBC status + age group + grade + LVI; sub-stage + HR and HER2 status + age group + grade +LVI; sub-stage + TNBC status + age group + grade +LVI, and compared their performance based on the Harrell's C-index, Uno's C-statistics and Akaike's information criterion (AIC). Our results indicated that the point system defined based on sub-stage + TNBC status + age + grade +LVI performed the best with the highest Harrell's C-index (0.7316) and Uno's C-statistics (0.6508) and the lowest AIC (488138.91). Our study also suggested that grouping breast cancer subjects by TNBC vs Non-TNBC has similar survival prognostic power to the more detailed BC classification based on HR/HER2 status. Our new staging system improves the prediction of all-cause survival over the traditional anatomic tumor, node and metastasis (TNM) system.

Statistical Analysis for validating and improving the staging system for breast cancer

By

Yiran Zhang

B.S, South China University of Technology

2016

MSPH, Emory University

Rollins School of Public Health

2018

Thesis Committee Chair: Limin Peng, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science of Public Health

in Department of Biostatistics

2018

Acknowledgements

I would like to express my sincere thanks of gratitude to my thesis advisor, Professor Limin Peng, who came up with such a great topic and gave me the opportunity to apply survival analysis to the real medical dataset and practice what I learned from the classes. And I also want to thank for her patience, motivation, enthusiasm and helping me in all the time of research and writing of this thesis. Professor Peng's accomplishments in the academic also drive me to think deep and work hard. I want to thank Dr. Xiaoxian Li for providing me all the clinical information and explain everything in this area with patience. I also want to thank Renjian Jiang to provide a great summary of the dataset and the useful information about the biomarkers in this study.

I also want to show my thanks to all the faculty and staff in Biostatistics department Rollins of Public health at Emory, they provide such a loving and encouraging study environment for us to study here for 2 years. Studying at Emory, I really learned lots of advanced data analysis techniques and fundamental theory in statistics which have significant influence on my future study or career.

Finally, I want to thank my family members for supporting me to study aboard and pursue my degree. My family members always inspired me to study harder and to learn more advanced knowledge, and they always accompany with me to face any difficulties.

Table of Contents

I Introduction

A cancer stage refers to the extent of the cancer. A cancer staging system is intended to inform the status of cancer and provide information to aid in treatment planning or selection. The TNM system is the most widely used cancer staging system; the TNM is the abbreviation of primary tumor (T), regional lymph nodes (N), and distant metastases (M). Each patient has his/her own TNM status and a parallel specific disease stage. Following guidelines such as those of the National Comprehensive Cancer Networks, a clinician usually sets up a treatment plan based on the patient's TNM status [2] [27] [28]. The American Joint Committee on Cancer and the International Union for Cancer Control updates the tumor–node–metastasis (TNM) cancer staging system regularly. The AJCC TNM system has been widely used around the world. However, it has been noted that, some biomarkers may carry additional prognostic information of cancer survival beyond that covered by the current TNM status. This is suggested by the observation that the cancer survival within each TNM stage may vary significantly by the value of these biomarkers.

Recently, many studies begin to examine the effect of primary tumor histologic grade and many other biologic tumor markers that related to prognostic of breast cancer. Those studies indicate that by including these factors, the AJCC TNM system could be refined. For example, Songjie et al. [24] proposed a model that contains miRNA and node status. This model can be used to stratify Triple Negative Breast Cancer (TNBC) patients into different prognostic subgroups for potentially individualized therapy. Jiehua et al. [26] also pointed out that androgen receptor (AR) is a favorable prognostic factors of disease free survival as well as overall survival. In addition, Huang et al. reported the clinical value of Cathepsin-D and Ki-67 index in predicting recurrence [25]. Many studies notice that estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) carry prognostic and predction value in patients with breast cancer [27][28][29]. Also, Li et al [1] shows that, when compared with non-TNBC (including ER positive and HER2 positive breast cancers), TNBC has worse prongosis in every stage.

Recently, Min et al. [2] proposed a novel staging system and suggested that using pathological stage, tumor grade and estrogen receptor (ER) status can contribute to a better predictive model of the 5-year disease specific survival. They built a new breast cancer prognostic staging group (PSG) and used different datasets to validate the result. They also found that adding progesterone receptor (PR) to the system can result in more refined subgroups than that from not using the pathological staging. The newly proposed breast cancer prognostic staging group (PSG) is largely based on studies from MD Anderson cancer center, which showed that incorporating biomarker status and tumor grade into the conventional TNM staging system improved the prognostic power. These studies are from single institution with relatively small patient cohort. With the rapid development of information and big data, more and more large scale datasets become available. Those datasets allow statisticians to extract valuable information on the prognosis of breast cancer. For example, the National Cancer Database (NCDB) is a nationally recognized dataset that represent more than 70 percent newly diagosed cancer cases nationwide and more than 34 million historical records. Using such a national database to build the prediction model for cancer survival can be more representative.

In this thesis, instead of using small patient cohorts, we utilize the national database NCDB to validate and improve the novel staging system presented by Min et al. [2] for predicting overall survival in breast cancer patients who receiving standard care. Furthermore, we aim to simplify this breast cancer prognostic staging group (PSG) by grouping patients into TNBC vs non-TNBC instead of incorporating ER, PR and HER2 status. We develop a new staging point system which accounts for age group, tumor grade, presence of LVI, HR/HER2 status, TNBC status and stage. We compare our new point system with the conventional anatomic TNM system by various statistical tools for evaluating model fits and prediction accuracy. We also investigate the utility of directly using the classification of TNBC vs non-TNBC versus more detail grouping based on BC subtypes for the prognosis purpose. We summarize the study cohort and describe the statistical methods in Chapter II, and present the results in Chapter III. These are followed by discussions in Chapter IV.

II Patients and Methods

2.1 Patient information

We searched the American college of Surgeon's National Cancer database (NCDB) for all female breast cancer patients diagnosed from 2010 and 2014 and identified 2,246,280 cases. We exclude patients who didn't receive any systematic treatment (i.e. HER2+ patients must receive chemotherapy from 2010-2012 and immunotherapy from 2013-2014, ER+ patients must receive at least hormonal treatments, and TNBC patients must receive at least chemotherapy) and patients who had missing information on estrogen receptor (ER), progesterone receptor (PR), HER2 status and overall survival time. Patients who had missing pathologic stage information, or were in stage 0 or stage NOS, are also excluded, or if they. There are total of 420,520 cases that meet our study criteria.

For all the cases included in our study, we collected the following information: age at diagnosis, tumor grade, hormonal receptor (ER or PR) and HER2 status, radiation information, presence of lymph vascular invasion (LVI), overall survival, and pathological stage using the American Joint Committee on Cancer (AJCC) Cancer Staging Manual edition during the year in which the case was diagnosed [2]. The definition of hormonal receptor (HR) is as follows: HR is positive when ER or PR status is positive; HR is negative when both ER and PR status are negative. [1]. We classified breast carcinomas into 4 subtypes by HR and HER2 status x: HR+/HER2+, HR+/HER2-, HR-/HER2+ and HR-/HER2-. The HR-/HER2- subtype was referred to as the TNBC cancer. The subtypes, HR+/HER2+, HR+/HER2- and HR-/HER2+ [2], were considered as non-TNBC. The follow up time was up to 72 months (median=36.3 months and mean=36.8 months). We calculated the numbers and percentages of subjects by each risk factors.

2.2 Survival Analysis

The survival outcome in our study is the overall survival time (OS) calculated as the time from breast cancer diagnosis to death resulting from any reason. The overall survival time is censored for any patient, who was alive at last follow-up visit.

We first evaluate the marginal association of OS times with each of the risk factors considered in this project: Age group, Grade, presence of LVI, HR/HER2 status, TNBC status and Stage. We dichotomize the age at diagnosis as ≤50 (low risk breast cancer group) and >50 (high risk breast cancer group). Then all risk factors are categorical variables. We used the Kaplan-Meier estimator to estimate the survival function curves for different factor levels. We then used log-rank tests to assess whether the OS in the different factor levels are significantly different. The below are briefly introductions of the Kaplan-Meier estimator and the log-rank test.

Kaplan-Meier estimator [3]: Given the number of events (call-cause death), d_i, and the total number of individuals who are at risk, n_i , at the e ith time point t_i, the Kaplan-Meier estimator of the survival function is given by

$$\hat{S}(t) = \prod_{i:t_i \le t} (1 - \frac{d_i}{n_i})$$

The log-rank test: is a nonparametric test to compare the survival distributions between/among different samples. The test statistics of log-rank test is constructed based on the differences between the observed and expected numbers of events (or failures) at all observed event times.

Suppose the risk factor of interest has j factor levels, and let $S_i(t)$ denote the OS function for the factor level i. The null hypothesis is given by

$$H_0: S_1(t) = S_2(t) = \cdots = S_j(t)$$

Denote the number of persons in group j at time t_i by $n_j(t_i)$. Then the expect numbers of failures is given by $E_i = d_i * \frac{n_j(t_i)}{n_i}$. The log-rank test statistic takes the form,

$$Z = \frac{\sum_{i=1}^{k}(d_i - E_i)}{\sqrt{\sum_{i=1}^{k} V_i}} \sim N(0,1) \text{ under } H_0$$

where V_i is the variance of the observed number of events.

Next, within in each pathological stage, we conduct the univariate Cox proportional hazard analysis for breast cancer subtypes. A multivariate Cox regression analysis were further conducted to study the OS across breast cancer subtypes and the OS of TNBC vs Non-TNBC, while adjusting for the significant risk factor (age group, tumor grade and presence of LVI) identified based on univariate analysis. We obtain the hazard ratio estimates and their 95% CI for every univariate and multivariate analysis. We use Wald test to test the significance of a covariate effect in the univariate and multivariate Cox proportional hazard model.

The univariate Cox proportional hazard model can be expressed as:

$$h(t|X) = h_0(t) * e^{\beta_1 X}$$

where $h(t|X)$ represent the hazard function given X, which represents a risk factor/covariate of interest [5].

The Wald test was used to test the significance of a covariate effect in a Cox proportional hazard model. When $\beta_1 = 0$, we will have: $h(t|X) = h_0(t)$, which means there is no effect of covariate X on the hazard function [6]. Thus, the null hypothesis is:

$$H_0 : \beta_1 = 0$$

The Wald test statistics is

$$W=\frac{\widehat{\beta_1}^2}{var(\widehat{\beta_1})} \sim \chi^2(1) \text{ under } H_0$$

where $\widehat{\beta_1}$ is the partial likelihood estimator of β_1.

Our multivariate Cox proportional hazard model takes the form,

$$h(t|Y_i) = h_0(t) * e^{\beta_1 Y_{i1} + \beta_2 Y_{i2} + \cdots + \beta_j Y_{ij}}$$

where $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{ij})$ is a j-dimensional vector of covariates for subject i. The Wald test follows the same rationale as that explained for the univariate Cox regression.

2.3 Building point system

Based on the multivariate Cox analyses, we built 4 different point systems using Age, Grade, LVI, along with the tumor subtype variables (i.e. HR and HER2 status or TNBC vs non-TNBC) and stage or sub-stage. Model 1 contains stage, HR and HER2 status, age group, grade and LVI; Model 2 contains stage, TNBC status, age group, grade and LVI; Model 3 contains sub-stage, HR and HER2 status, age group, grade and LVI. Model 4 contains sub-stage, TNBC status, age group, grade and LVI. We use multivariate Cox hazard regression to fit those 4 models, and obtain the hazard ratio estimates and p-values from the Wald tests. A prognostic score of 0 to 3 was assigned to each factor by considering the magnitude of the hazard ratio (HR) [2].

Specifically, the risk factor level associated with an estimated HR less than 1.15 gets 0 point; the risk factor level associated with an estimated HR between 1.15 to 2.5 and p value <0.05 is assigned 1 point; the risk factor level associated with an estimated HR greater than 2.5 and less than or equal to 6 and a p value <0.05 is assigned 2 points; finally, the risk factor level associated with an estimated HR greater than 6 and a p value <0.05 is assigned 3 points. The overall staging score is calculated as the sum of the total points assigned according to the risk factor values. We evaluate the OS functions stratified by the overall staging score [2]. We apply the prognostic point staging systems developed based on Models 1-4 to the NCDB dataset. Specifically, we first calculate the diagnostic point for each subject based on point assignment rules. Then we fit the Cox model:

$$h(t) = h_o(t) * e^{\beta_1 * point}$$

where "point" denotes the calculated diagnostic point for each subject, which takes values from 0 to 8 with 0 representing the lowest risk of death and 8 indicating the highest risk of death. We shall refer these four models as four point system models.

2.3 Evaluating point system performance

We evaluate the four point system models by Harrell's concordance index (C-index) [7], Uno's concordance index (Uno's C-statistics) [8] and Akaike information criterion (AIC) [9]. Since the traditional C-statistics in logistic regression [11] is designed to deal with binary outcomes, C-statistics cannot handle the time-to-event data. In addition, our dataset has a great proportion of right-censored cases. Therefore, to evaluate the predictive performance of our models, we considered Harrell's concordance index and Uno's concordance index will be used [10]. Those two versions of C-statistics are designed specifically for right-censored data. The major difference between Harrell's method and Uno's method is how they order the survival times in the presence of censoring [12]. Harrell's method provided a direct method by giving up those data which are incomparable due to censoring. If the subject i has survival time T_i and censor time C_i, the Harrell's index can be expressed as followed:

$$C_H = \frac{\sum_{i\neq j} \varDelta_i I\left(X_i < X_j\right) * [I\left(\widehat{\boldsymbol{\beta}'}\boldsymbol{Z_i} > \widehat{\boldsymbol{\beta}'}\boldsymbol{Z_j}\right) + 0.5 * I(\widehat{\boldsymbol{\beta}'}\boldsymbol{Z_i} = \widehat{\boldsymbol{\beta}'}\boldsymbol{Z_j})]}{\sum_{i\neq j} \varDelta_i I\left(X_i < X_j\right)}$$

where $I\ (\cdot)$ is indicator function, $X_i = \min\ (T_i, C_i)$, $\varDelta_i = I\ (X_i = T_i)$, $\widehat{\boldsymbol{\beta}'}$ is the maximum partial likelihood estimator of the vector of true Cox regression parameters $\boldsymbol{\beta}'$, and $\boldsymbol{Z_i}$ is the vector of covariates.

The Limitation of Harrell's method is that the index simply ignores the censored cases. The Uno's index overcame this barrier [8] by modeling the censoring distribution and using it to weight the uncensored observations to avoid the bias from ignoring censored cases [12]. The Uno's index has the following expression:

$$C_U = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \varDelta_i \hat{G}(X_i)^{-2} I\left(X_i < X_j,\ X_i < \tau\ \right) * [I\left(\widehat{\boldsymbol{\beta}'}\boldsymbol{Z_i} > \widehat{\boldsymbol{\beta}'}\boldsymbol{Z_j}\right) + 0.5 * I\left(\widehat{\boldsymbol{\beta}'}\boldsymbol{Z_i} = \widehat{\boldsymbol{\beta}'}\boldsymbol{Z_j}\right)]}{\sum_{i=1}^{n}\sum_{j=1}^{n} \varDelta_i \hat{G}(X_i)^{-2} I\left(X_i < X_j,\ X_i < \tau\ \right)}$$

where τ is user specify time, if not specified, then τ takes the largest event time, $\hat{G}(t)$ is the Kaplan-Meier estimate of the censoring distribution (assuming no covariates).

AIC is also a traditional model selection criterion. The construction of AIC makes the trade-off between the goodness of fit of the model and the simplicity of the model [13] [14]. AIC can be expressed as:

$$AIC = 2k - 2ln(\hat{L})$$

where k is the number of covariates in the model and $\hat{L}$ is the maximum value for the likelihood function.

A higher value of Harrell's C-index or Uno's C-index suggests more accurate survival prediction that the model is expected to produce for OS. A lower AIC value indicates a better balance between the model goodness-of-fit and the model fitness. A P-value below 0.05 was regarded as statistically significant. Data cleaning, data management, data analysis including fitting Cox regression models, conducting Wald tests, calculating the Harrell's C-index, Uno's index and AIC, were performed by SAS (version 9.4, SAS Institute, www.sas.com). Kaplan-Meier plots and log-rank tests were obtained from using R 3.4.3 (The R Foundation for Statistical Computing, www.r-project.org). The specific R packages include "survival" (Therneau & Lumley, www.r-project.org), "KMsurv" (Klein & Jun, www.r-project.org), "survivalMPL" (Dominique-Laurent & Jun, www.r-project.org).

III Results

(All of the tables/figures are in Appendix A)

3.1 Clinicopathological Characteristics of selected cohorts.

Table 1 summarizes the demographic information of the NCDB subjects included in our study. It is shown that the majority of patients are in the high risk age group (>50: 76.63%) and didn't show the presence of Lymph Vascular Invasion (67.2%). Meanwhile, HR+/HER2- subtype carcinomas took up 78.08% of cases and TNBC carcinomas accounted for 11.4% of all the patients. Over half of patients were at the Stage I (53.8%) and took the radiation therapy (65.4%). Most cases were in Grade II (42.55%), which means the tumor was moderately differentiated, moderately well differentiated or intermediate differentiation.

3.2 Results from univariate and multivariate survival analysis.

The univariate analysis shows that all of our risk factors: age group, Grade, LVI, subtype carcinomas, TNBC level, and pathological stage are significantly different across their factor levels (Log-rank test<0.0001, Figure 1). It is also shown that the low risk age group (≤50), lower tumor grade level, and no presence of LVI are significantly associated with better OS (Figure 1, plots A, B, C). The plot D of Figure 1 further shows that OS demonstrates the following pattern across the four different subtype carcinomas: HR+/HER2+<HR+/HER2-<HR-/HER2+<TNBC. and the OS curves for HR+/HER2+ and HR+/HER2- are similar.

Plot E of Figure 1 indicates that TNBC patients have significant worse OS than non-TNBC patients. Figure 1 plot F shows that patients have significant worse OS if they are in higher stage of breast cancer.

Table 2 presents the estimates of hazard ratio and its 95% confidence interval of each non-TNBC subtype carcinomas (i.e. HR+/HER2+, HR+/HER2- and HR-/HER2+) vs TNBC obtained from the Cox regression models which is stratified by stage. While Table 3 shows the estimates of hazard ratio and its 95% confidence interval of each non-TNBC subtype carcinomas vs TNBC obtained from the multivariate Cox regression models which is stratified by stage and adjusts for age group, tumor grade and presence of LVI.

From Table 2, we notice that in each stage and sub-stage, the hazard ratio is less than 1. It indicates that when only consider the subtypes in the cox regression model, the TNBC always has worse overall survival times than any other 3 non-TNBC subtypes (HR+/HER2+, HR+/HER2-, HR-/HER2+) in each stage and sub-stage with all p-values<0.0001. Table 2 also show us the trends that in each stage or sub-stage, the hazard ratio between non-TNBC subtypes to TNBC is increasing in the order HR+/HER2+<HR+/HER2-<HR-/HER2+ which meet the same result shows in Figure 1. Table 3 also show the same result that the non-TNBC patients has better OS compare with TNBC patients (all p-values<0.0001). While, the hazard ratio between non-TNBC to TNBC is decreasing when the stage is increasing. This means when in higher stage of breast cancer, the difference of non-TNBC and TNBC patients on overall survival times gets bigger.

Similar to the univariate analysis results, Table 3 shows that TNBC has worse OS than all other three subtypes (HR+/HER2+, HR+/HER2-, HR-/HER2+) in every stage and sub-stage (most of p-values <0.0001).

Table 4 presents the estimates of hazard ratio and its 95% confidence interval of non-TNBC vs TNBC obtained from the univariate Cox regression models and multivariate Cox regression models adjust for age group, tumor grade and LVI which are stratified by stage. Both univariate analysis and multivariate analysis show TNBC had worse OS than non-TNBC in every stage and sub-stage (all P-values <0.0001, Table 4).

3.3 Construction of prognostic staging systems incorporating all risk factors

Based on the univariate and multivariate analysis results, we understand that all the risk factors we choose (age group, tumor grade, presence of LVI, stage/sub-stage, subtypes/TNBC status) are significantly correlated with the OS. We build 4 different prognostic staging system models following the approach used in Min et al [2].

Tables 5-8 shows the details of how we constructed the four point system models (i.e. Model 1-Model 4) based on slightly different sets of risk factors. The results of multivariate analysis for Model 1-Model 4 accordingly show in the Table 5-8, all of 4 tables present the estimates of hazard ratio and the associated p-value. More specifically, the point assignment was based on the hazard ratio estimates and the associated p-values shown in each table. If HR less than 1.15, 0 point is assigned; if HR between 1.15 to 2.5 and p value <0.05, 1 point is assigned; if HR greater than 2.5 and less than or equal to 6 and a p-value <0.05, 2 points are assigned; finally, if HR greater than 6 and a p-value <0.05, 3 points are assigned. Except for the grade II for model 2 (p-value: 0.0691) and model 4 (p-value: 0.1259), all the hazard ratios are significant: p-value <0.05.

Tables 5-8 also present Harrell's C-statistics, Uno's C-statistics and AIC associated with the four point system models. Model 1 includes stage, HR and HER2 status, age, grade and LVI (Table 5) with Harrell's C-statistics: 0.7407; Uno's C-statistics: 0.6602 and AIC: 533178.81. Model 2 includes stage, TNBC vs non-TNBC, age, grade and LVI (Table 6) with Harrell's C-statistics: 0.7377; Uno's C-statistics: 0.6559 and AIC: 533538.15. Model 3 includes sub-stage, HR and HER2 status, age, grade and LVI (Table 7) with Harrell's C-statistics: 0.7446; Uno's C-statistics: 0.6646 and AIC: 515986.03. Model 4 includes sub-stage, TNBC vs non-TNBC, age, grade and LVI (Table 8) with Harrell's C-statistics: 0.7417; Uno's C-statistics: 0.6606 and AIC: 516342.77. Judging based on the Harrell's C-statistics and Uno's C-statistics, we rank the four point system models (poorest to best) in the order, Model 2, Model 1,Model 4, Model 3.Based on AIC, we would rank the four models (from the poorest to the best) as Model 2, Model 1, Model 4, and <Model 3. Based on these results, we recommend using model 3 as the final model for defining our proposed prognostic point staging system.

3.4 Application of the prognostic point staging system

The OS curves stratified by the prognostic point, Harrell's C-index, Uno's C-index and AIC are shown in the Figure 2 for each of the prognostic point systems constructed in Tables 5-8. The prognostic point system that contains sub-stage, TNBC status, age group, tumor grade and presence of LVI which is developed from model 3 has the smallest AIC: 488138.91. Even though this system Harrell's C-index and Uno's C-index are slightly smaller than the prognostic point system developed from model 4 that contains sub-stage, HR/HER2 status, age group, tumor grade and presence of LVI (Harrell's C-index: 0.7316 vs 0.7325, Uno's C-index: 0.6508 vs 0.6509), the later one has a bigger AIC: 498087.73. The prognostic point system developed from model 1 that contains stage, HR/HER2 status, age group, tumor grade and presence of LVI has the largest AIC: 516853.87 as with relatively smaller Harrell's C-index: 0.7282, and Uno's C-index: 0.6434. The prognostic point system developed from model 2 that contains stage, TNBC status, age group, tumor grade and presence of LVI has a smaller AIC: 507710.93 as with relatively larger Harrell's C-index: 0.7272, and Uno's C-index: 0.6448. The log-rank test for all 4 prognostic systems are strongly significant: all p-value<0.0001. We also constructed the benchmark model for anatomic TNM staging system which only adjust for sub-stage, and it has Harrell's C-index: 0.7160, Uno's C-index: 0.641 and AIC: 688536.49. The lower Harrell's C-index and Uno's C-index, combined with the much larger AIC associated the TNM staging suggests clear improvement resulted from the new prognostic staging systems that take into account HR/HER2 status or TNBC status and the presence of LVI.

IV Discussion

The conventional anatomic TNM tumor staging system predicted prognosis based on tumor size, lymph node status and distant metastasis. For the past decades, it is clear that breast cancer patient survival is greatly affected by the cancer biomarker status. The newly proposed breast cancer PSG incorporated the biomarker status and better predicted patient survival. [15]. The rapid development of cancer biology and biomarker measurements makes the prediction of treatment response more accurate. The previous study examined the effectiveness of point system along with pathological stage, tumor grade and ER status. They found the improvement in guesstimate between stages relative to disease specific survival [2].

In the current analysis, we showed that the age group, tumor grade, presence of LVI, pathological stage/sub-stage and HR/HER2 status/TNBC status were significantly correlated with overall survival. Consistent with other publications, we found that older age, presence of LVI, higher tumor grade (usually grade III), higher pathological stage, being TNBC were associated with poor prognosis [2] [17]. As shown in other studies, age is a stronger predictor for breast cancer, probably becauseolder patients would have more comorbidities and higher chance to have breast cancer.

Bloom et al [18] shows Grade III breast cancer also significantly affected the survival rates of breast cancer patients [19]. In our study, grade III was a significant predictor in both univariate and multivariate analysis. Our analysis also indicated the strongly effect of LVI in OS. HR/HER2 status is shown a significant predictive and prognostic value on OS both in our study and other studies. In our study, we notice that the difference between HR+/HER2+ and HR+/HER2- is small in Figure 1. In the later univariate and multivariate analysis, the hazard ratios between HR+/HER2+ vs TNBC and HR+/HER22 vs TNBC are similar in every stage and sub-stage. Although ER, PR, HER2 value are easily to get in the experiment, but group them together is slightly cumbersome in clinic perspective. Our study indicates that the TNBC status also showed the strong prediction and prognosis in the OS.

We validate the novel staging system for predicting disease specific survival [2] with the national datasets NCDB for predicting overall survival. The novel staging system considered ER, PR and HER2 status in their model while we noticed that TNBC status had strong prediction power from our analysis, we build 4 staging system that all include age group, grade and presence of LVI. The only 2 difference between systems were that using stage or sub-stage and using HR/HER2 status or TNBC status. Our results indicated that only consider TNBC status would have much smaller AIC (Model 2: 507710.93 vs Model 1: 516853.87; Model 4: 488138.91 vs Model 3: 498087.73) and similar in Harrell's C-index and Uno's C-index. We also notice that using sub-stage rather than pathological stage information will produce better predictive performance. Our final recommendation is a simplified staging system that includes sub-stage, TNBC status, age group, grade and presence of LVI to predict the OS.

Our study has several strengths including national wide dataset that has a large number of cases ; succeed in controlling patient-relative, treatment-relative variables; the contemporary nature of the data (modern era chemotherapy and targeted therapies for ER positive and HER2 positive breast cancers). We also came up with an improved staging system based on the observations of our study. We also have several limitations. First, we didn’t perform systematic variable selections; instead we considered predictors suggested by previous publications [2]. Second, the NCDB dataset doesn’t contain the cause-specific mortality information; thus we don’t have the opportunity to evaluate the breast cancer specific survival. Finally, we compared our candidate predictive models by using C-index, AIC and K-M curves. We can potentially apply a more rigorous statistical framework to assess the predictive performance of our models. This constitutes a sensible direction for future work.

References

1. Li, Xiaoxian, Jing Yang, Limin Peng, Aysegul A. Sahin, Lei Huo, Kevin C. Ward, Ruth O'Regan, Mylin A. Torres, and Jane L. Meisel. "Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer." *Breast cancer research and treatment* 161, no. 2 (2017): 279-287.
2. Yi, Min, Elizabeth A. Mittendorf, Janice N. Cormier, Thomas A. Buchholz, Karl Bilimoria, Aysegul A. Sahin, Gabriel N. Hortobagyi et al. "Novel staging system for predicting disease-specific survival in patients with breast cancer treated with surgery as the first intervention: time to modify the current American Joint Committee on Cancer staging system." *Journal of Clinical Oncology* 29, no. 35 (2011): 4654-4661.
3. Dinse, Gregg E. "An Alternative to Efron's Redistribution-of-Mass Construction of the Kaplan—Meier Estimator." *The American Statistician* 39, no. 4 (1985): 299-300.
4. Mantel, Nathan. "Evaluation of survival data and two new rank order statistics arising in its consideration." *Cancer Chemother. Rep.* 50 (1966): 163-170.
5. Breslow, Norman E. "Analysis of survival data under the proportional hazards model." *International Statistical Review/Revue Internationale de Statistique* (1975): 45-57.
6. Cowles, Mary Kathryn. "Modelling survival data in medical research." *Journal of the American Statistical Association* 99, no. 467 (2004): 905-907.
7. Harrell Jr, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. "Evaluating the yield of medical tests." *Jama* 247, no. 18 (1982): 2543-2546.
8. Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data." *Statistics in medicine* 30, no. 10 (2011): 1105-1117.
9. Akaike, Hirotugu. "A new look at the statistical model identification." *IEEE transactions on automatic control* 19, no. 6 (1974): 716-723.
10. Gönen, Mithat, and Glenn Heller. "Concordance probability and discriminatory power in proportional hazards regression." *Biometrika* 92, no. 4 (2005): 965-970.
11. Austin, Peter C., and Ewout W. Steyerberg. "Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable." *BMC medical research methodology* 12, no. 1 (2012): 82.
12. Changbin Guo, Ying So and Woosung Jang. "Evaluating Predictive Accuracy of Survival Models with PROC PHREG." *SAS Institute Inc*, Paper SAS462-2017.

13. Aho, Ken, DeWayne Derryberry, and Teri Peterson. "Model selection for ecologists: the worldviews of AIC and BIC." *Ecology* 95, no. 3 (2014): 631-636.
14. Akaike, Hirotogu. "Information theory and an extension of the maximum likelihood principle." In *Selected Papers of Hirotugu Akaike*, pp. 199-213. Springer, New York, NY, 1998.
15. Sobin LH TNM: principles, history, and relation to other prognostic factors. *Cancer 91 (8 Suppl)*: (2001): 1589-1592.
16. Veronesi, Umberto, Stefano Zurrida, Giuseppe Viale, Viviana Galimberti, Paolo Arnone, and Franco Nolè. "Rethinking TNM: a breast cancer classification to guide to treatment and facilitate research." *The breast journal* 15, no. 3 (2009): 291-295.
17. Frkovic-Grazio, S., and M. Bracko. "Long term prognostic value of Nottingham histological grade and its components in early (pT1N0M0) breast carcinoma." *Journal of clinical pathology* 55, no. 2 (2002): 88-92.
18. Bloom, H. J. G., and W. W. Richardson. "Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years." *British journal of cancer* 11, no. 3 (1957): 359.
19. Horak, Elizabeth R., N. Klenk, R. Leek, S. LeJeune, K. Smith, N. Stuart, A. L. Harris, M. Greenall, and K. Stepniewska. "Angiogenesis, assessed by platelet/endothelial cell adhesion molecule antibodies, as indicator of node metastases and survival in breast cancer." *The Lancet* 340, no. 8828 (1992): 1120-1124.
20. Schoppmann, Sebastian F., Guenther Bayer, Klaus Aumayr, Susanne Taucher, Silvana Geleff, Margaretha Rudas, Ernst Kubista et al. "Prognostic value of lymphangiogenesis and lymphovascular invasion in invasive breast cancer." *Annals of surgery* 240, no. 2 (2004): 306.
21. Varriale, Elisa, Guido Pettinato, Luigi Panico, Giuseppe Pefrella, and II A. Rafaele Bianco. "The prognostic value of lymphatic and blood vessel invasion in operable breast cancer." (1995).
22. Crowe, Joseph P., Nahida H. Gordon, Robert R. Shenk, Robert M. Zollinger, Dorothy J. Brumberg, and Jerry M. Shuck. "Age does not predict breast cancer outcome." *Archives of surgery* 129, no. 5 (1994): 483-488.
23. Foulkes, William D., Ian E. Smith, and Jorge S. Reis-Filho. "Triple-negative breast cancer." *New England journal of medicine* 363, no. 20 (2010): 1938-1948.

24. Shen, Songjie, Qiang Sun, Zhiyong Liang, Xiaojiang Cui, Xinyu Ren, Huan Chen, Xiao Zhang, and Yidong Zhou. "A prognostic model of triple-negative breast cancer based on miR-27b-3p and node status." *PLoS One* 9, no. 6 (2014): e100664.
25. Huang, Liang, Zhebin Liu, Sheng Chen, Yin Liu, and Zhiming Shao. "A prognostic model for triple-negative breast cancer patients based on node status, cathepsin-D and Ki-67 index." *PLoS One* 8, no. 12 (2013): e83081.
26. He, Jiehua, Roujun Peng, Zhongyu Yuan, Shusen Wang, Jiewen Peng, Guinan Lin, Xiaomei Jiang, and Tao Qin. "Prognostic value of androgen receptor expression in operable triple-negative breast cancer: a retrospective analysis based on a tissue microarray." *Medical oncology* 29, no. 2 (2012): 406-410.
27. He, Jiehua, Roujun Peng, Zhongyu Yuan, Shusen Wang, Jiewen Peng, Guinan Lin, Xiaomei Jiang, and Tao Qin. "Prognostic value of androgen receptor expression in operable triple-negative breast cancer: a retrospective analysis based on a tissue microarray." *Medical oncology* 29, no. 2 (2012): 406-410.
28. Edge, Stephen B., and Carolyn C. Compton. "The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM." *Annals of surgical oncology* 17, no. 6 (2010): 1471-1474.
29. Häberle, Lothar, Alexander Hein, Matthias Rübner, Michael Schneider, Arif B. Ekici, Paul Gass, Arndt Hartmann et al. "Predicting triple-negative breast cancer subtype using multiple single nucleotide polymorphisms for breast cancer risk and several variable selection methods." *Geburtshilfe und Frauenheilkunde* 77, no. 06 (2017): 667-678.
30. Uematsu, Takayoshi, Masako Kasami, and Sachiko Yuen. "Triple-negative breast cancer: correlation between MR imaging and pathologic findings." *Radiology* 250, no. 3 (2009): 638-647.
31. Hortobagyi, Gabriel N. "Treatment of breast cancer." *New England Journal of Medicine* 339, no. 14 (1998): 974-984.
32. Ismail-Khan, Roohi. "Estrogen Receptor (ER) and Progesterone Receptor (PR) Positive Breast Cancer." Dr Susan Love Foundation. June 05, 2017. https://www.drsusanloveresearch.org/estrogen-receptor-er-and-progesterone-receptor-pr-positive-breast-cancer.
33. Chun, Christina. "HER2-Positive Breast Cancer: Survival Rates and Other Statistics." Healthline. October 31, 2014. https://www.healthline.com/health/breast-cancer/her2-positive-survival-rates-statistics.

34. Quamos, Hope. "HER2-Positive Breast Cancer Chemotherapy." Healthline. September 22, 2015. https://www.healthline.com/health/breast-cancer/chemotherapy-for-her2-positive-breast-cancer#1.
35. Ryan, P. D., N. M. Tung, S. J. Isakoff, M. Golshan, A. Richardson, A. D. Corben, B. L. Smith, R. Gelman, E. P. Winer, and J. E. Garber. "Neoadjuvant cisplatin and bevacizumab in triple negative breast cancer (TNBC): safety and efficacy." *Journal of Clinical Oncology* 27, no. 15_suppl (2009): 551-551.
36. Mahamodhossen, Yashin A., Wei Liu, and Zhou Rong-Rong. "Triple-negative breast cancer: new perspectives for novel therapies." *Medical oncology* 30, no. 3 (2013): 653.
37. Albergaria, André, Sara Ricardo, Fernanda Milanezi, Vítor Carneiro, Isabel Amendoeira, Daniella Vieira, Jorge Cameselle-Teijeiro, and Fernando Schmitt. "Nottingham Prognostic Index in triple-negative breast cancer: a reliable prognostic tool?" *BMC cancer* 11, no. 1 (2011): 299.

Appendix A

Table 1: Demographic and clinic pathological characteristics

Characteristics	Cohort=420,520 n	%
Stage		
I	226257	53.8
IA	202349	48.12
IB	12361	2.94
Unknown	11547	2.75
II	140098	33.32
IIA	95537	22.72
IIB	42435	10.09
Unknown	2126	0.51
III	45625	10.85
IIIA	28782	6.84
IIIB	4475	1.06
IIIC	11799	2.81
Unknown	569	0.14
IV	8540	2.03
Lymph Vascular Invasion		
No	281815	67.02
Yes	79836	18.99
Unknown	58869	14
Radiation		
No	143613	34.15
Yes	275007	65.4
Unknown	1900	0.45
Sub-type		
HR+/HER2-	328356	78.08
HR+/HER2+	29101	6.92
HR-/HER2+	15122	3.6
TNBC	47941	11.4
TNBC status		
Non-TNBC	372579	88.6
TNBC	47941	11.4
Grade		
I	92688	22.04
II	178935	42.55
III	122494	29.13
Age Group		
≤50 (Low risk)	98260	23.37
>50 (High risk)	322260	76.63

Abbreviations: HR: hormonal receptor; TNBC: triple negative breast cancer

Table 2. Stratified Univariate analysis of correlation of subtypes vs TNBC with overall survival

	Analysis results			
Stage	**Hazard ratio**	**95% CI**		**P-value**
I				
HR+/HER2+ vs TNBC	0.36	0.32	0.41	<0.0001
HR+/HER2- vs TNBC	0.61	0.57	0.65	<0.0001
HR-/HER2+ vs TNBC	0.66	0.57	0.76	<0.0001
II				
HR+/HER2+ vs TNBC	0.26	0.24	0.29	<0.0001
HR+/HER2- vs TNBC	0.43	0.41	0.45	<0.0001
HR-/HER2+ vs TNBC	0.54	0.49	0.60	<0.0001
III				
HR+/HER2+ vs TNBC	0.17	0.15	0.19	<0.0001
HR+/HER2- vs TNBC	0.26	0.25	0.27	<0.0001
HR-/HER2+ vs TNBC	0.37	0.34	0.41	<0.0001
IV				
HR+/HER2+ vs TNBC	0.21	0.18	0.24	<0.0001
HR+/HER2- vs TNBC	0.36	0.33	0.38	<0.0001
HR-/HER2+ vs TNBC	0.40	0.36	0.46	<0.0001
IA				
HR+/HER2+ vs TNBC	0.35	0.30	0.41	<0.0001
HR+/HER2- vs TNBC	0.63	0.59	0.68	<0.0001
HR-/HER2+ vs TNBC	0.67	0.57	0.78	<0.0001
IB				
HR+/HER2+ vs TNBC	0.33	0.21	0.50	<0.0001
HR+/HER2- vs TNBC	0.38	0.30	0.48	<0.0001
HR-/HER2+ vs TNBC	0.43	0.26	0.70	<0.0001
IIA				
HR+/HER2+ vs TNBC	0.29	0.25	0.33	<0.0001
HR+/HER2- vs TNBC	0.48	0.45	0.50	<0.0001
HR-/HER2+ vs TNBC	0.55	0.48	0.62	<0.0001
IIB				
HR+/HER2+ vs TNBC	0.22	0.19	0.26	<0.0001
HR+/HER2- vs TNBC	0.34	0.32	0.37	<0.0001
HR-/HER2+ vs TNBC	0.51	0.44	0.59	<0.0001
IIIA				
HR+/HER2+ vs TNBC	0.17	0.15	0.19	<0.0001
HR+/HER2- vs TNBC	0.24	0.23	0.26	<0.0001
HR-/HER2+ vs TNBC	0.34	0.30	0.39	<0.0001
IIIB				
HR+/HER2+ vs TNBC	0.25	0.19	0.32	<0.0001

HR+/HER2- vs TNBC	0.37	0.32	0.41	<0.0001
HR-/HER2+ vs TNBC	0.39	0.32	0.49	<0.0001
IIIC				
HR+/HER2+ vs TNBC	0.16	0.13	0.19	<0.0001
HR+/HER2- vs TNBC	0.27	0.25	0.29	<0.0001
HR-/HER2+ vs TNBC	0.37	0.32	0.43	<0.0001

Abbreviation: CI: confidence interval; HR: hormonal receptor; TNBC: triple negative breast cancer

Table 3. Stratified Multivariate analysis of correlation of subtypes vs TNBC with overall survival adjusted for age, grade and LVI

	Analysis result			
Stage	**Hazard ratio**	**95% CI**		**P-value**
I				
HR+/HER2+ vs TNBC	0.42	0.36	0.49	<0.0001
HR+/HER2- vs TNBC	0.67	0.62	0.73	<0.0001
HR-/HER2+ vs TNBC	0.63	0.54	0.75	<0.0001
II				
HR+/HER2+ vs TNBC	0.29	0.26	0.33	<0.0001
HR+/HER2- vs TNBC	0.50	0.47	0.53	<0.0001
HR-/HER2+ vs TNBC	0.51	0.46	0.57	<0.0001
III				
HR+/HER2+ vs TNBC	0.20	0.18	0.22	<0.0001
HR+/HER2- vs TNBC	0.34	0.32	0.37	<0.0001
HR-/HER2+ vs TNBC	0.35	0.32	0.39	<0.0001
IV				
HR+/HER2+ vs TNBC	0.19	0.16	0.23	<0.0001
HR+/HER2- vs TNBC	0.36	0.32	0.40	<0.0001
HR-/HER2+ vs TNBC	0.36	0.30	0.42	<0.0001
IA				
HR+/HER2+ vs TNBC	0.40	0.34	0.47	<0.0001
HR+/HER2- vs TNBC	0.68	0.62	0.74	<0.0001
HR-/HER2+ vs TNBC	0.64	0.54	0.76	<0.0001
IB				
HR+/HER2+ vs TNBC	0.43	0.26	0.69	0.00024
HR+/HER2- vs TNBC	0.49	0.36	0.67	<0.0001
HR-/HER2+ vs TNBC	0.44	0.25	0.78	0.0023
IIA				
HR+/HER2+ vs TNBC	0.32	0.28	0.37	<0.0001
HR+/HER2- vs TNBC	0.54	0.50	0.58	<0.0001
HR-/HER2+ vs TNBC	0.53	0.46	0.62	<0.0001
IIB				

HR+/HER2+ vs TNBC	0.25	0.21	0.29	<0.0001
HR+/HER2- vs TNBC	0.43	0.39	0.47	<0.0001
HR-/HER2+ vs TNBC	0.48	0.40	0.56	<0.0001
IIIA				
HR+/HER2+ vs TNBC	0.20	0.17	0.23	<0.0001
HR+/HER2- vs TNBC	0.32	0.30	0.35	<0.0001
HR-/HER2+ vs TNBC	0.32	0.27	0.37	<0.0001
IIIB				
HR+/HER2+ vs TNBC	0.31	0.24	0.41	<0.0001
HR+/HER2- vs TNBC	0.47	0.40	0.55	<0.0001
HR-/HER2+ vs TNBC	0.37	0.28	0.48	<0.0001
IIIC				
HR+/HER2+ vs TNBC	0.17	0.14	0.21	<0.0001
HR+/HER2- vs TNBC	0.35	0.32	0.39	<0.0001
HR-/HER2+ vs TNBC	0.37	0.31	0.44	<0.0001

Abbreviation: TNBC: triple negative breast cancer; CI: confidence interval; HR: hormonal receptor

Table 4. Univariate and multivariate analysis of correlation of TNBC vs non-TNBC with overall survival adjusted for age, grade and LVI

	Univariate analysis				Multivariate analysis			
Stage	**Hazard ratio**	**95% CI**		**P-value**	**Hazard ratio**	**95% CI**		**P-value**
I	0.60	0.56	0.64	<0.0001	0.64	0.59	0.69	<0.0001
II	0.41	0.40	0.43	<0.0001	0.47	0.44	0.49	<0.0001
III	0.26	0.25	0.27	<0.0001	0.32	0.30	0.34	<0.0001
IV	0.34	0.32	0.37	<0.0001	0.33	0.29	0.36	<0.0001
IA	0.61	0.57	0.66	<0.0001	0.64	0.58	0.70	<0.0001
IB	0.38	0.30	0.47	<0.0001	0.48	0.35	0.64	<0.0001
IIA	0.46	0.43	0.49	<0.0001	0.50	0.47	0.54	<0.0001
IIB	0.34	0.31	0.36	<0.0001	0.40	0.37	0.44	<0.0001
IIIA	0.24	0.23	0.26	<0.0001	0.30	0.28	0.32	<0.0001
IIIB	0.35	0.32	0.40	<0.0001	0.43	0.37	0.49	<0.0001
IIIC	0.26	0.24	0.28	<0.0001	0.32	0.29	0.35	<0.0001

Abbreviation: TNBC: triple negative breast cancer; CI: confidence interval;

Table 5. Model 1: stage + (HR and HER2 Status) + age + grade + LVI

Model 1: C-statistics: 0.7407; Uno's C-statistics: 0.6602; AIC: 533178.81			
	Multivariate Analysis		
Factor	**Hazard ratio**	**P-value**	**Assigned points**
Stage			
I	Reference		0
II	1.71	<.0001	1
III	4.992	<.0001	2
IV	13.962	<.0001	3
Sub-type			
HR+/HER2-	Reference		0
HR+/HER2+	0.649	<.0001	0
HR-/HER2+	1.155	<.0001	1
TNBC	2.749	<.0001	2
Age			
≤50 (Low risk)	Reference		0
>50 (High risk)	1.857	<.0001	1
Grade			
I	Reference		0
II	1.065	0.004	0
III	1.564	<.0001	1
LVI			
No	Reference		0
Yes	1.38	<.0001	1

Abbreviation: HR: hormonal receptor; LVI: lymph vascular invasion;
C-index: Harrell’s concordance index; Uno’s C-index: Uno’s concordance index; AIC: Akaike’s information criterion.
The points were assigned based on the hazard ratio. A 0 point was assigned when the hazard ratio was <1.15; point 1: 1.15-2.5; point 2: >2.5-6; point 3: >6.

Table 6. Model 2: stage + TNBC + age + grade + LVI

Model 2: C-statistics: 0.7377; Uno's C-statistics: 0.6559; AIC: 533538.15			
	Multivariate Analysis		
Factor	**Hazard ratio**	**P-value**	**Assigned points**
Stage			
I	Reference		0
II	1.697	<.0001	1
III	4.959	<.0001	2
IV	13.783	<.0001	3
Sub-type			
Non-TNBC	Reference		0

TNBC	2.825	<.0001	2
Age			
≤50 (Low risk)	Reference		0
>50 (High risk)	1.885	<.0001	1
Grade			
I	Reference		0
II	1.04	0.0691	0
III	1.466	<.0001	1
LVI			
No	Reference		0
Yes	1.37	<.0001	1

Abbreviation: HR: hormonal receptor; LVI: lymph vascular invasion; TNBC: triple negative breast cancer. C-index: Harrell's concordance index; Uno's C-index: Uno's concordance index; AIC: Akaike's information criterion.

The points were assigned based on the hazard ratio. A 0 point was assigned when the hazard ratio was <1.15; point 1: 1.15-2.5; point 2: >2.5-6; point 3: >6.

Table 7. Model 3: Sub-stage + (HR and HER2 Status) + age + grade + LVI

Model 3: C-statistics: 0.7446; Uno's C-statistics: 0.6647; AIC: 515986.03 ;			
	Multivariate Analysis		
Factor	**Hazard ratio**	**P-value**	**Assigned points**
Stage			
IA	Reference		0
IB	1.101	0.0594	0
IIA	1.526	<.0001	1
IIB	2.256	<.0001	1
IIIA	4.046	<.0001	2
IIIB	7.278	<.0001	3
IIIC	7.125	<.0001	3
IV	14.43	<.0001	3
Sub-type			
HR+/HER2-	Reference		0
HR+/HER2+	0.645	<.0001	0
HR-/HER2+	1.141	<.0001	0
TNBC	2.77	<.0001	2
Age			
≤50 (Low risk)	Reference		0
>50 (High risk)	1.832	<.0001	1
Grade			
I	Reference		0
II	1.059	0.0097	0
III	1.552	<.0001	1

LVI			
No	Reference		0
Yes	1.311	<.0001	1

Abbreviation: HR: hormonal receptor; LVI: lymph vascular invasion;
C-index: Harrell's concordance index; Uno's C-index: Uno's concordance index; AIC: Akaike's information criterion.
The points were assigned based on the hazard ratio. A 0 point was assigned when the hazard ratio was <1.15; point 1: 1.15-2.5; point 2: >2.5-6; point 3: >6.

Table 8. Model 4: Sub-stage + TNBC + age + grade + LVI

Model 4: C-statistics: 0.7417; Uno's C-statistics: 0.6606; AIC: 516342.77;			
	Multivariate Analysis		
Factor	**Hazard ratio**	**P-value**	**Assigned points**
Stage			
IA	Reference		0
IB	1.101	0.0599	0
IIA	1.515	<.0001	1
IIB	2.238	<.0001	1
IIIA	4.016	<.0001	2
IIIB	7.246	<.0001	3
IIIC	7.072	<.0001	3
IV	14.229	<.0001	3
Sub-type			
Non-TNBC	Reference		0
TNBC	2.85	<.0001	2
Age			
≤50 (Low risk)	Reference		0
>50 (High risk)	1.86	<.0001	1
Grade			
I	Reference		0
II	1.035	0.1259	0
III	1.452	<.0001	1
LVI			
No	Reference		0
Yes	1.302	<.0001	1

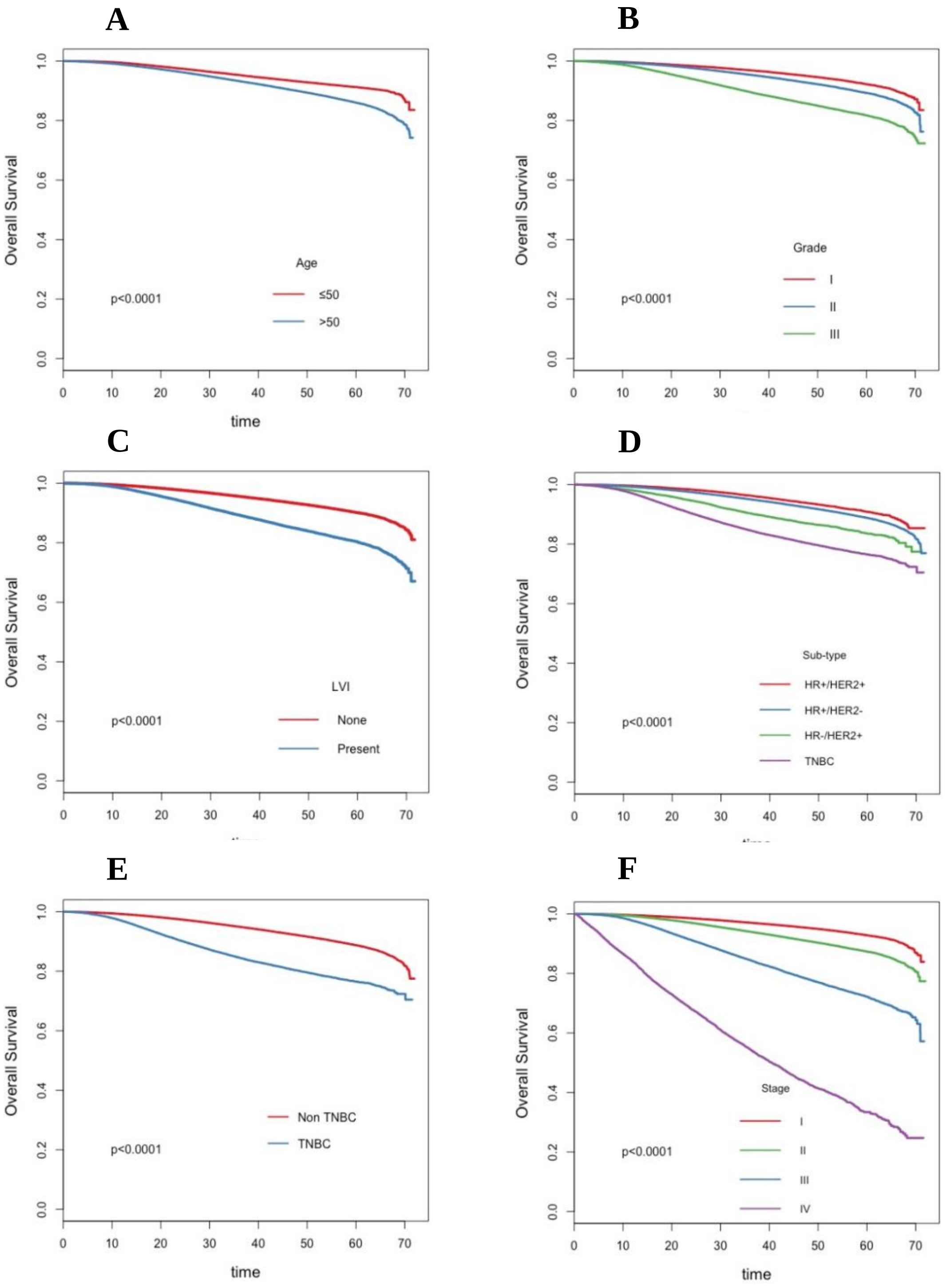
Abbreviation: HR: hormonal receptor; LVI: lymph vascular invasion; TNBC: triple negative breast cancer.
C-index: Harrell's concordance index; Uno's C-index: Uno's concordance index; AIC: Akaike's information criterion.
The points were assigned based on the hazard ratio. A 0 point was assigned when the hazard ratio was <1.15; point 1: 1.15-2.5; point 2: >2.5-6; point 3: >6.

Table 9. C-statistics and AIC for each prognostic staging system model

	C-index	Uno's C-index	AIC
Model 1: Stage + (HR and HER2 Status) + age + grade + LVI	0.7282	0.6434	516853.87
Model 2: Stage + TNBC + age + grade + LVI	0.7272	0.6448	507710.93
Model 3: Sub-stage + (HR and HER2 Status) + age + Grade + LVI	0.7325	0.6508	498087.73
Model 4: Sub-stage + TNBC + age + grade + LVI	0.7316	0.6509	488138.91
Anatomic TNM system	0.716	0.641	688536.49

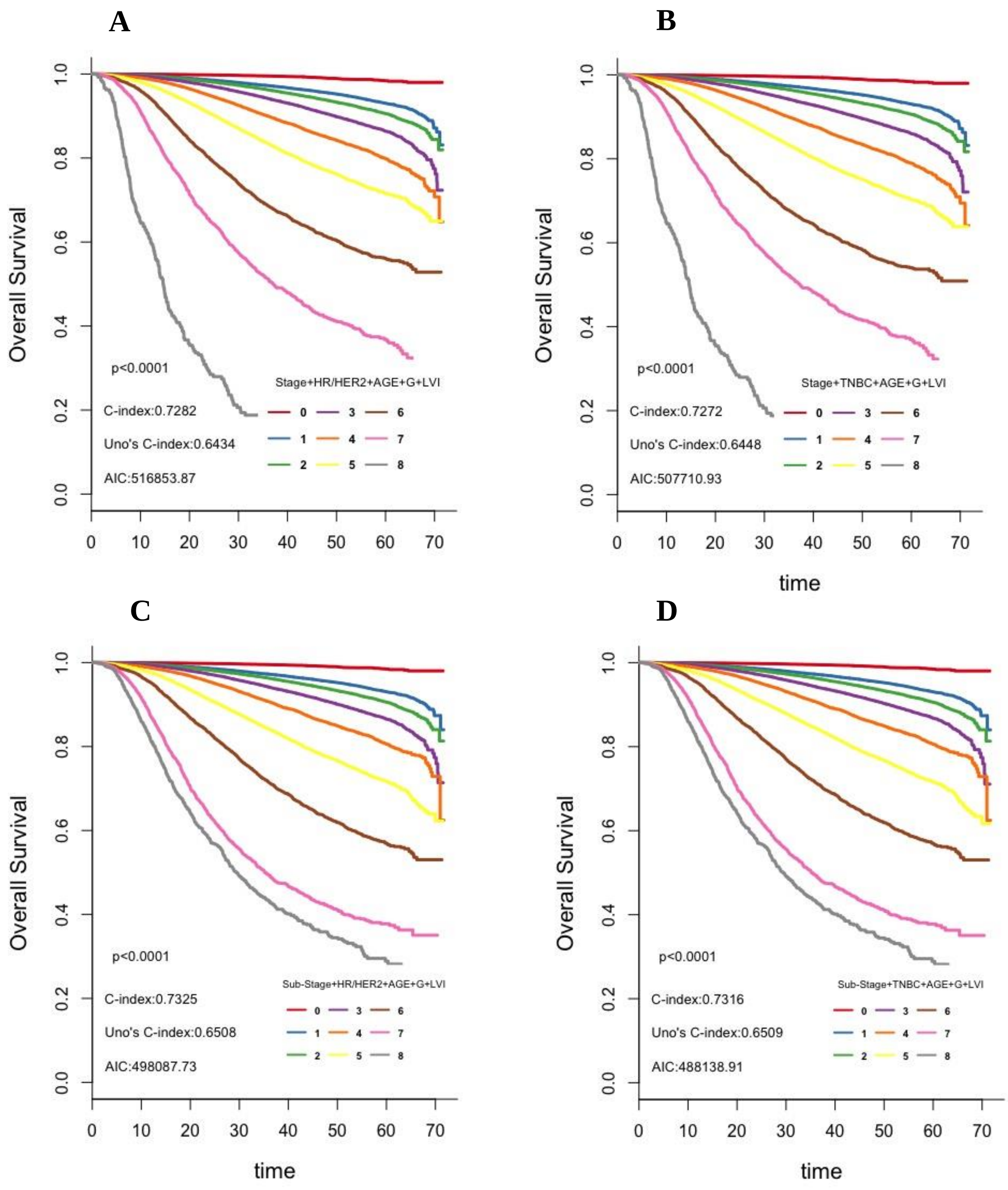
Abbreviation: HR: hormonal receptor; LVI: lymphovascular invasion; TNBC: triple negative breast cancer; C-index: Harrell’s concordance index; Uno’s C-index: Uno’s concordance index; AIC: Akaike’s information criterion

Figure 1. Kaplan-Meier Curves and Log-rank test results for risk factors



Abbreviation: HR: hormonal receptor; LVI: lymphovascular invasion; TNBC: triple negative breast cancer;

Figure 2. Kaplan-Meier Curves for 4 staging systems.



Abbreviation: HR: hormonal receptor; LVI: lymph vascular invasion; G: Grade; TNBC: triple negative breast cancer.
C-index: Harrell's concordance index; Uno's C-index: Uno's concordance index; AIC: Akaike's information criterion.