A New Survival Prediction Model for Patients with Synchronous Colorectal Carcinomas Based on SEER

YuXin Xu
fujian medical university
XiaoJie Wang
Xiehe Affiliated Hospital of Fujian Medical University
Ying Huang
Xiehe Affiliated Hospital of Fujian Medical University
DaoXiong Ye
Xiehe Affiliated Hospital of Fujian Medical University
Pan Chi (chipan363@163.com)
Fujian Medical University

Research

Keywords: Colorectal cancer (CRC), synchronous colorectal carcinoma (SCC), Prediction Model, National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER)

Posted Date: April 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-20493/v1

Abstract

Introduction The nomogram for postoperative prediction of overall survival (OS) in patients' synchronous colorectal carcinomas (SCC) was developed and validated by LASSO regression combined with COX regression. Methods The data was obtained from the SEER database of patients diagnosed with colorectal cancer (CRC) more than one time between 2004 and 2013. The cut-off points for the continuous variable were identified by the K-adaptive partitioning algorithm and x-tile software. Using LASSO regression combined with the Cox regression, a model for predicting the overall survival of SCC was built, internally and externally validated, and measured through a calibration curve, C-index, AIC, BIC, IDI, NRI, timeROC, timeAUC, and decision curve analysis (DCA), and results compared to the model developed by the Cox regression. Results Patients with SCC were found to be older, more often men, and likely to have a depth of invasion by T3. In addition, there were no significant differences between the model developed by LASSO regression combined with Cox regression and the Cox regression in the calibration curve, C-index, AIC, BIC, IDI, NRI, timeROC, and DCA. Besides, the model developed by LASSO regression combined with Cox regression was found to perform better than the Cox regression in the timeAUC. Moreover, the model developed by LASSO regression combined with Cox regression showed good calibration, C-index, AIC, BIC, IDI, NRI, timeROC, timeAUC and had a larger net benefit compared to both the first time TNM staging and the combination of two times TNM staging. Discussion This present study indicates that a close follow-up of older patients, male, and T3 should be made. LASSO regression combined with COX regression decreases the variables of the model, avoids overfitting and collinearity and has clinical significance.

Introduction

Colorectal cancer (CRC) is the third common cancer and ranks third as a cause of cancer-related death for males in America[1, 2]. The definition of multiple primary colorectal carcinomas (MPCC) is the presence of 2 or more primary invasive adenocarcinomas diagnosed in patients. Synchronous colorectal carcinomas (SCC) is identified as the second invasive adenocarcinomas diagnosis within 6 months after the first[3]. Metachronous colorectal carcinomas (MCC) is identified as the second invasive adenocarcinomas diagnosis after more than 6 months after the first[4]. Among patients suffering from colorectal cancer, synchronous colorectal carcinomas contribute 1–8% [5].

Patients with familial adenomatous polyposis, hereditary nonpolyposis colorectal cancer (HNPCC), hereditary non-polyposis colorectal cancer, serrated polyps/hyperplastic polyposis and inflammatory bowel diseases (ulcerative colitis and Crohn's disease) have a higher risk of synchronous colorectal carcinomas. These predisposing factors contribute to slightly more than 10% of synchronous colorectal carcinomas[6]. Compared to solitary colorectal cancer, synchronous colorectal carcinomas are more common in the right colon and sigmoid colon. During the pathological examination, synchronous colorectal carcinomas are usually found to be mucinous adenocarcinomas. Most of the patients with synchronous colorectal carcinoma have two cancers but only 6 cases have been reported. Compared to

patients with a solitary colorectal carcinoma, patients with synchronous colorectal carcinoma have a higher percentage of microsatellite instability. Also, compared to patients with solitary tumors, synchronous colorectal carcinomas are enriched with MSI-H tumors, particularly those arising from SSAs. Moreover, MSI contributes to the improvement of the overall survival of patients with synchronous tumors. It is notable that SSA-associated SCC has a predilection for elderly women. Besides, SSA is associated with a favorable prognosis and is more likely to be MSI-H and BRAF V600E positive[3]. What's more, there is no appreciable differences between patients with synchronous tumors and single neoplasm in survival when compared to individuals who had single neoplasms. In contrast, individuals who had metachronous carcinomas have been observed to show a poor clinical outcome after the development of the second carcinoma[7]. Therefore, patients with colorectal cancer must be fully studied endoscopically[8].

Currently, there only exists literature of mostly small series (< 80 patients) in which epidemiology and clinicopathology are described[3–7, 9]. However, there are few reports on the impact factors of synchronous colorectal carcinoma's overall survival and formulation of prognostic models. In this study, we evaluated the impact factors of synchronous colorectal carcinomas on the overall survival and made a prognostic model with a large cohort of patients.

In this study, our aim was to develop and validate a nomogram based on treatment variables, surgical variables, clinical characteristics and tumor characteristics to predict the survival of synchronous colorectal carcinomas patients. The data was obtained from the population-based Surveillance, Epidemiology, and End Results (SEER) database which contains a large sample size and has a long follow-up time. However, if the prediction model is associated with the first and second-time treatment variables, surgical variables and tumor characteristics, there would be multiple mutual linear problems. Furthermore, because of incorporating too many variables, there may be over-fitting in predicting model. For these reasons, we selected the least absolute shrinkage and selection operator (LASSO) method to deal with the above concerns. In order to determine whether the prediction model fitted with LASSO combined with Cox was better, we compared the LASSO model (fit by Cox regression after variables selection by LASSO Cox regression) to the COX model (fit by Cox regression), TNM model (established in first time T,N,M grade) and TTNNMM model (established in first and second times T,N,M grade) by considering the discrimination and calibration.

Methods

Data Source

We identified the survivors from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) 13-registry database by analyzing patients diagnosed from 2004 to 2013. SEER is a publicly available, nationally representative, population-based cancer database that contains more than 8 million cancer cases, with data that spans 4 decades and covers 28% of the United States population. It is considered a valid source of cancer incidence and survival data in the United States. In addition, the

SEER has developed and maintained high-quality, validated data on causes of death among cancer survivors, providing insight into relative and cause-specific deaths in this population[10, 11]. Our study was determined and it exempted the data from Colorectal Surgery Union Hospital in Fuzhou because publicly available de-identified data were used. Data was retrieved using SEER*Stat 8.3.5 (Surveillance Research Program, National Cancer Institute, Bethesda, MD).

Patients

Patients over 18 years old who were diagnosed with colorectal carcinoma between 2004 and 2013 with surgery were initially analyzed. Patients diagnosed with colorectal carcinoma less than twice were excluded to explore synchronous colorectal carcinomas. Patients who survival time was unknown were excluded to explore the epidemiology, pathogenesis and factors that influenced the survival of synchronous colorectal carcinomas. Patients with an unknown grade of the tumor, unknown T stage, unknown N stage, and unknown M stage were excluded for further comparison of the feasibility of the TNM model and TTNNMM model. Patients with unknown prognostic characteristics (including race, tumor size and location) were also excluded. The clinicopathologic variables were then collected from the SEER 13 database, including gender, race, sex, delta t, months survived and first and second times age, marital, location of tumor, TNM staging[12], histologic grade, number of lymph nodes examined, number of positive lymph nodes, tumor size, radiation sequence, chemotherapy, and surgical related variables. Then, patients who had colorectal cancer more than 3 times or multiple metachronous primary carcinomas were excluded. Lastly, we excluded patients who survived less than 1 month or other variable unknown (Figure 1).

In the construction of the survival predicting model, the internal cohort included patients from SEER database, while the external validation cohort consisted of patients from Colorectal Surgery Union Hospital in Fuzhou database.

Statistical Analysis

Statistical analysis was carried out with R software (version 3.4.2; http://www.Rproject.org) and SPSS (Statistical Product and Service Solutions, version 22.0). The packages in R used in this study are as follows. statistical significance levels were all two-sided, with statistical significance set at .05.

Variables Selection and Models Constructing

The least absolute shrinkage and selection operator (LASSO) method, which are suitable for the regression of high-dimensional data[13, 14], were used to select the most useful predictive variables from the primary data set. The “glmnet” package was used to perform the LASSO Cox regression model analysis[15]. To compare the differences in the Cox method and LASSO combined with Cox method, we separately used the Cox regression or LASSO combined with Cox regression to construct models. The COX model was selected and constructed using the internal cohort by backward Cox analysis using Akaike’s information criterion (AIC) selection criteria and the best model was selected with the least AIC[16-

[18]. The LASSO model was selected and constructed using the internal cohort by LASSO combined with Cox regression. The TNM model was established using the internal cohort by the first time T, N and M staging and that of TTNNMM model was established using the internal cohort by the first and second times T, N and M staging[19].

Compare Models

The LASSO model using LASSO combined with Cox regression and COX model using backward Cox analysis was first internally validated in the internal cohort using a bootstrap method (1,000 bootstraps resamples) and then externally validated in the external cohorts. The 3- and 5-year OS calibration of the LASSO model and COX model were performed by comparing the observed survival with the predicted survival in the internal and external cohorts. Then for survival testing with the LASSO model and COX model of the specificity, time-dependent receiver operating characteristic (timeROC) curves were estimated for two cohorts by inverse probability of censoring weighting estimators (KM-weight) at 3-, and 5- year[20, 21]. Sequential AUCs were compared among the four models using identically and independently distributed representations of the AUC estimators[22]. Also, the overall prognostic performance of the four models was assessed using the Bayesian Information Criterion (BIC) via bootstrap-resampling analysis. Lastly, four models were evaluated with AIC, C-index[23], the net reclassification improvement (NRI)[24, 25] and integrated discrimination improvement (IDI)[26].

Clinical Use

The net benefit and clinical usefulness of the four models above were estimated with decision curve analysis (DCA) throughout the whole cohort[27].

Nomogram for a visualization model

For the purpose of illustration and clinical applicability, we created a nomogram based on the LASSO model. In the nomogram, model-based score points for each predictor variable category were displayed, which has to be summarized for any individual patient. From the resulting total number of points, the corresponding predicted survival probabilities from the nomogram could be easily read.

Results

Clinical Characteristics

From the data obtained from 2004 to 2013, 4616 patients with synchronous colorectal carcinomas (SCC) in the SEER database were found. Patient characteristics are shown in Table 1. There are significant correlations in age and slight correlation in pN, pM, examined lymph nodes, Surg Prim Site, and chemotherapy between the twice synchronous colorectal. Patients with SCC were mostly older (>65 years), more often men, and likely to have a depth of invasion by T3. Tumors were mostly situated in the cecum, ascending colon and sigmoid colon.

Results of the selected variables with Cox regression and LASSO combined with Cox regression are listed in Table 2. Table 2 indicates that the age of the first time SSC diagnosis, sex, first time size, first time surgery, second time marital, second time grade, second time chemotherapy, first and second times pT, pN, pM, regional nodes examined and site of disease was significantly associated with overall survival (OS) by Cox regression. Table 2 also indicates that the age of first time SSC diagnosis, sex, second-time chemotherapy, first and second times pT, pN, pM, regional nodes examined were significantly associated with overall survival (OS) by LASSO combined with Cox regression.

Results from the relation between first and second times pT, pN, pM, grade and regional nodes are listed in Table 1.

Predictive Variable Selection

31 variables were reduced to 11 or 16 potential predictors on the basis of 4616 patients by LASSO combined with Cox regression or Cox regression in the internal cohort (Figure 2A, 2B, 2C) and were featured with nonzero coefficients in the LASSO Cox regression model or the minute AIC in Cox regression model (Figure 2D).

Development of COX Model and LASSO Model

The multivariable regression model for age, sex, marital, race, site, pT, pN, pM, radiation, chemotherapy, surgery, nodes examined, etc. were included in the Cox regression after variables were selected by the LASSO Cox regression or Cox regression. We showed hazard ratios with 95% CIs for covariates which are included in Table 2.

Apparent Performance of the LASSO Model or COX Model in the Internal Cohort

The calibration curves of the LASSO model and COX model for the probability of overall survival (OS) in 3-5 years between prediction and observation in the internal cohort (Figure 3A,3B, 3C,3D) were plotted to assess the calibration of the COX model and LASSO model, which were accompanied with the Hosmer-Lemeshow test (A significant test statistic implies that the model calibrates perfectly).

Validation of the LASSO Model and COX Model

Internally validation was tested using the internal cohort. The external validation was tested in the external cohort. The LASSO model was formed in the internal cohort and was applied to all the patients of the external cohort. The calibration curves in 3-5 years (Figure 4A, 4B) were derived on the basis of the regression analysis.

C-index and AIC

To quantify the discrimination performance of the COX model, LASSO model, TNM model, and TTNNMM model, Harrell's C-index and AIC were applied (Table 3). The C-index for the COX model, LASSO model, TNM model and TTNNMM model were 0.710 (95% CI, 0.703 to 0.717), 0.712 (95% CI, 0.705 to 0.719),

0.637 (95% CI, 0.631 to 0.644) and 0.651 (95% CI, 0.644 to 0.657), which were confirmed to be 0.710, 0.712, 0.637 and 0.651 via bootstrapping validation. The AIC for the COX model, LASSO model, TNM model, and TTNNMM model were 33431, 33420, 34043, 33994. The 1-,3-,5- years AUC for the COX model, LASSO model, TNM model, and TTNNMM model are shown in Table 4.

Predictive Accuracy of COX Model and LASSO Model

According to the survROC curves for 1-,3-,5- years overall survival (OS) for the COX model, LASSO model, TNM model, and TTNNMM model (Figs 5A,5B,5C,5D), the ROC curve (a general measure of predictiveness) was found to be greater in 3- and 5- years.

Whether Apparent Different Performance of The LASSO and COX Model

TimeAUC

Time-dependent ROC curves were generated to compare the sequential trends of the LASSO, COX, TNM and TTNNMM model for OS. The time-dependent ROC curve of the LASSO model was continuously superior to that of the COX model, TNM model and TTNNMM model (Figure 6).

BIC

The prognostic performances of the LASSO, COX, TNM, and TTNNMM model were compared using BIC, which is not only a measure of the goodness of fit of an estimated statistical model but also accurately considers the number of parameters included in the model. As shown in Figure 7, there was no significant difference between the COX and LASSO model after the bootstrap analysis (BIC 4.49,95% CI, –2.92–11.91) but there was a significant difference between the TNM and LASSO model (BIC 1178.76,95% CI,1171.15–1186.37), also TTNNMM and LASSO model (BIC 1098.57,95% CI,1092.05–1105.09).

Net Reclassification Improvement (NRI) and Integrated Discrimination Improvement (IDI)

The discriminant ability for LASSO model, COX model, TNM model, and TTNNMM model was calculated using NRI and IDI (Table 5). Compared to the TNM model and TTNNMM model, LASSO model was found to be a higher discriminant and possess reclassification indices (integrated discrimination improvement 0.072 and 0.064; $p < 0.001$; net reclassification improvement 0.525 and 0.466) (Table 4). In addition, compared to the COX model, the LASSO model doesn't significantly decrease the discriminant and reclassification indices (integrated discrimination improvement -0.002, p, 0.058; NRI -0.009) (Table 5).

Clinical Use

Decision curve analysis was conducted to determine the clinical usefulness of the LASSO model by quantifying the net benefits at different threshold probabilities. We also plotted the decision curve for the four models in 3-5 years (Figure 8A, 8B).

Visualization of SCC Survival Prediction Model

Survival prediction model of the nomogram was established based on factors selected by LASSO combined with the Cox regression (Figure 9). The nomogram showed that first time age had the most contribution to prognosis, followed by first- and second-times T stage, N stage, metastases and examined lymph nodes. Sex had a modest effect on survival. Each subtype of the variables was assigned a score. A straight line can be drawn down at each time point on the total point scale to determine the estimated probability of survival, according to the total number of points. For each predictor, the points assigned on the 0–10 scale at the top are read and then these points are added. The number on the “Total Points” scale were found and then the corresponding predictions of 3-, and 5-year risk are recorded.

Discussion

In this study, we developed and validated a prognostic model about SSC which based on large data from SEER by combining LASSO regression and COX regression. Our results show that the OS is associated with age, sex, second-time chemotherapy and first and second times pT, pN, pM.

Notably, the second time examined lymph nodes didn’t show enough predictive strength on the basis of Cox regression, which makes a common strategy to exclude this variable for model development. However, it may be a result of nuances in the data set or confounding by other predictors that reject important predictors, [17, 28] for which no significant statistical association with OS does not definitively imply that examined lymph nodes are unimportant. In addition, more lymph nodes examined may mean a better quality of operation. Therefore, we kept the second time examined lymph nodes as a candidate factor in the process of model development. For the same reason, we kept surgery, sex, the first-time tumor size and the second-time grade, pN, and the site in the COX model.

Grade, size, surgery and marital which may be multi-collinearity bias with pT, pN, pM, and age were not included in the LASSO model. Grade, size, surgery may be associated with TNM grading[12]. Besides, the old aged were more likely to be widowed. Also, because of overfitting site was not included in the LASSO model.

From Table 1, we found that patients with SCC were generally older (>65 years), more often male, and likely with the depth of invasion of T3. There may be less estrogen to protect in male[29] and a high probability of microsatellite instability (MSI) in older patients[30]. Besides, it may be as a result of tumor biologic characteristics with the depth of invasion by T3. Therefore, patients who are older (>65 years), men, and depth of invasion by T3 should closely monitor the postoperative enteroscopy for early detection of SCC.

Table 3-5 and Figure 3-8 show that there was no significant difference between the LASSO model and COX model (NRI, IDI, c index, ROC, AIC, and BIC) but the LASSO model was obviously better than the TNM model and TTNNMM model (NRI, IDI, c index, AIC, ROC, and BIC). Compared to the COX model, the LASSO model significantly reduced the variables included which minimized overfitting and collinearity. Moreover, Figure 6 shows that the LASSO model performs better than the Cox model in the timeAUC. Although the

LASSO model included fewer variables, the LASSO model performed better in the timeAUC compared to the COX model. (Table 6)

Some studies have also indicated that males are more susceptible to SCC[5]. Besides, SCC occurs more often in the right hemi colon and sigmoid colon[4-6], therefore there may exist short-term postoperative complications in SCC which contribute to the offset within 1 year[5].

We also found that the AUC for timeROC in 3-5 years OS was larger than that in 1- year. There is a possibility that most patients die of postoperative complications within 1 year, so we didn't perform prognosis of patient's OS within 1 year.

The most important and final argument for the use of the nomogram is based on the need to interpret the individual need for additional treatment or care. However, the clinical consequences of a particular level of discrimination or degree of miscalibration cannot be adequately assessed by the risk-prediction discrimination, performance, and calibration [17, 31, 32]. Therefore, in order to justify the clinical usefulness, it is crucial to ascertain whether the LASSO model-assisted decisions can improve patient outcomes. With this aim, in this study, the application of the decision curve analysis instead of the multi-institutional prospective for the validation of the model was performed. This novel method offers an insight into the clinical consequences on the basis of threshold probability, from which the net benefit could be derived. (Net benefit is defined as the proportion of true positives minus the proportion of false positives, weighted by the relative harm of false-positive and false-negative results.)[17, 33] Through the decision curve plot, we can ascertain whether the probability of threshold of a patient or doctor is 5% using the LASSO model is more beneficial than either the TNM model or TTNNMM model and not inferior compared to the COX model. (Figure 8A, 8B)

There are some limitations in the present work that should be discussed. The collection of the SEER database is retrospective. There is a lack of molecular data and data for biological prognostic factors that might also influence the prognosis of SCC patients. In recent years, increased research with gene markers, such as MSI, SSA, BRAF V600E associated with SCC has been proposed.[3, 6] in this regard, there might be some increase in the bias that we excluded all patients who had missing data from the collected variables. The study didn't incorporate detailed chemotherapy and radiation methods due to the lack of adequate information and large bias of the information. Finally, although this nomogram performed well in both internal and external cohorts, due to the influence of deaths related to the operation, the data should be used with caution when predicting 1-year risk. Even so, although it didn't include the genomic characteristics, excluded patients who had missing data and was retrospect to analysis, it was the first model to perform a prognosis OS of SCC.

Based on the database content, the main influencing factors were screened for the LASSO model. Due to the limitations of the database, some important factors weren't covered. In the future, we hope to have relevant data to incorporate it into our research.

In conclusion, this study presents a prognosis nomogram that incorporates both the first-time and the second-time variables and can be conveniently used to facilitate the prediction of OS in patients with SCC.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors reviewed and approved the manuscript.

Availability of data and materials

The datasets are available in SEER database to select the eligible cases. The data are also available from the corresponding author.

Conflict of interest

Mr. YuXin Xu have no conflict of interest or financial ties to declare.

Prof. Pan Chi have no conflict of interest or financial ties to declare.

Dr. XiaoJie Wang have no conflict of interest or financial ties to declare.

Dr. Ying Huang have no conflict of interest or financial ties to declare.

Dr. DaoXiong Ye have no conflict of interest or financial ties to declare.

Acknowledgements

Not applicable.

Author Contributions

YuXin Xu processed the data and carried out computational simulations.

XiaoJie Wang, Ying Huang , DaoXiong Ye helped on data collections and analyses.

YuXin Xu , Pan Chi, XiaoJie Wang, Ying Huang , DaoXiong Ye analyzed the results.

YuXin Xu , Pan Chi, XiaoJie Wang drafed the manuscript.

All authors reviewed and approved the manuscript.

References

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA Cancer J Clin, 2018. **68**(1): p. 7-30.
2. Huang, Y.Q., et al., *Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer.* J Clin Oncol, 2016. **34**(18): p. 2157-64.
3. Hu, H., et al., *Clinicopathologic features of synchronous colorectal carcinoma: A distinct subset arising from multiple sessile serrated adenomas and associated with high levels of microsatellite instability and favorable prognosis.* American Journal of Surgical Pathology, 2013. **37**(11): p. 1660.
4. Brandariz, L., et al., *New Perspectives in Multiple Primary Colorectal Cancer: A Surgical Approach.* Digestion, 2016. **94**(2): p. 57-65.
5. Leersum, N.J., Van, et al., *Synchronous colorectal carcinoma: a risk factor in colorectal cancer surgery.* Diseases of the Colon & Rectum, 2014. **57**(4): p. 460-6.
6. Lam, A.K., S.S. Chan, and M. Leung, *Synchronous colorectal cancer: clinical, pathological and molecular implications.* World J Gastroenterol, 2014. **20**(22): p. 6815-20.
7. Fante, R., ., et al., *Frequency and clinical features of multiple tumors of the large bowel in the general population and in patients with hereditary colorectal carcinoma.* Cancer, 2015. **77**(10): p. 2013-2021.
8. Papadopoulos, V., et al., *Synchronous and metachronous colorectal carcinoma.* Tech Coloproctol, 2004. **8 Suppl 1**: p. s97-s100.
9. Liu, Y.L., et al., *Prognostic significance of lymph node status in patients with metastatic colorectal carcinoma treated with lymphadenectomy.* J Surg Oncol, 2014. **109**(3): p. 234-8.
10. Howlader, N., et al., *Improved Estimates of Cancer-Specific Survival Rates From Population-Based Data.* J Natl Cancer Inst, 2010. **102**(20): p. 1584-1598.
11. Mariotto, A.B., et al., *Cancer survival: an overview of measures, uses, and interpretation.* J Natl Cancer Inst Monogr, 2015. **2014**(49): p. 145-186.
12. Amin, M.B., et al., *The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging.* Ca A Cancer Journal for Clinicians, 2017. **67**(2): p. 93-99.
13. Sauerbrei, W., P. Royston, and H. Binder, *Selection of important variables and determination of functional form for continuous predictors in multivariable model building.* Statistics in Medicine, 2010. **26**(30): p. 5512-5528.
14. Tibshirani, R., *Regression shrinkage and selection via the lasso: A retrospective.* Journal of the Royal Statistical Society: Series B Statistical Methodology, 2011. **73**(3): p. 273-282.
15. Tibshirani, R., . *The lasso method for variable selection in the Cox model.* Statistics in Medicine, 1997. **16**(4): p. 385-395.
16. Venables, W.N. and B.D. Ripley, *Modern Applied Statistics with S.* Statistics & Computing, 2002. **52**(1): p. 704–705.

17. Collins, G.S., et al., *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD Statement*. 2015. 146–154.
18. Sauerbrei, W., A.L. Boulesteix, and H. Binder, *Stability investigations of multivariable regression models derived from low- and high-dimensional data.* Journal of Biopharmaceutical Statistics, 2011. **21**(6): p. 1206-1231.
19. Rao, S.J., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis by Frank E. Harrell.* Publications of the American Statistical Association, 2005. **98**(461): p. 257-258.
20. Paul, B., D. Jean-Fran?Ois, and J.G. Hélène, *Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks.* Statistics in Medicine, 2013. **32**(30): p. 5381-5397.
21. Heagerty, P.J., T. Lumley, ., and M.S. Pepe, *Time-dependent ROC curves for censored survival data and a diagnostic marker.* Biometrics, 2000. **56**(2): p. 337-344.
22. Rodríguez-Álvarez, M.X., et al., *Nonparametric estimation of time-dependent ROC curves conditional on a continuous covariate.* Statistics in Medicine, 2016. **35**(7): p. 1090-1102.
23. Hanley, J.A. and B.J. Mcneil, *A method of comparing the areas under receiver operating characteristic curves derived from the same cases.* Radiology, 1983. **148**(3): p. 839-843.
24. Pencina, M.J., E.W. Steyerberg, and R.B.D.A. Sr, *Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers.* Statistics in Medicine, 2011. **30**(1): p. 11-21.
25. Navdeep, T., et al., *A predictive model for progression of chronic kidney disease to kidney failure.* Jama the Journal of the American Medical Association, 2011. **305**(15): p. 1553-9.
26. Chambless, L.E., C.P. Cummiskey, and C. Gang, *Several methods to assess improvement in risk prediction models: extension to survival analysis.* Statistics in Medicine, 2011. **30**(1): p. 22-38.
27. Vickers, A.J., et al., *Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers.* Bmc Medical Informatics & Decision Making, 2008. **8**(1): p. 53.
28. Moons, K.G., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration.* European Journal of Clinical Investigation, 2015. **45**(2): p. 204-214.
29. Fernandez, E., ., et al., *Hormone replacement therapy and risk of colon and rectal cancer.* Cancer Epidemiol Biomarkers Prev, 1998. **7**(4): p. 329-333.
30. Hyung Min, S., et al., *Clinicopathologic characteristics and outcomes of gastric cancers with the MSI-H phenotype.* Journal of Surgical Oncology, 2010. **99**(3): p. 143-147.
31. Van, C.B. and A.J. Vickers, *Calibration of risk prediction models: impact on decision-analytic performance.* Medical Decision Making An International Journal of the Society for Medical Decision Making, 2015. **35**(2): p. 162.
32. *<dca1.pdf>.*

33. Balachandran, V.P., et al., *Nomograms in oncology: more than meets the eye.* Lancet Oncology, 2015. **16**(4): p. e173-e180.

Tables

Table 1. Characteristics of patients with Synchronous colorectal carcinomas

First Time Characteristics	Total	Second Time Characteristics	Total	rho
	4616(100%)		4616(100%)	
Age(years)		Age(years)		0.999
0-49	369(8.0%)	0-49	368(8.0%)	
50-59	619(13.4%)	50-59	618(13.4%)	
60-64	454(9.8%)	60-64	450(9.7%)	
65-69	573(12.4%)	65-69	577(12.5%)	
70-74	652(14.1%)	70-74	646(14.0%)	
75-79	721(15.6%)	75-79	727(15.7%)	
80-84	656(14.2%)	80-84	654(14.2%)	
85+	573(12.4%)	85+	576(12.5%)	
Site		Site		0.087
Large Intestine, NOS	34(0.7%)	Large Intestine, NOS	38(0.8%)	
Rectum	468(10.1%)	Rectum	599(13.0%)	
Rectosigmoid Junction	228(4.9%)	Rectosigmoid Junction	286(6.2%)	
Sigmoid Colon	816(17.7%)	Sigmoid Colon	786(17.0%)	
Descending Colon	287(6.2%)	Descending Colon	372(8.1%)	
Splenic Flexure	163(3.5%)	Splenic Flexure	165(3.6%)	
Transverse Colon	531(11.5%)	Transverse Colon	583(12.6%)	
Hepatic Flexure	268(5.8%)	Hepatic Flexure	221(4.8%)	
Ascending Colon	807(17.5%)	Ascending Colon	766(16.6%)	
Cecum	1014(22.0%)	Cecum	800(17.3%)	
pT		pT		0.205
Tis/T0	95(2.1%)	Tis/T0	253(5.5%)	
T1	624(13.5%)	T1	1229(26.6%)	
T2	702(15.2%)	T2	909(19.7%)	
T3	2620(56.8%)	T3	1917(41.5%)	
T4	575(12.5%)	T4	308(6.7%)	
pN		pN		0.638
N0	2746(59.5%)	N0	3032(65.7%)	
N1	1163(25.2%)	N1	983(21.3%)	
N2	707(15.3%)	N2	601(13.0%)	
pM		pM		0.460
M0	4083(88.5%)	M0	4067(88.1%)	
M1	533(11.5%)	M1	549(11.9%)	
Examined lymph nodes		Examined lymph nodes		0.770
1-14	1652(35.8%)	1-14	1813(39.3%)	
15-39	2558(55.4%)	15-39	2417(52.4%)	
40+	406(8.8%)	40+	386(8.4%)	
Grade		Grade		0.306
unkwon	233(5.0%)	unkwon	499(10.8%)	
Grade I	360(7.8%)	Grade I	483(10.5%)	
Grade II	3006(65.1%)	Grade II	2924(63.3%)	
Grade III	880(19.1%)	Grade III	629(13.6%)	
Grade IV	137(3.0%)	Grade IV	81(1.8%)	
Tumor Size		Tumor Size		0.172
1-29mm	860(18.6%)	1-29mm	1445(31.3%)	
30-99mm	3035(65.7%)	30-99mm	2238(48.5%)	
100-900mm	192(4.2%)	100-900mm	96(2.1%)	
unknown	529(11.5%)	unknown	837(18.1%)	
Surg Prim Site		Surg Prim Site		0.559
Radical resection	4159(90.1%)	Radical resection	4158(90.1%)	
Combined organ resection	403(8.7%)	Combined organ resection	361(7.8%)	
Partial resection	29(0.6%)	Partial resection	63(1.4%)	

Surgical resection	13(0.3%)	Surgical resection	22(0.5%)	
Radical resection+Ostomy	12(0.3%)	Radical resection+Ostomy	12(0.3%)	
Chemotherapy		Chemotherapy Recode		0.796
No/Unknown	3063(66.4%)	No/Unknown	3237(70.1%)	
Yes	1553(33.6%)	Yes	1379(29.9%)	
Sex				
Male	2498(54.1%)			
Female	2118(45.9%)			
Race				
White	3773(81.7%)			
Black	489(10.6%)			
Other	354(7.7%)			
Marital				
single	581(12.6%)			
unmarri	4(0.1%)			
marry	2425(52.5%)			
separat	43(9.0%)			
Divorce	386(8.4%)			
Widowed	1000(21.7%)			
unknown	177(3.8%)			

Table 2. Cox regression and LASSO combine Cox regression of clinical characteristics for prognosis of SSC for overall survival (OS)

variable	COX				LASSO			
	adj.p	HR	95%CI L	95%CI U	adj.p	HR	95%CI L	95%CI U
First Time Characteristics								
Age(years)	<0.001	2.293	2.091	2.514	<0.001	2.327	2.129	2.544
Sex	<0.001	0.794	0.727	0.868	<0.001	0.816	0.749	0.889
Site	0.023	1.092	1.012	1.178				
pT	<0.001	1.253	1.188	1.323	<0.001	1.247	1.182	1.316
pN	<0.001	1.733	1.490	2.015	<0.001	1.688	1.450	1.964
pM	<0.001	2.444	2.114	2.824	<0.001	2.422	2.096	2.798
Examined lymph nodes	<0.001	0.691	0.597	0.799	0.035	0.778	0.617	0.982
Tumor Size	0.134	1.132	0.962	1.332				
Surg Prim Site	0.013	1.681	1.118	2.527				
Second Time Characteristics								
Marital	0.046	1.055	1.001	1.111				
Site	0.043	0.926	0.860	0.998				
Grade	0.091	1.204	0.971	1.493				
pT	<0.001	1.312	1.197	1.438	<0.001	1.349	1.236	1.473
pN	0.07	1.159	0.988	1.359	0.034	1.190	1.013	1.397
pM	0.003	1.253	1.079	1.457	0.003	1.255	1.081	1.458
Examined lymph nodes					0.312	0.887	0.704	1.119
Chemotherapy	<0.001	0.715	0.642	0.795	<0.001	0.721	0.649	0.802

Table 3. AIC and C-index for four models

	AIC	C-index	Concordance
LASSO	33431	0.710	0.710 (0.703-0.717)
COX	33420	0.712	0.712 (0.705-0.719)
TNM	34043	0.637	0.637 (0.631-0.644)
TTNNMM	33994	0.651	0.651 (0.644-0.657)

Table 4. 1-,3-,5- years timeAUC for four models

	AUC		
	1-years	3-years	5-years
LASSO	0.681	0.698	0.712
COX	0.638	0.646	0.651
TNM	0.646	0.676	0.672
TTNNMM	0.659	0.696	0.694

Table 5. NRI and IDI for comparing LASSO model and other models

	NRI			IDI			
	value	95% CI L	95% CI U	value	95% CI L	95% CI U	P value
LASSO&COX	-0.009	-0.054	0.012	0.002	0.000	0.006	0.058
LASSO&TNM	0.525	0.464	0.589	-0.072	-0.087	-0.061	<0.001
LASSO&TTNNMM	0.466	0.415	0.538	-0.064	-0.079	-0.052	<0.001

Figures

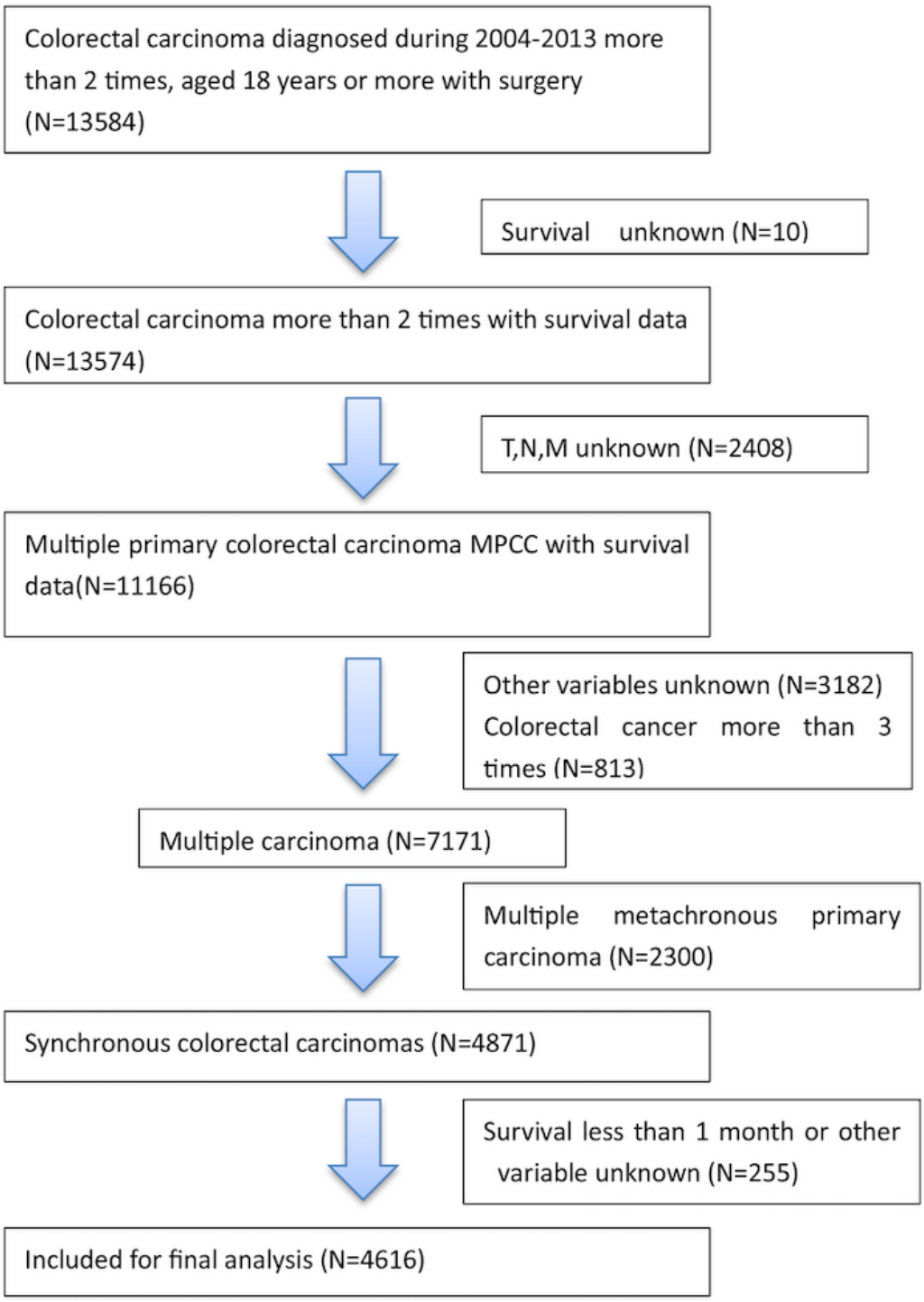

Figure 1

Flowchart of patient selection for this study

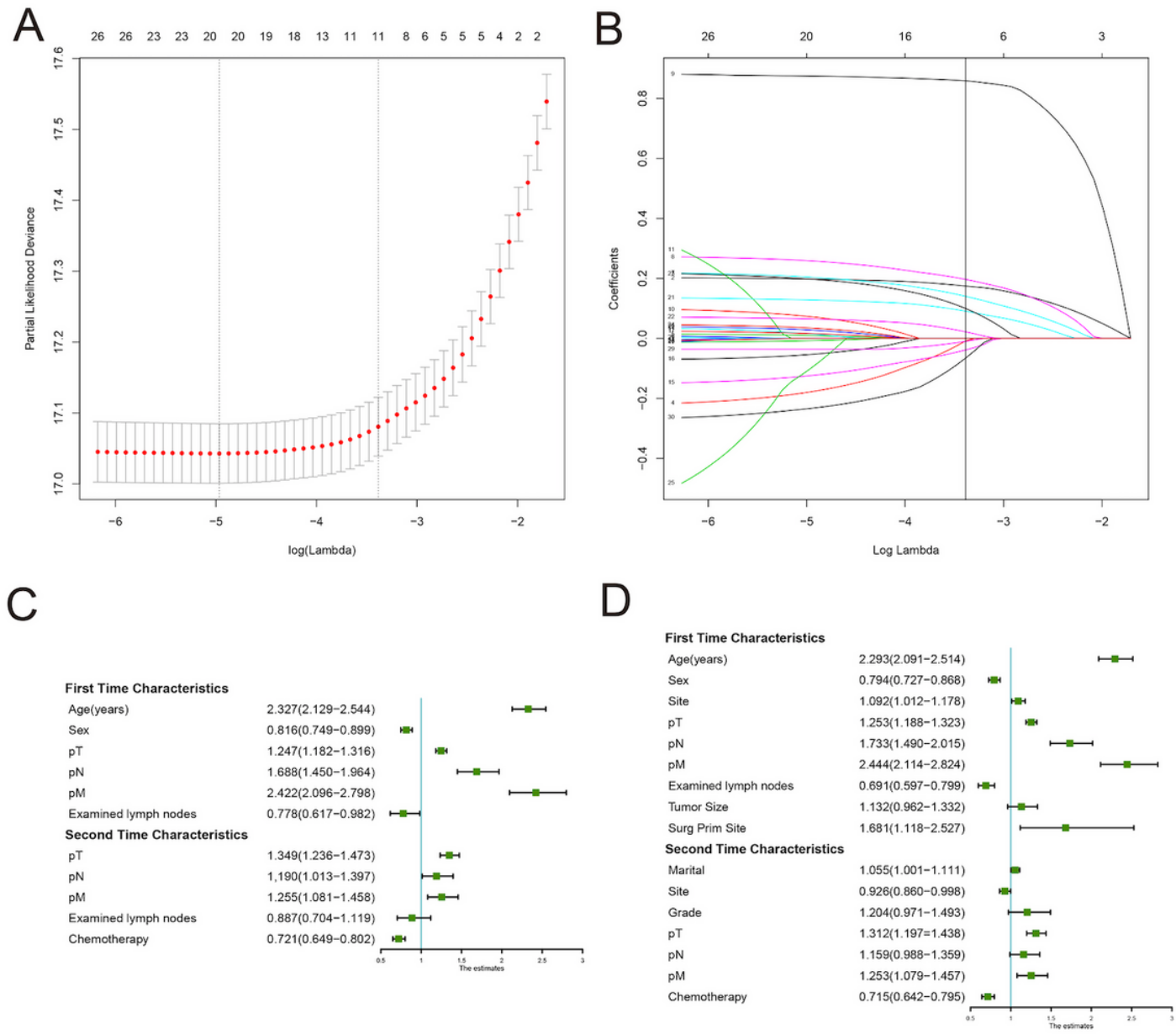

Figure 2

The least absolute shrinkage and selection operator (LASSO) Cox regression model was used to select predicted variables. (A) Tuning parameter (λ) selection in the LASSO model used 10-fold cross-validation via minimum criteria. The partial likelihood deviance curve was plotted versus log (λ). Dotted vertical lines were drawn at the optimal values using the minimum criteria and the 1 standard error of the minimum criteria (the 1-SE criteria). A λ value of 0.034, with log (λ), -3.385 was chosen (1-SE criteria) according to 10-fold cross-validation. (B) LASSO coefficient profiles of the 31 variables. A coefficient profile plot was produced against the log (λ) sequence. The vertical line was drawn at the value selected using 10-fold cross-validation, where optimal λ resulted in 11 nonzero coefficients. Multivariable analysis of factors affecting overall survival (OS) by Cox regression and LASSO combine Cox regression. (C) The

plot shows the hazard ratios (HRs; squares) and 95% CIs (lines) of LASSO combine multivariable Cox regression. (D)The plot shows the hazard ratios (HRs; squares) and 95% CIs (lines) of multivariable Cox regression. The figure shows all of the significant covariates. The vertical line represents an HR of 1 for reference.

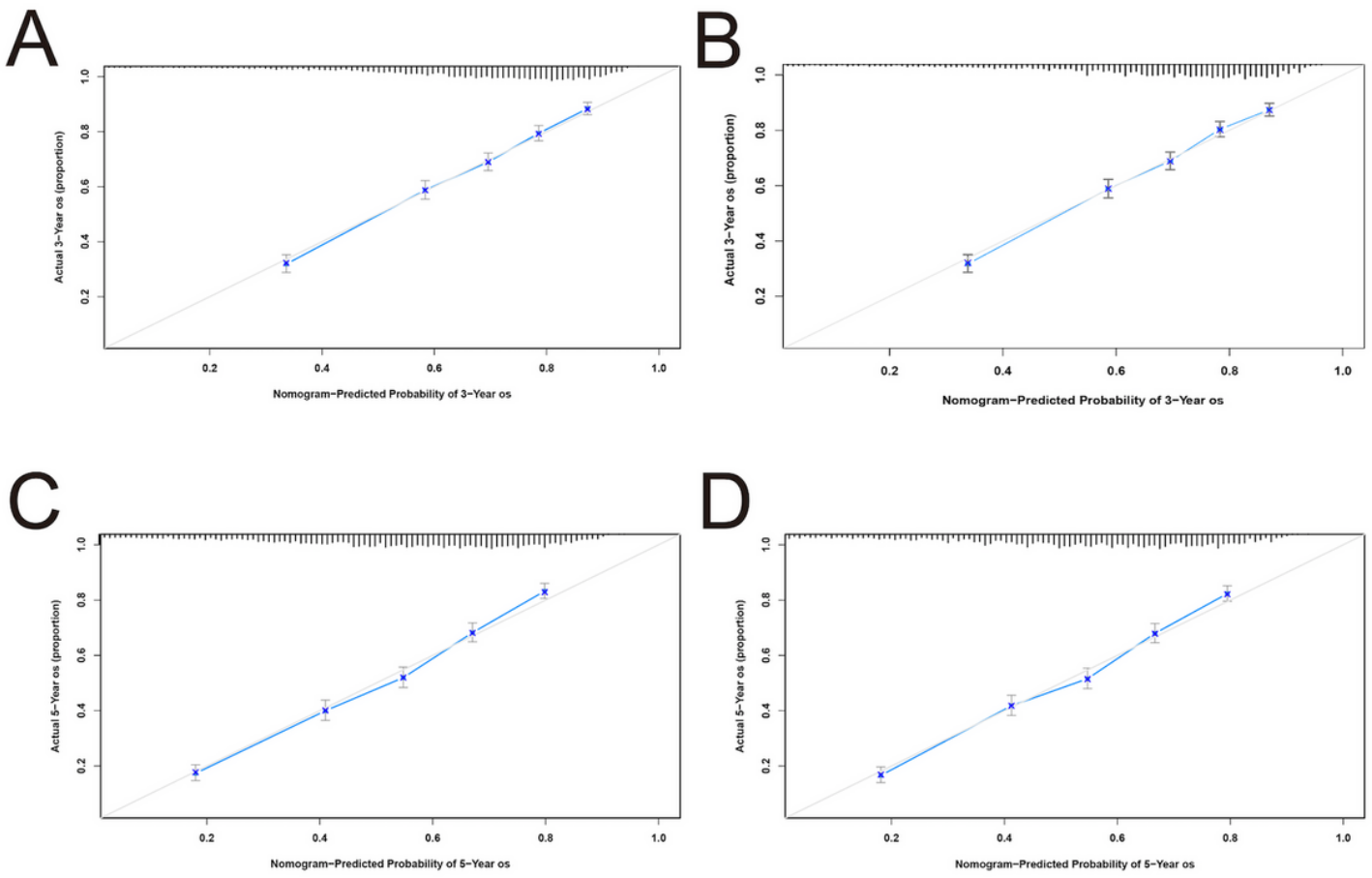

Figure 3

Calibration curves of the Cox nomogram and the LASSO model in each cohort. (A) Calibration curve of 3-years OS of the Cox model in the primary cohort. (B) Calibration curve of 3-years OS of the LASSO nomogram in the primary cohort. (C) Calibration curve of 5-years OS of the Cox model in the primary cohort. (D) Calibration curve of 5-years OS of the LASSO nomogram in the primary cohort. The y-axis represents the actual overall survival(OS). Calibration curves depict the calibration of each model in terms of the agreement between the predicted risks of predicted overall survival(OS) and actual overall survival(OS). The y-axis represents the actual overall survival(OS). The x-axis represents the predicted overall survival(OS). The diagonal dotted line represents a perfect prediction by an ideal model. The blue solid line represents the performance of the model, of which a closer fit to the diagonal dotted line represents a better prediction.

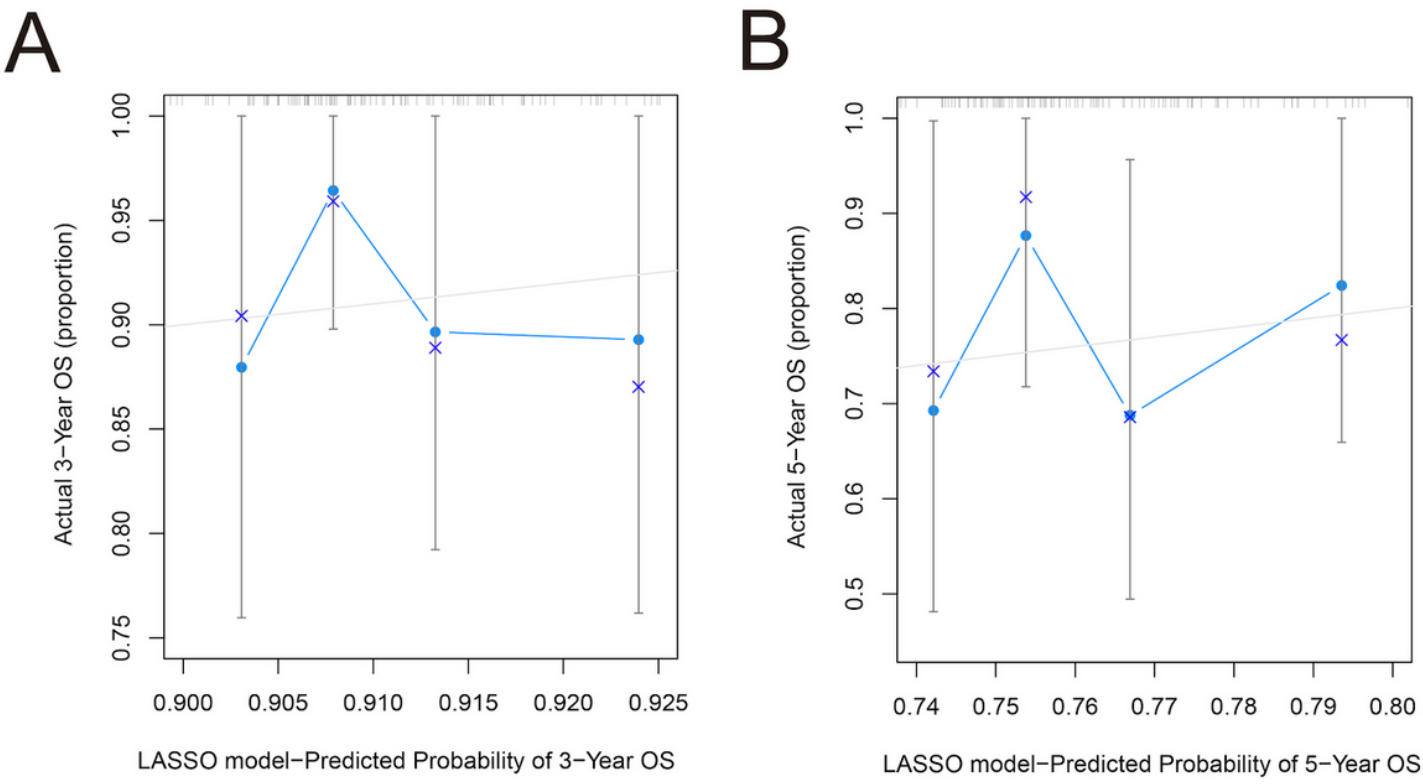

Figure 4

(A) Calibration curve of 3-years OS of the LASSO model in the validation cohort. (B) Calibration curve of 5-years OS of the LASSO model in the validation cohort.

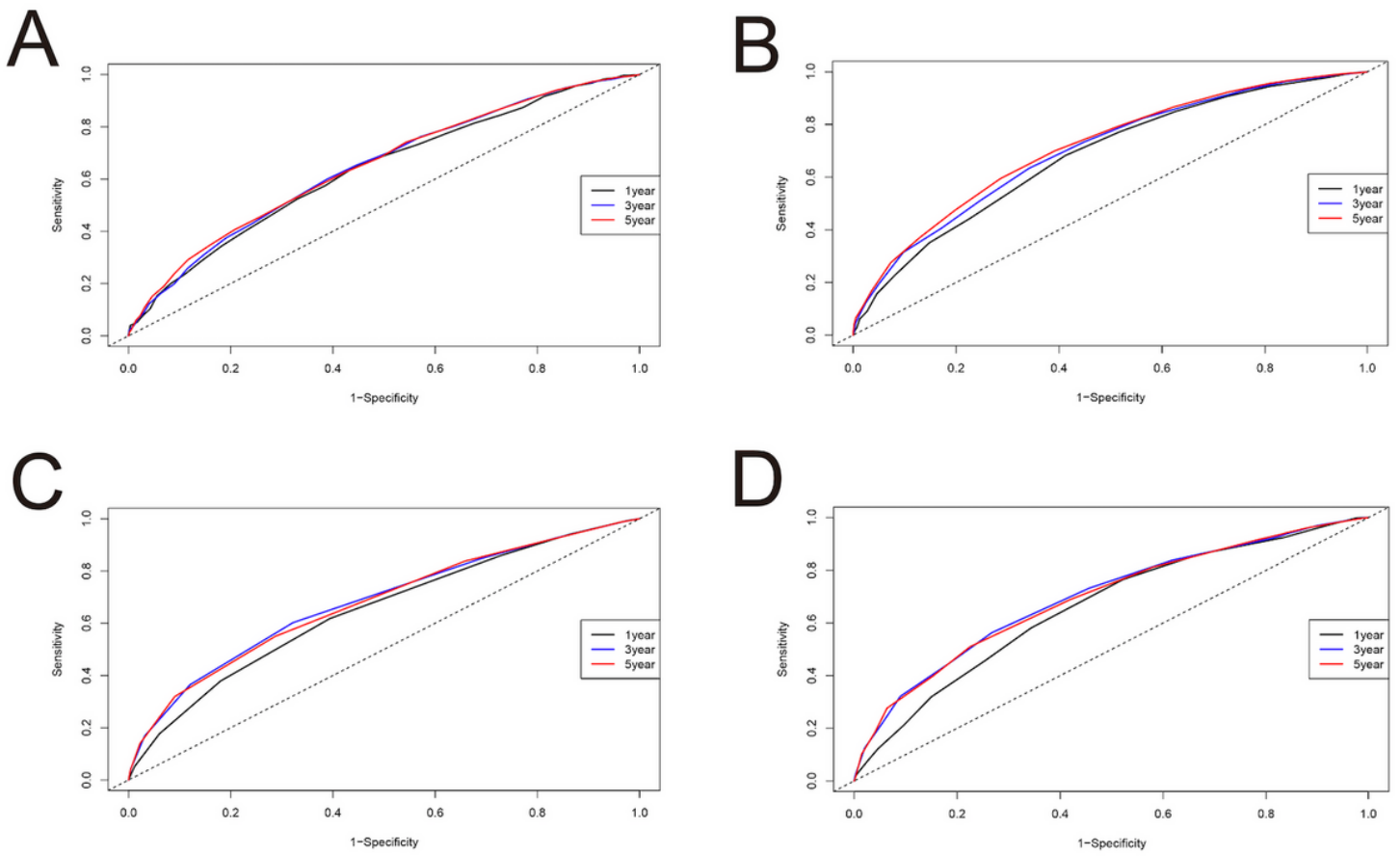

Figure 5

(A) ROC curve of the LASSO model in prediction of prognosis of patients at 1, 3 and 5-year point in the 2004–2013 primary cohort. (B) ROC curve of the COX model in prediction of prognosis of patients at 1, 3 and 5-year point in the 2004–2013 primary cohort. (C) ROC curve of the TNM Stage model in prediction of prognosis of patients at 1, 3 and 5-year point in the 2004–2013 primary cohort. (D) ROC curve of the TTNNMM stage model in prediction of prognosis of patients at 1, 3 and 5-year point in the 2004–2013 primary cohort.

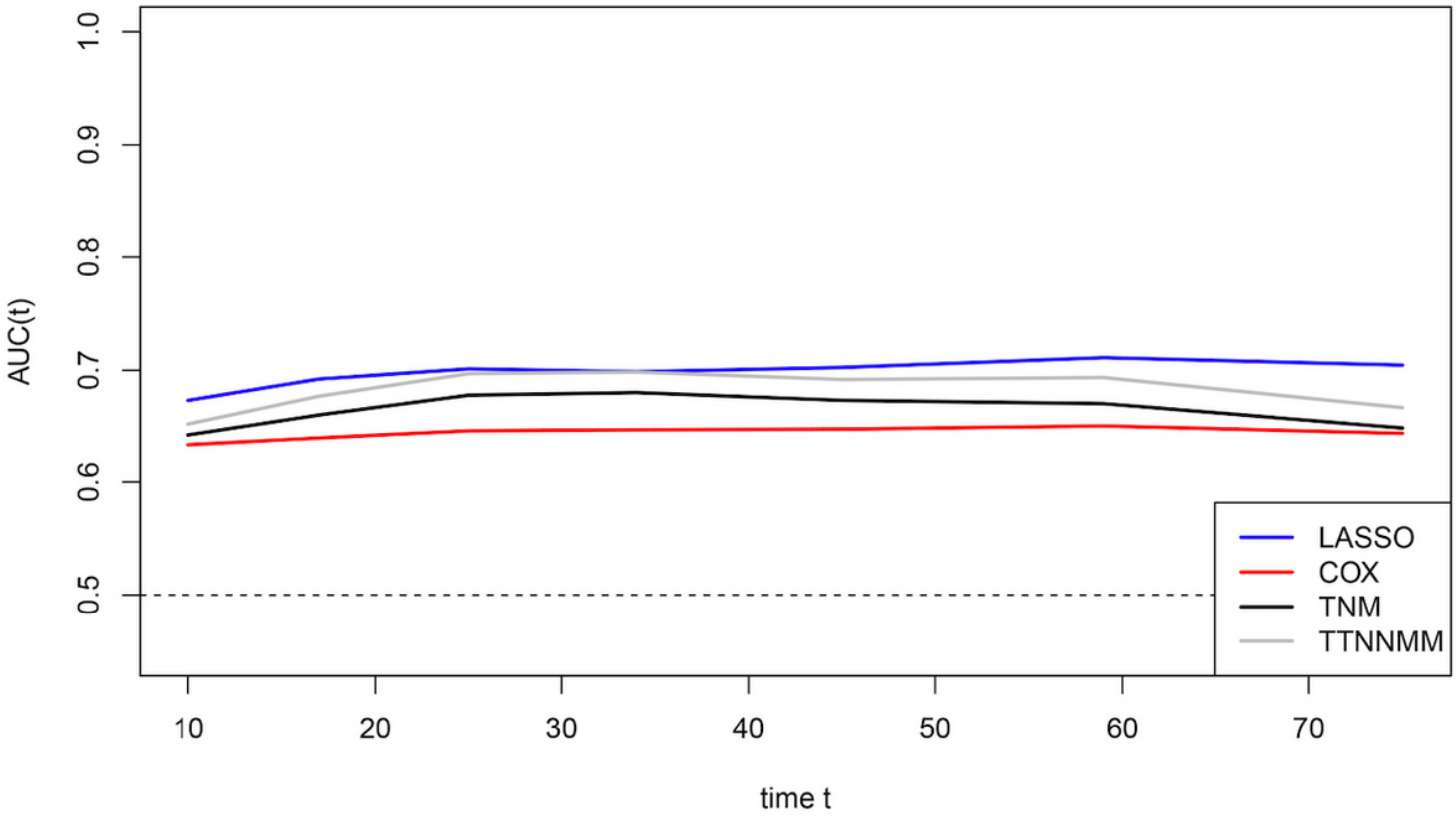

Figure 6

Time-dependent ROC curves for the LASSO, COX, TNM, and TTNNMM model. The horizontal axis represents year after diagnosis and the vertical axis of the estimated area under the ROC curve for survival at the time of interest. Blue, red, black and gray solid lines represent the estimated AUCs of the LASSO, COX, TNM and TTNNMM model.

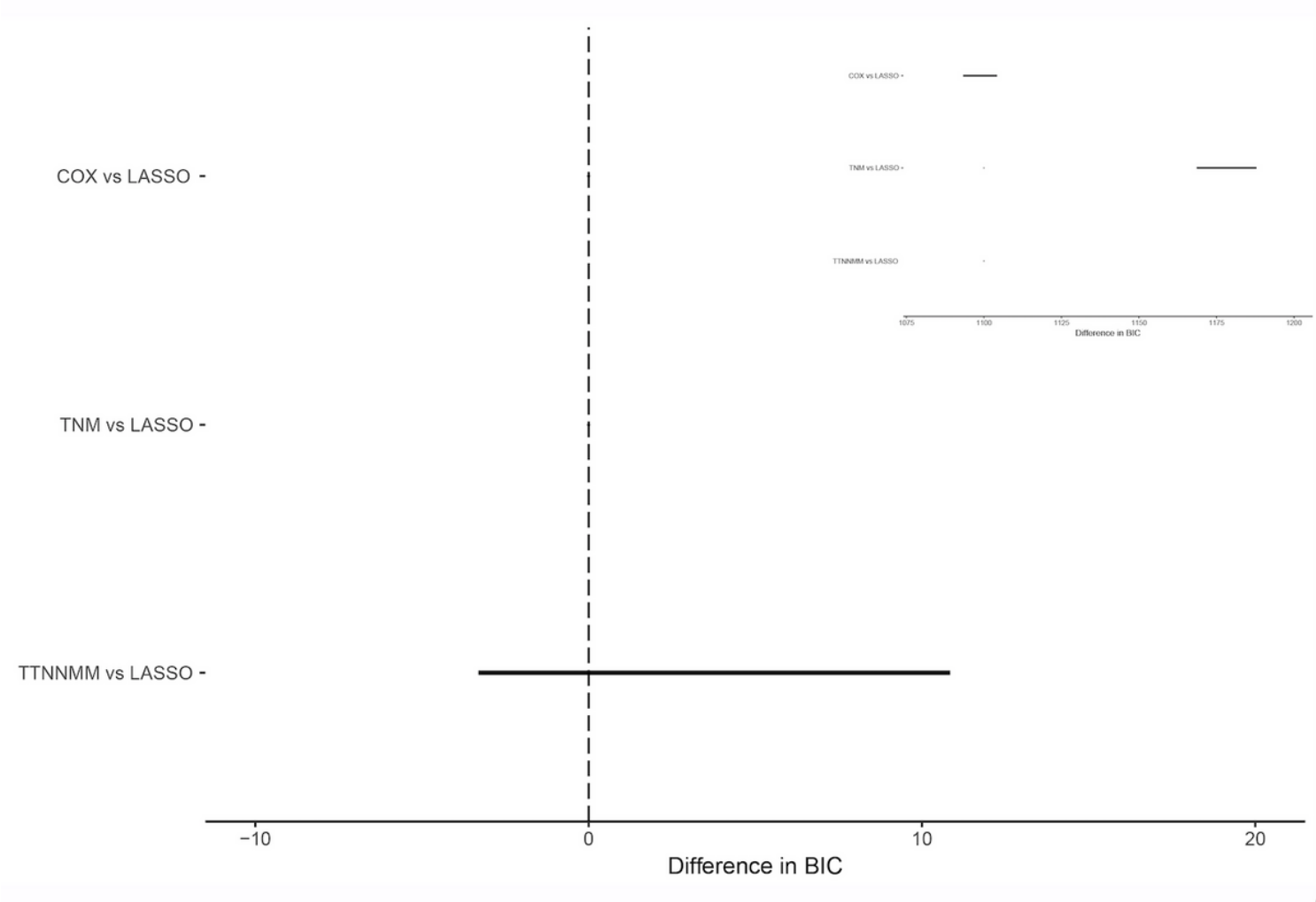

Figure 7

Results of the BIC level of the 4 different models. By the BIC via bootstrap analysis (1000 samples, 95% CI limits are shown).

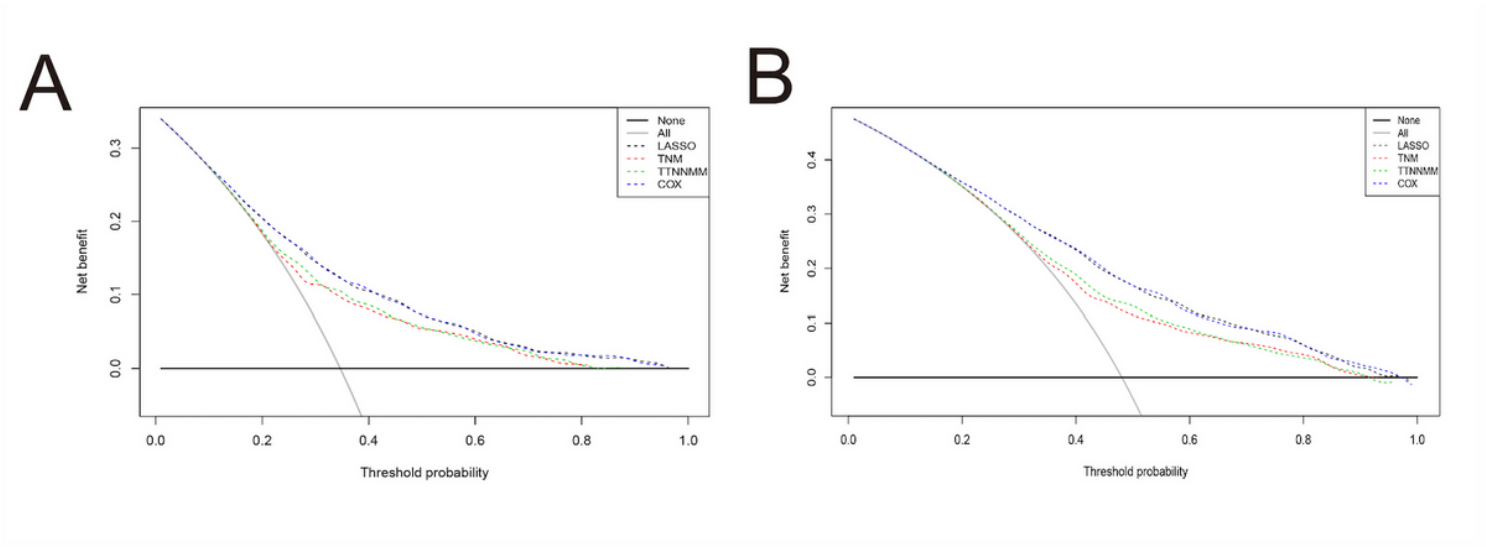

Figure 8

Decision curve analysis for the TNM model, TTNNMM model, Cox model and LASSO model in the prediction of prognosis of patients at 3- and 5-years point. The y-axis measures the net benefit. The black dotted line represents the LASSO model. The blue dotted line represents the COX model. The green dotted line represents the TTNNMM model. The red dotted line represents the TNM model. The black line represents the assumption that no patients died. The grey line represents the assumption that all patients died. The net benefit was calculated by subtracting the proportion of all patients who are false positive from the proportion who are truly positive, weighting by the relative harm of forgoing treatment compared with the negative consequences of unnecessary treatment. The decision curve showed that if the threshold probability of a patient or doctor is >5%, using the LASSO model in the current study to predict OS more benefit than the treat-all-patients scheme or the treat-none scheme. The net benefit was not comparable, with several overlaps, on the basis of the LASSO model and the COX model.

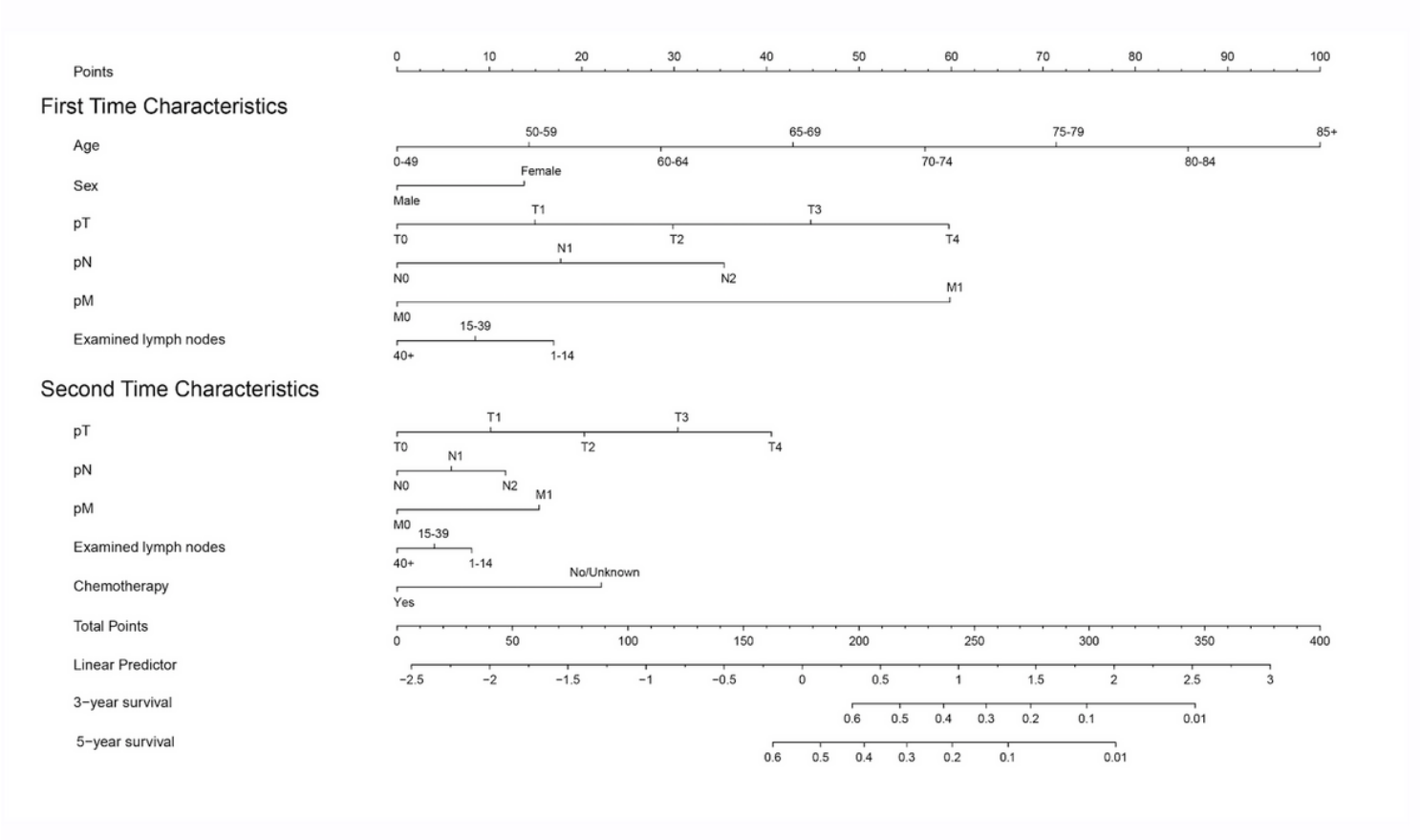

Figure 9

Developed LASSO model nomogram. The LASSO model nomogram was developed in the primary cohort, with the first time age and sex, the second time chemotherapy and the first and second times pT, pN, pM, and regional nodes examined incorporated. LASSO model nomograms to predict 3- and 5-year overall survival probability with SCC. For each predictor, read the points assigned on the 0–10 scale at the top and then add these points. Find the number on the “Total Points” scale and then read the corresponding predictions of 3- and 5-year risk.