Bioinformatics, 31(20), 2015, 3330–3338
doi: 10.1093/bioinformatics/btv374
Advance Access Publication Date: 18 June 2015
Original Paper

OXFORD

Systems biology

NCC-AUC: an AUC optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data

Meng Zou, Zhaoqi Liu, Xiang-Sun Zhang and Yong Wang*

Academy of Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 10080, China

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on March 9, 2015; revised on May 28, 2015; accepted on June 14, 2015

Abstract

Motivation: In prognosis and survival studies, an important goal is to identify multi-biomarker panels with predictive power using molecular characteristics or clinical observations. Such analysis is often challenged by censored, small-sample-size, but high-dimensional genomic profiles or clinical data. Therefore, sophisticated models and algorithms are in pressing need.

Results: In this study, we propose a novel Area Under Curve (AUC) optimization method for multi-biomarker panel identification named Nearest Centroid Classifier for AUC optimization (NCC-AUC). Our method is motived by the connection between AUC score for classification accuracy evaluation and Harrell's concordance index in survival analysis. This connection allows us to convert the survival time regression problem to a binary classification problem. Then an optimization model is formulated to directly maximize AUC and meanwhile minimize the number of selected features to construct a predictor in the nearest centroid classifier framework. NCC-AUC shows its great performance by validating both in genomic data of breast cancer and clinical data of stage IB Non-Small-Cell Lung Cancer (NSCLC). For the genomic data, NCC-AUC outperforms Support Vector Machine (SVM) and Support Vector Machine-based Recursive Feature Elimination (SVM-RFE) in classification accuracy. It tends to select a multi-biomarker panel with low average redundancy and enriched biological meanings. Also NCC-AUC is more significant in separation of low and high risk cohorts than widely used Cox model (Cox proportional-hazards regression model) and L_1-Cox model (L_1 penalized in Cox model). These performance gains of NCC-AUC are quite robust across 5 subtypes of breast cancer. Further in an independent clinical data, NCC-AUC outperforms SVM and SVM-RFE in predictive accuracy and is consistently better than Cox model and L_1-Cox model in grouping patients into high and low risk categories.

Conclusion: In summary, NCC-AUC provides a rigorous optimization framework to systematically reveal multi-biomarker panel from genomic and clinical data. It can serve as a useful tool to identify prognostic biomarkers for survival analysis.

Availability and implementation: NCC-AUC is available at http://doc.aporc.org/wiki/NCC-AUC.

Contact: ywang@amss.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Survival analysis studies the time to occurrence of a certain event. Identifying significant biomarkers for prediction is longstanding as an important topic to provide better understanding. Here one key challenge is that the outcome variable of interest for survival analysis is sometimes censored, which means the survive time is only partially known. In addition, the candidate features are always in high-dimension with the rapid development of high-throughput measurement technologies. As a result, discovering feature combinations is suffering from 'curse of dimensionality'. How to overcome the above difficulties and identify the most accurate prognostic multi-biomarker panel from survival data to assist clinicians is the problem we want to address.

The most widely used model for survival analysis is Cox proportional-hazards regression model (Cox model). It can handle the censored survival data in a rigorous way (Breslow, 1975; David, 1972) and provides an estimation of the hazard ratio and its confidence interval. However, Cox model directly utilizes the noisy and censored data to predict the survival ratio thus often leads to poor prognosis. Also Cox model uses all features with high dimensionality in general. It is trained on a particular dataset and usually generalized poorly to other independent datasets. To avoid overfitting, L_1 penalized Cox model (L_1-Cox model) was proposed to perform variable selection and shrinkage, which makes it very useful for de-noising and identifying the most meaningful features (Goeman, 2010). Nonetheless, both Cox model and L_1-Cox model heavily depend on the quality of input survival time.

In reality, Harrell's Concordance Index (CI) is a widely used performance measure for assessing prediction models in survival analysis. The main idea is that it doesn't consider the value of survival time but the rank of patients by their survival times (Harrell *et al.*, 1982, 1996). In DREAM Breast Cancer Prognosis Challenge, challenge models will be scored by calculating the CI between the predicted survival and the true survival information in the validation dataset. The final assessment of models and the determination of the best performer will be based on the CI of predictions on the test dataset (http://www.the-dream-project.org/). Recently, CI is used to assess the clinical utility of cancer genomic and proteomic data across tumor types in The Cancer Genome Atlas (TCGA) project (Yuan *et al.*, 2014).

In this sense, it is natural to directly optimize CI to select biomarkers and construct predictive models. However maximizing CI is a non-convex optimization problem with non-differentiable objective function and hard to solve (Steck *et al.*, 2008). We notice that CI and AUC (Area Under ROC Curve) are identical and equate the Mann-Whitney statistics in the absence of censored data (Koziol and Jia, 2009; Wolf *et al.*, 2011). We could use AUC to assess the classification accuracy and further to approximate CI. With this connection, our motivation is to convert survival time regression to a binary classification problem and then maximizing AUC. For example, we choose the 5-year threshold in practice without loss of generality, i.e. we predict whether the patient can live longer than 5 years. In clinical study, the 5-year survival rate is commonly used as a survival rate for estimating the prognosis of a particular disease. It serves as a general accepted standard by the American Cancer Society to assess the cancer malignancy. In addition, 5-year threshold leads to the balanced positive and negative datasets (Supplemental Table S4). In this way we focus to solve a classification problem about maximizing AUC.

Classical classification methods, such as support vector machine (Shivaswamy *et al.*, 2007; Van Belle *et al.*, 2011), logistic regression (Efron, 1988), k-nearest neighbor (Parry *et al.*, 2010), have been widely used in survival analysis and achieved good performance. Moreover, Heagerty *et al.* (2000) proposed the time-dependent AUC, which was adopted as evaluation for the model performance (Bonato *et al.*, 2011; Cerhan *et al.*, 2007; Simon *et al.*, 2011). Nevertheless, none of these methods optimize AUC directly to construct classifiers.

AUC is an important performance measure that has been widely used (Metz, 1978). Some algorithms have been developed to optimize AUC based on surrogate loss (Herschtal and Raskutti, 2004; Zhao *et al.*, 2011). Those technologies can be borrowed in optimizing AUC to deal with the censored survival data. Meanwhile, the rapidly developed high-throughput measuring techniques usually generate high dimensional genomic or clinical data. Feature selection is necessary along with model estimation to reduce data dimension and model complexity. Also it helps to provide intuitive understanding. Here, we denote selected features having potential diagnostic usage as biomarkers. Importantly, features combine in linear or nonlinear ways to improve diagnostic accuracy. This useful feature combination is named as multi-biomarker panel.

It's a challenging task to select a biomarker subset to maximize accuracy in computation. Here wrapper methods give a good example for its complexity. The idea of these state-of-art feature selection algorithms is to use each subset to train a model and then test on a hold-out dataset. Since wrapper methods need to go through all the possible subsets, they are computationally intensive. For example if there are 50 features, it should train 2^{50} models to select the best subset by comparing models' error rates on the hold-out validation dataset. To reduce the computational complexity, some heuristic methods have been proposed. For example, Support Vector Machine-based Recursive Feature Elimination (SVM-RFE) has become one of the leading methods and is being widely used (Guyon *et al.*, 2002). It adopts an intuitive strategy to remove a feature with the smallest coefficient in SVM model one step. We note that this stepwise feature selection strategy may miss some critical feature combinations.

In this article, we convert the survival time regression to a classification problem aiming to select a small group of biomarkers and maximize AUC score. Specially, we propose a single optimization model based on nearest centroid classifier with feature selection, which maximizes AUC and minimizes the number of selected features simultaneously. To make the computation tractable, maximizing AUC is slacked to minimize a loss function. Meanwhile, we use L_1 'norm' to minimize the number of selected features instead of L_0 'norm' thus make it could be efficiently solved by a linear programming. Finally our method predicts whether a patient survives longer than 5 years in a fast and efficient way. To validate the performance, we apply the new method to gene expression data of breast cancer and clinical data of stage IB Non-Small-Cell Lung Cancer (NSCLC). The results show that our method can identify panels of biomarkers with good survival prediction performance.

2 Methods

2.1 Overview of NCC-AUC

We propose a novel method based on nearest centroid classifier and maximize classification accuracy by AUC. Simultaneously, we minimize the number of features to select the best multi-biomarker panel for survival prediction. To make the computation efficient, we approximate AUC by a loss function. Meanwhile the number of features (L_0 'norm') is approximated by the sum of probability for

selecting the features (L_1 'norm'). To balance these two objectives, a parameter λ is introduced to formulate the model as a single-objective-function optimization problem. Solving the optimization problem allows us to select a panel of biomarkers to assist survival prediction (Fig. 1). We name our method as NCC-AUC (Nearest Centroid Classifier for AUC optimization).

2.2 Survival time classification

As shown in Figure 1, the identification of patients with longer survival time is treated here as a binary classification problem. We apply a threshold (5 years) to the survival time to each patient and classify all the patients into two groups: positive samples are patients living shorter than 5 years and negative samples are those patients living longer than 5 years. For the censored data, we only keep the patients with censored time above 5 years and classified them as negative samples.

Assume that we have n positive samples $X_1, X_2, \cdots, X_n$ and m negative samples $Y_1, Y_2, \cdots, Y_m$. Each sample is represented by a p-dimensional feature vector. Then the centroids of the positive and negative classes are $\mu_+ = \frac{\sum_{i=1}^{n} X_i}{n}$ and $\mu_- = \frac{\sum_{j=1}^{m} Y_j}{m}$ respectively.

The idea of nearest centroids is intuitive. If X is a positive sample, it should satisfy,

$$\|X - \mu_+\| < \|X - \mu_-\| \tag{1}$$

Dealing with the L_2 'norm' leads to

$$2X^{\mathrm{T}}(\mu_- - \mu_+) < \|\mu_-\|^2 - \|\mu_+\|^2 \tag{2}$$

Similarly, if X is a negative sample, it should satisfy,

$$2X^{\mathrm{T}}(\mu_+ - \mu_-) < \|\mu_+\|^2 - \|\mu_-\|^2 \tag{3}$$

Let $w = \mu_+ - \mu_-$ and $b = -(\|\mu_+\|^2 - \|\mu_-\|^2)/2$, then we get a linear classifier,

$$C(X) = w^{\mathrm{T}}X + b = (\mu_+ - \mu_-)^{\mathrm{T}}X - (\|\mu_+\|^2 - \|\mu_-\|^2)/2 \tag{4}$$

When X is a positive sample, $C(X) = w^{\mathrm{T}}X + b > 0$. When X is a negative sample, $C(X) = w^{\mathrm{T}}X + b < 0$.

It is not fair to evaluate a classifier by only a cutoff such as zero. AUC is a good settlement for such problem by considering pairwise comparison. Specially, the AUC of the classier $C(X)$ should be

$$\mathrm{AUC} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I\Big(C(X_i) - C(Y_j)\Big) \tag{5}$$

where $I(x) = \begin{cases} 1 & x > 0 \\ 0 & others \end{cases}$

However, the classifier deals with all the features without difference. Noise will be added and the dominate features won't be found. We introduce a feature selection variable θ, which is a p-dimensional vector.

$$\theta = \{\theta_k,\ k = 1, 2, \ldots p\}, \quad \theta_k \in [0, 1],$$

θ_k can be treated as the probability to select the k-th feature as a member in the multi-biomarker panel. Let $\sigma_k = \theta_k w_k$, $k = 1, 2, \ldots, p$, is also a p-dimensional vector and defined as follows,

$$\sigma = \{\sigma_k,\ k = 1, 2, \ldots p\},$$

Then the above linear classifier will be

$$C(\theta, X) = \sigma^{\mathrm{T}}X + b \tag{6}$$

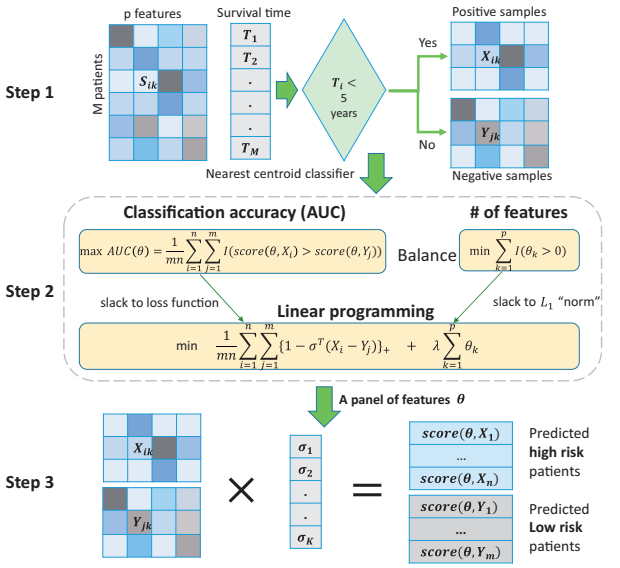

Fig. 1. The flowchart of our NCC-AUC method. Step 1, we group patients into positive and negative samples by thresholding the survival time. Step 2, by utilizing nearest centroid classifier, we maximize the classification accuracy (AUC) and minimize the number of selected features simultaneously. To balance the two objectives, we introduce a parameter λ to formulate it as an optimization problem with a single objective function. Step 3, by solving the optimization model, we could select a panel of features which maximizes its AUC score in classification

2.3 Constructing optimization model

Specifically, we want to find a vector θ to maximize the objective function as follows,

$$\mathrm{AUC} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I\Big(C(\theta, X_i) - C(\theta, Y_j)\Big) \tag{7}$$

As the paper (Zhao *et al.*, 2011) states, the value of AUC(θ) keeps the same for the classifier $c_0 C(\theta, X)$ where $c_0 > 0$, therefore, we only consider the situation $\theta \geq 0$.

Ideally, we would want the score for positive samples to be higher than that for negative samples, which yields 1 for the AUC and completely differentiates positive and negative samples. This could be measured by the Mann–Whitney U test (Siegel, 1956) to test whether the positive samples ranked higher than negative samples $P(X > Y)$.

We note that $1 - \mathrm{AUC}(\theta)$ can be interpreted as the misclassification rate. Then we formulate the feature selection and classification problem aimed for minimizing the misclassification rate (maximizing the AUC score).

$$\min_{\theta \geq 0} \quad 1 - \mathrm{AUC}(\theta)$$

However the objective function of this optimization problem is non-convex. For efficient computation, it is sensible to minimize a convex surrogate loss function. We choose the hinge loss function used in support vector machine as,

$$\min_{\theta \geq 0} \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\{1 - \Big(C(\theta, X_i) - C(\theta, Y_j)\Big)\}_+ \tag{8}$$

where $x_+ = xI(x > 0)$.

So the problem can be simplified as follows,

$$\min_{\theta\geq 0}\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\{1-W^{\mathrm{T}}(X_i-Y_j)\}_+ \quad (9)$$

In addition, we minimize the number of selected features. It is well known, L_0 'norm' is the number of non-zero entries

$$\min\|\theta\|_0=\min\sum_{k=1}^{p}\mathrm{I}(\theta_k) \quad (10)$$

In certain conditions, L_1 minimizer is equivalent to L_0 minimizer (Candes and Tao, 2005; Donoho, 2006). In practice, the following L_1-approximtion to the original L_0 -problem is heuristically used beyond these conditions.

$$\min\|\theta\|_1=\min\sum_{k=1}^{p}\theta_k \quad (11)$$

Since we have two objectives to be simultaneously minimized, a parameter λ then is introduced as follows,

$$\min_{\theta\geq 0}\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\{1-\sigma^{\mathrm{T}}(X_i-Y_j)\}_+ +\lambda\sum_{k=1}^{p}\theta_k \quad (12)$$

It is equivalently to minimize the following linear programming problem,

$$\min_{\theta,\xi_{ij}^1,\xi_{ij}^2}\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\xi_{ij}^1+\lambda\sum_{k=1}^{p}\theta_k \quad (13)$$

s.t.

$$\begin{cases}\xi_{ij}^1\geq 0,\ \xi_{ij}^2\geq 0,\ 1\leq i\leq n,\ 1\leq j\leq m\\ \xi_{ij}^1-\xi_{ij}^2=1-\sigma^{\mathrm{T}}(X_i-Y_j)\\ \sigma_k=\theta_k w_k\ \ k=1,2,\ldots,p\\ \theta\geq 0\end{cases}$$

where ξ_{ij}^1,ξ_{ij}^2 denotes the positive and negative part for the term $\sigma^{\mathrm{T}}(X_i-Y_j)$. This linear programming can be solved by many efficient algorithms.

2.4 Parameter tuning

Parameter λ is a single parameter to be tuned in our model. We chose λ by two steps: primary adjustment and further adjustment. In primary adjustment, we tested 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10 000, 100 000 for λ. Then the best parameter λ is obtained by comparing the sum of training and validating AUCs. Next step we selected the interval containing the best parameter λ for further adjustment. For example, we selected interval [0.001, 0.1] if the best parameter $\lambda=0.01$ was obtained by the primary adjustment. Further, we set step length as 0.01 and selected optimal λ in interval [0.001, 0.1]. Larger λ means fewer selected features and larger misclassification rate. In this way we balanced these two terms in the objective function and selected parameter λ with smaller misclassification and relative fewer features. All the calculations were calculated by MATLAB R2013a environment on a computer 3.40 GHz Inter Core i7-2600CPU and 16 GB memory.

2.5 Datasets

We validated our method both in genomic data and clinical data.

1. Gene expression profiles of breast cancer: Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) contains patients' overall survival time, expression profiles, and PAM50 subtype annotations of 1981 breast tumors collected from participants of the METABRIC trial (Curtis *et al.*, 2012). This dataset was accessed through Synapse (synapse.sagebase.org). The gene expression profiles contain 49 576 probe sets of 1981 patients with 328 basal-like tumors, 238 HER2+ tumors, 719 luminal A, 490 luminal B and 200 normal-like tumors. In each subtype, we randomly selected 50–60% samples as training cohort (Supplemental Table S4).
2. Clinical data of lung cancer: The clinical data of lung cancer was downloaded from http://jco.ascopubs.org/content/27/7/1091/suppl and it contained 148 stage IB NSCLC patients with 37 biomarkers (Zhu *et al.*, 2009). The 37 biomarkers included age, gender, cancer-cell type, tumor diameter and 33 immunohistochemistry biomarkers. We randomly selected approximate 2/3 samples as training cohort and the others were validation cohort.

3 Results

3.1 Performance on gene expression profiles of breast cancer

3.1.1 Functional gene panel identified by NCC-AUC in basal-like subtype

Frist of all, we divided the patients into two groups by the fact that if a patient could survive more than 5 years or not. We selected significance genes by *t*-test and kept the 2101 genes with *P*-value <0.001. In Basal-like subtype, we randomly selected 200 samples from 328 samples as training cohort and the others are treated as validation cohort. NCC-AUC selects 97 genes with training AUC $=1$ and validation AUC $=0.7207$ (Fig. 2A) by setting $\lambda=0.001$. Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) reveals that these genes contain 22 up-regulated genes with *q*-value $<10^{-15}$ and 15 down-regulated genes with *q*-value $<10^{-7}$ in basal-like subtype (Smid *et al.*, 2008; Supplementary Table S1). GSEA also shows that these genes contain 16 genes which could discriminate basal-like subtype from luminal subtype with *q*-value $<10^{-13}$ (Farmer *et al.*, 2005). These evidences support that the selected genes are closely correlated with basal-like subtype.

3.1.2 NCC-AUC outperforms SVM and SVM-RFE in basal-like subtype

The main advantage of NCC-AUC is to optimize classification accuracy and select the best biomarker panel in a single model. Here we demonstrate the performance gain of our NCC-AUC by comparing with SVM and SVM-RFE. SVM is a supervised learning model and has been widely used in machine learning (Cortes and Vapnik, 1995) to achieve high accuracy in classifying a small set of samples. SVM-RFE is one of the most successful wrapper methods in feature selection (Guyon *et al.*, 2002). SVM-RFE conducts feature selection in a sequential backward elimination manner, which starts with all the features and discards one feature at a time by checking the classification accuracy. SVM and SVM-RFE are implemented by MATLAB R2013a with default parameters.

SVM gets AUC $=1$ in training cohort while 0.5187 in validation cohort in basal-like subtype dataset (Fig. 2A). This shows that it's necessary to remove redundancy genes to avoid overfitting. If we add the feature selection option in SVM, SVM-RFE identifies a panel of 53 genes with performance of AUC $=1$ in training cohort and AUC $=0.5513$ in validation cohort (Fig. 2A). NCC-AUC performs better with AUC $=0.7207$ in validation cohort and the same AUC in training cohort. The improvement is due to the fact that

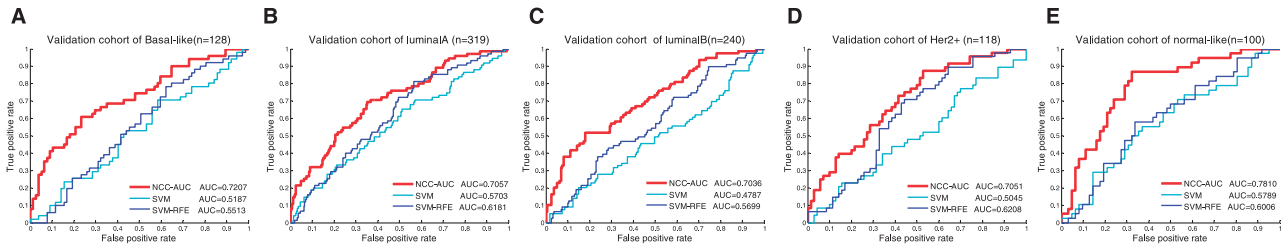

Fig. 2. Receiver operating characteristics curves for NCC-AUC, SVM and SVM-RFE in validation cohort for five subtypes of breast cancer; (**A**) for basal-like subtype, (**B**) for Her2+ subtype, (**C**) for lunminalA subtype, (**D**) for luminalB subtype, (E) for normal-like subtype. NCC-AUC consistently outperforms other methods

NCC-AUC optimizes AUC directly while SVM and SVM-RFE optimize the classification error. Also NCC-AUC regularizes the number of selected feature in a simultaneous way while SVM-RFE in a stepwise way. In conclusion, NCC-AUC shows advantage in classification and outperforms SVM-RFE in basal-like subtype.

3.1.3 NCC-AUC outperforms Cox model and L_1-Cox model in basal-like subtype

To show the advantage to treat the survival time regression problem as a binary classification problem, we compared our NCC-AUC with Cox model and L_1 penalized in Cox model. In survival analysis, Cox model and L_1-Cox model are typical methods to fully make use of the survival time while our NCC-AUC only uses whether the survival time is longer than 5 years. Cox model is one of most classical model in survival analysis, and shows efficiency for censored data (Efron, 1977). L_1 penalized Cox model has the property that it simultaneously performs variable selection and shrinkage, which makes it very useful for finding interpretable prediction rules in high-dimensional data (Goeman, 2010). Cox model and L_1-Cox model are implemented by R package 'survival' and 'penalized' with default parameters.

To investigate Kaplan-Meier analysis for the multiple-biomarker panel by NCC-AUC, we selected cutoff score based on ROC analysis. The score close to the point with best sensitivity and specificity balance was selected as the cutoff. Then we could classify samples into two classes in training and validation cohorts with the panel selected by NCC-AUC. In basal-like subtype, there are 114 predicted low risk samples and 86 predicted high risk samples in training cohort. Log-rank test demonstrates that the two groups have significant P-value < 0.00001 (Fig. 3A). In validation cohort, there are 78 predicted low risk samples and 50 predicted high risk samples. Log-rank test shows that P-value is significant (0.009) (Fig. 3B). In comparison, Cox model also shows significant difference between predicted low risk patients and predicted high patients with P-value < 0.00001 in training cohort but not significant with P-value $= 0.235$ in validation cohort (Fig. 3C, 3D). With feature selection strategy, L_1-Cox model identifies a panel of 53 genes and gets P-value < 0.00001 in training cohort and P-value $= 0.735$ in validation cohort (Fig. 3E, 3F). Here we selected the median value as the cutoff for Cox model and L_1-Cox model to convenient compare with NCC-AUC. Therefore, NCC-AUC performs better than Cox model and L_1-Cox model in basal-like subtype.

3.1.4 NCC-AUC outperforms SVM, SVM-RFE, Cox model and L_1-Cox model across other subtypes

We applied NCC-AUC to other subtypes of breast cancer and compared with SVM, SVM-RFE in classification accuracy. In Her2+ subtype, NCC-AUC identifies 57 genes with AUC $= 0.7051$ in validation cohort. This is better than SVM with AUC $= 0.5045$ and

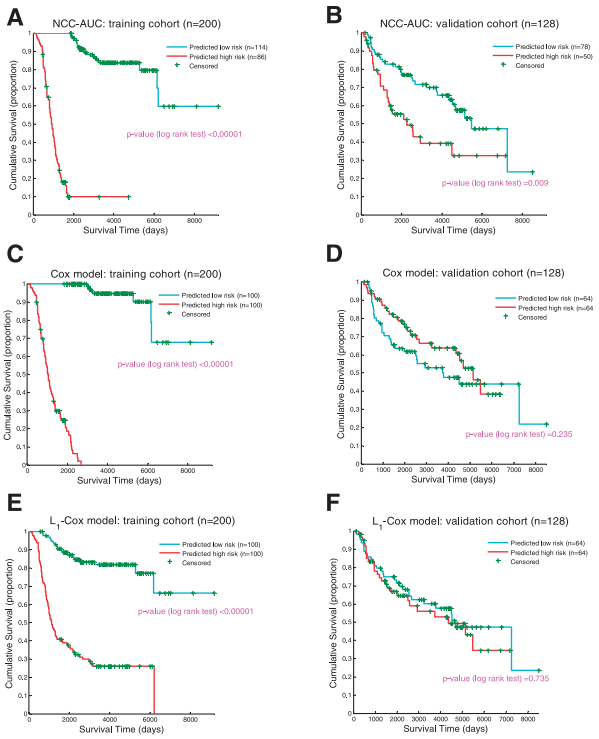

Fig. 3. Comparing NCC-AUC with existing methods in Kaplan-Meier survival estimation for basal-like subtype in training and validation cohorts. (**A**) and (**B**) show the results in training and validation cohorts for NCC-AUC. (**C**) and (**D**) show the results in training and validation cohorts for Cox model. (**E**) and (**F**) show the results in training and validation cohort for L_1-Cox model

SVM-RFE with AUC $= 0.6208$ (Fig. 2B). In other three subtypes, we also get consistent improvement (Fig. 2C–E). Especially in normal-like subtype, our method improves AUC from 0.6006 (SVM-RFE) to 0.7810 in validation cohort (Fig. 2E). These demonstrate that NCC-AUC outperforms classical classification model SVM and SVM-RFE across different subtypes.

In addition, we compared NCC-AUC with Cox model and L_1-Cox model to show its advantage in survival analysis. In luminalA subtype, log-rank test shows that Cox model (P-value $= 0.630$) is non-significant and L_1-Cox model (P-value $= 0.001$) indicates significant difference between predicted high risk patients and predicted low risk patients in validation cohort (Fig. 4). Nevertheless, NCC-AUC shows better significance than L_1-Cox model in validation cohort (P-value < 0.00001). In luminalB subtype, both Cox model (P-value $= 0.063$) and L_1-Cox model (P-value $= 0.303$) are not significant but NCC-AUC (P-value $= 0.00005$) shows significance in validation cohort. In normal-like and Her2+ subtype, NCC-AUC also shows much better

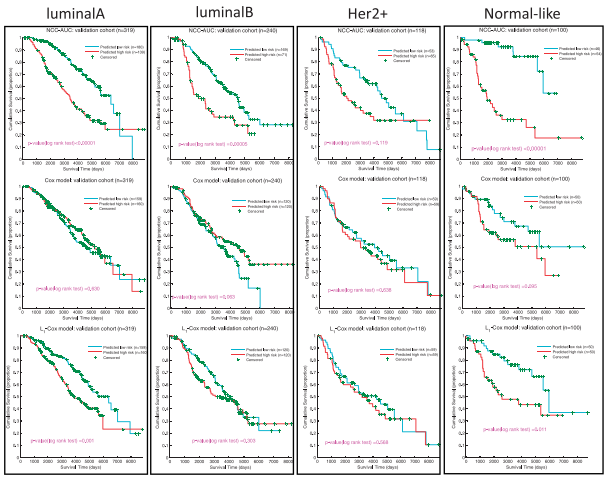

Fig. 4. Kaplan-Meier survival estimation for other subtypes of breast cancer in validation cohorts by NCC-AUC, Cox model and L_1-Cox model. NCC-AUC shows consistent improvement over existing methods

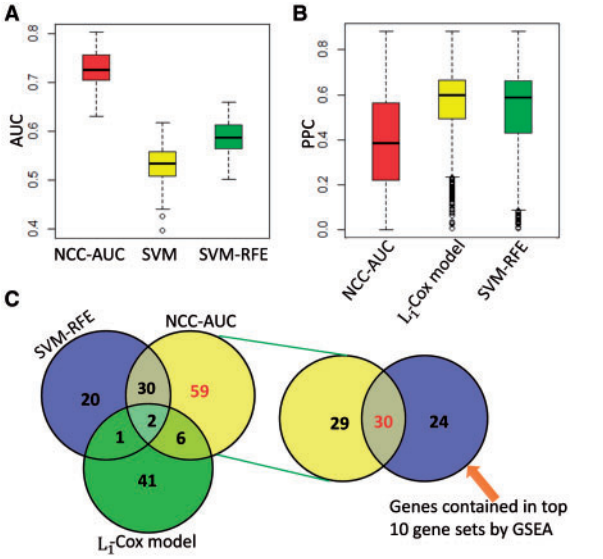

Fig. 5. (**A**) The boxplot of AUC in validation cohorts by randomly selecting 200 samples as training cohort and others as validation cohort 100 times. SVM and SVM-RFE serve as control. (**B**) The distribution of Pearson correlation coefficients (PCCs) among genes in the selected multi-biomarker panel by NCC-AUC, L_1-Cox model and SVM-RFE. (**C**) Venn diagram of genes by NCC-AUC, L_1-Cox model and SVM-RFE and Venn diagram of genes by unique genes selected by NCC-AUC and top 10 gene sets by GSEA

performance than Cox model and L_1-Cox model (Fig. 4). These evidence strongly support that NCC-AUC outperforms classical survival analysis method Cox model and L_1-Cox model across subtypes of breast cancer.

3.1.5 The performance gain by NCC-AUC is robust

The performance gain by NCC-AUC is robust to the random splitting of training and validation cohorts. To demonstrate this, we randomly selected 200 samples from 328 samples as training cohort in basal-like subtype and the others are in validation cohort. The procedure was repeated by 100 times. SVM and SVM-RFE serve as control (Fig. 5A). The mean AUC of NCC-AUC is 0.7264 and the variance is 0.0012. It significantly improves the results of SVM 0.5337 (P-value $< 10^{-89}$, Student's t-test) and SVM-RFE 0.5879 (P-value $< 10^{-74}$, Student's t-test), respectively. Those side-by-side comparisons show that NCC-AUC is a robust method and it is not influenced by randomly sampling the training cohort.

To look at the consistent improvement for the different thresholds, we also took threshold as 3, 4, 6, 7 years to divide the patients into two groups in the basal-like subtype. NCC-AUC consistently outperforms SVM and SVM-RFE at different thresholds (Supplementary Fig. S1). Moreover, we compared NCC-AUC with Cox model and L_1-Cox model. Results show that the improvement of NCC-AUC is robust (Supplementary Fig. S2).

To demonstrate NCC-AUC is robust at different thresholds, we randomly selected 200 samples from 328 samples as training cohort in basal-like subtype 100 times. Then the time-dependent AUC (Heagerty *et al.*, 2000) of validation cohort at different thresholds can be obtained. The results demonstrate that NCC-AUC is robust at the threshold = 5, 6, 7 years by Student's t-test (Supplementary Fig. S3).

3.1.6 NCC-AUC selects biological meaningful biomarkers with low redundancy

NCC-AUC is efficient in selecting biomarkers with low redundancy and identifying plausible biomarkers consistent with existing annotations and previous study. Good multi-biomarker panel should be a set of non-redundant and complementary biomarkers that maintain the maximal classification ability. To show NCC-AUC can select low redundancy biomarkers, we proposed a mean redundancy score and compared NCC-AUC with L_1-Cox model and SVM-RFE in basal-like subtype. Here we evaluated biomarker redundancy by the mean redundancy score among the identified biomarkers. Given a set of biomarkers, the average redundant score is defined as the average of pairwise Pearson Correlation Coefficients (PCC). The biomarker redundancy of NCC-AUC is 0.3897 and the improvements compared to L_1-Cox model and SVM-RFE are significant: 0.5504 (P-value $< 10^{-141}$, Student's t-test) and 0.5251 (P-value $< 10^{-107}$, Student's t-test), respectively (Fig. 5B).

To demonstrate that NCC-AUC can select more plausible biomarkers than L_1-Cox model and SVM-RFE, we further applied GSEA for the genes identified L_1-Cox model, and SVM-RFE and compared with NCC-AUC. Results show that SVM-RFE selects one gene set (SMID_BREAST_CANCER_BASAL_UP) closely associated with basal-like subtype (Supplementary Table S2). While NCC-AUC selects three gene sets related to basal-like subtype (SMID_BREAST_CANCER_BASAL_UP, FARMER_BREAST _CANCER_BASAL_VS_LULMINAL, SMID_BREAST_CANC-ER_BASAL_DN). Although SMID_BREAST_CANCER_BASA- L_UP shows up in both results, our NCC-AUC gets a 22 gene overlap with a significant q-value (1.79E-16). SVM-RFE obtains a 9 gene overlap with a q-value 1.10E-04. For L_1-Cox model, it can select 50 genes which overlap significantly with the three gene sets related to basal-like subtype (Supplementary Table S3). However, the significance levels are all less than the 97 genes identified by NCC-AUC. For example, this 50 gene set overlaps with SMID_BREAST_CANCER _BASAL_UP by 12 genes with a q-value 1.11E−08. Moreover, we investigated the genes which were uniquely identified by NCC-AUC in basal-like subtype. We found that more than half of these genes are not contained in L_1-Cox model and SVM-RFE (Fig. 5C). By checking the functional annotation, half of these genes are contained in top 10 gene sets by GSEA from the 97 genes identified by NCC-AUC.

To look at the consistency, we compared prognosis result by NCC-AUC in validation cohort in all subtypes. We found that

normal-like subtype shows the best result, followed by lunminalA, luminalB, basal-like, Her2+. Apart from Her2+, the order of other four subtypes are consistent with two studies for the same dataset (Jerby-Arnon *et al.*, 2014; Liu *et al.*, 2014). Except for normal-like subtype, the validation AUC of the other four subtypes are low, which further confers the limitation of gene expression data in breast cancer prognosis prediction. In addition, we also found that more than half of the genes selected by SVM-RFE are consistent with NCC-AUC in basal-like subtype (Fig. 5C). Taken together, NCC-AUC selects biologically meaningful genes, which are consistent with other research in basal-like subtype.

3.2 Performance in clinical data of lung cancer

3.2.1 Multi-biomarker panel identified by NCC-AUC in clinical data

In addition to genomic data, we further tested NCC-AUC in clinical data for multiple-biomarker panel identification. The training cohort contains 42 positive samples and 58 negative samples, and validation cohort contains 20 positive samples and 28 negative samples. NCC-AUC selects a panel of 12 biomarkers with AUC = 0.8284 in training cohort and AUC = 0.7589 in validation cohort by setting $\lambda = 0.013$ (Fig. 6). The panel consists of p21ras, NM23-H1, CD34-MVD, BAX, CD44v6, EMA, CEA, cyclin-D1, p27kip1, VEGF, cancer-cell type and PCNA. p21ras is an oncogene and has been reported as an independent predictors of prognosis for NSCLC (Gessner, 1998; Kim *et al.*, 1998). Besides, NM23-H1 increases CD34-MVD's expression and VEGF and may be associated with NSCLC metastasis (Che *et al.*, 2005). BAX has important regulatory roles in apoptosis and is associated with early stage NSCLC (Krajewski *et al.*, 1994; Milas *et al.*, 2003). In addition, CD44v6 is related to CEA and EMA and may play role in differentiation and progression of NSCLC (Nguyen *et al.*, 1999). These biomarkers combine together to form a panel to assist survival time prediction.

3.2.2 NCC-AUC outperforms SVM and SVM-RFE in clinical data

Compared with NCC-AUC, SVM gets AUC = 0.9020 and AUC = 0.4786 in training and validation cohort respectively (Fig. 6). It shows that NCC-AUC successfully avoids overfitting by selecting a panel of biomarkers to de-noise. SVM-RFE identifies a panel of 8 biomarkers with performance of AUC = 0.8173 in training cohort and AUC = 0.5037 in validation cohort (Fig. 6). NCC-AUC is better with training AUC = 0.8284 and validation AUC = 0.7589. Those comparison shows that NCC-AUC possesses advantage in feature selection and performs better than SVM-RFE in both training and validation cohorts of clinical data.

3.2.3 NCC-AUC outperforms Cox model and L_1-Cox model in clinical data

To investigate the survival analysis by NCC-AUC, we selected a cut-off score similar as before and then classified training and validation cohort into two groups. The training cohort contains 55 predicted low risk and 45 predicted high risk patients. Similarly, the validation cohort contains 25 predicted low risk and 23 predicted high risk patients. Log-rank test shows P-value < 0.00001 in training cohort and P-value = 0.0158 in validation cohort (Fig. 7A, 7B). Compared to NCC-AUC, Cox model shows significance in training cohort (P-value < 0.00001) while non-significance in validation cohort (P-value = 0.5939) (Fig. 7C, 7D). With feature selection strategy, L_1-Cox model identifies a panel of 17 biomarkers with P-value < 0.00001 in training cohort and 0.7166 in validation cohort

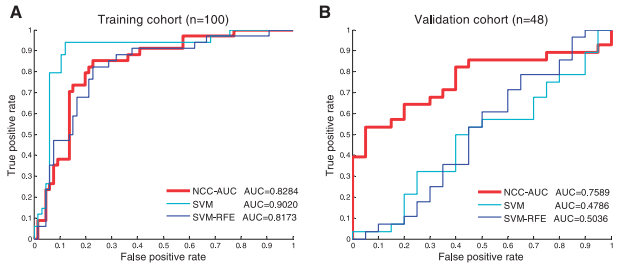

Fig. 6. Receiver operating characteristics curves for NCC-AUC, SVM and SVM-RFE in training cohort (**A**) and validation cohort (**B**) in clinical data. NCC-AUC outperforms the existing methods in validation

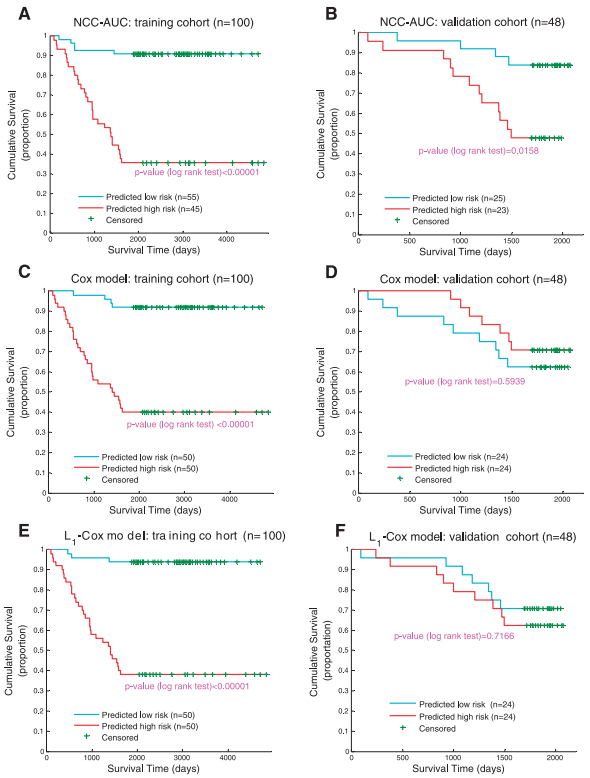

Fig. 7. Comparing NCC-AUC with existing methods in Kaplan-Meier survival estimation for Stage IB NSCLC patients in training and validation cohorts. (A) and (B) show the results in training and validation cohorts for NCC-AUC. (C) and (D) show the results in training and validation cohorts for Cox model. (E) and (F) show the results in training and validation cohort for L_1-Cox model

(Fig. 7E, 7F). NCC-AUC performs better than both Cox model and L_1-Cox model in validation cohort of clinical data. Therefore, converting survival time prediction to a classification problem with evaluating AUC accuracy significantly improves patient survival prediction and achieves better performance than classical survival models.

4 Discussion and conclusion

Selecting multi-biomarker panel in cancer survival analysis is a challenging task because of the complexity of cancer pathogenesis and the difficulty of estimating the time until event occurs. Cox model is a classical model and widely used in survival analysis and L_1-Cox model could select a panel of features for survival risk ratio estimation. However, the quality of survival time is very crucial for Cox

model and L_1-Cox model. Inspired by the fact that CI is a good measure for survival estimation, we aim to optimize CI directly. Noticing that AUC is a good approximation for CI, we treat the problem as a classification problem. This key idea allows us to propose NCC-AUC to identify meaningful biomarkers for survival analysis. NCC-AUC optimizes AUC and the number of biomarkers simultaneously thus we could select a multi-biomarker panel that has more powerful prognosis-predicting ability.

We demonstrated the good performance of NCC-AUC on not only gene expression profiles of breast cancer but also clinical data of Stage IB NSCLC. We observed that the multi-biomarker panel identified by NCC-AUC is more plausible in biology. For example, in basal-like subtype NCC-AUC identifies 97 genes and GSEA annotation shows that these genes were closely related with this subtype (Supplementary Table S1). Furthermore, these genes also show lower average redundancy than L_1-Cox model and SVM-RFE. This further demonstrates NCC-AUC's ability to remove redundant features. In addition, NCC-AUC shows robust performance gain across tumor subtypes and training cohort sampling, and it also shows consistent improvement than SVM, SVM-RFE, Cox model and L_1-Cox model at different thresholds. Finally, we validated NCC-AUC in clinical data and NCC-AUC selects the biomarkers closely associated with stage IB NSCLC. Thus NCC-AUC holds the promise to identify plausible multi-biomarker panel to assist survival analysis.

Our results show that the binary classification framework works well in real survival data analysis. The comparisons with traditional survival analysis, such as Cox model and L_1-Cox model, clearly demonstrate this point. Unlike Cox model, NCC-AUC converts survival time prediction to a classification problem and evaluates classification by AUC. Here AUC is a good approximation of CI which is a classical index in survival analysis. Importantly NCC-AUC performs better than Cox model and L_1-Cox model in validation cohort for both gene expression dataset and clinical dataset (Supplemental Tables S5, S6). Therefore, we have a good reason to believe that NCC-AUC is a useful tool for survival analysis. Especially when the survival data is noisy, NCC-AUC can be more useful to ignore the detailed values.

Further comparison with existing feature selection method SVM-RFE reveals the advantage of NCC-AUC in feature selection. Unlike SVM-RFE's stepwise way to eliminate one feature in each step, NCC-AUC allows us to select the best feature combination by solving a single optimization model. We showed that NCC-AUC is better than SVM-RFE in both datasets (Supplemental Tables S5, S6). In addition, we maximize AUC to improve the classification accuracy instead of classification error with a specified cutoff in SVM-RFE. Furthermore, NCC-AUC is formulated as a linear programming and could be solved rapidly and effectively.

We acknowledge that our current NCC-AUC is limited in several ways, which suggests future improvements. On one hand, NCC-AUC is limited in identifying much biological insights. NCC-AUC directly selects the molecular signatures for the samples with different phenotypes (whether a patient survive more than 5 years) while a posteriori justification may happen from molecular level to phenotype level then the ambiguous molecular signature may be useless. (Piao *et al.*, 2011) proposed a method to reduce the justification, which may help to solve the problem. In addition, two improvements may enhance the biological insights in the future work. One is that we should integrate the pathway information into NCC-AUC and then we will identify the network biomarker with good interpretability; the other is that, we should also take the values of survival time into account rather than only take a threshold. We tried to maximize the sum of PCCs of the selected features and survival time, which shows no significant improvement (P-value = 0.9186, Student's t-test) (Supplementary Fig. S4) in the basal-like subtype. But it integrates more information and may have a significant improvement in other cases. NCC-AUC is also limited in model construction. We classify patients by assessing whether the patient's survival time is longer than 5 years. We assume that the two classes are stable and have its own centroid distribution, which allows us to design simple nearest centroid based classifier. Besides, NCC-AUC failed to consider some patients with censored survival time. This ignores the information from censored patients with survival time less than five years. In addition, NCC-AUC is based on linear classifier and we should seek other nonlinear classification methods. Finally, AUC is only a good measure of CI and we should consider to optimize CI directly in future.

In conclusion, we convert the survival time regression problem to a classification problem motivated by optimizing CI. In practice, we evaluate classification accuracy by AUC and propose a novel optimization model to maximize it. Our main contribution is to simultaneously optimize two objectives: AUC and the number of selected features. We slack AUC by a loss function and the number of features by L_1 'norm' and solve the problem via a linear programming. We applied NCC-AUC to gene expression profiles of breast cancer and stage IB NSCLC. The results fully demonstrate the advantages of NCC-AUC by outperforming SVM, SVM-RFE, Cox model and L_1-Cox model. Our novel method holds the promise to predict the reliable prognosis of breast cancer and stage IB NSCLC. We also note that our new method can be widely extended in analyzing cancer genomics data and studying other complex diseases.

Funding

This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040700). The authors are also supported by The National Natural Science Foundation of China (NSFC) under Grants Nos. 11131009, 11422108 and 61171007.

Conflict of Interest: none declared.

References

Bonato,V. *et al.* (2011) Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, **27**, 359–367.

Breslow,N.E. (1975) Analysis of survival data under the proportional hazards model. *Int. Stat. Rev. Revue Int. de Stat.*, **43**, 45–57.

Candes,E.J. and Tao,T. (2005) Decoding by linear programming. *IEEE Trans. Inf. Theory*, **51**, 4203–4215.

Cerhan,J.R. *et al.* (2007) Prognostic significance of host immune gene polymorphisms in follicular lymphoma survival. *Blood*, **109**, 5439–5446.

Che,G. *et al.* (2005) Transfection of nm23-H1 increased expression of beta-Catenin, E-Cadherin and TIMP-1 and decreased the expression of MMP-2, CD44v6 and VEGF and inhibited the metastatic potential of human non-small cell lung cancer cell line L9981. *Neoplasma*, **53**, 530–537.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2 000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

David,C.R. (1972) Regression models and life tables (with discussion). *J. R. Stat. Soc.*, **34**, 187–220.

Donoho,D.L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory*, **52**, 1289–1306.

Efron,B. (1977) The efficiency of Cox's likelihood function for censored data. *J. Am. Stat. Assoc.*, **72**, 557–565.

Efron,B. (1988) Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Am. Stat. Assoc.*, **83**, 414–425.

Farmer,P. *et al.* (2005) Identification of molecular apocrine breast tumours by microarray analysis. *Breast Cancer Res.*, **7**, P2. 11.

Gessner,C. (1998) [Detection of mutations of the K-ras gene in condensed breath of patients with non-small-cell lung carcinoma (NSCLC) as a possible noninvasive screening method]. *Pneumologie (Stuttgart, Germany)*, **52**, 426–427.

Goeman,J.J. (2010) L1 penalized estimation in the cox proportional hazards model. *Biometrical J.*, **52**, 70–84.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Harrell,F.E.Jr. *et al.* (1982) Evaluating the yield of medical tests. *JAMA*, **247**, 2543–2546.

Harrell,F.E.Jr., Lee,K.L. and Mark,D.B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, **15**, 361–387.

Heagerty,P.J. *et al.* (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–344.

Herschtal,A. and Raskutti,B. (2004) Optimising area under the ROC curve using gradient descent. In: *Proceedings of the Twenty-First International Conference on Machine learning*. ACM, p. 49.

Jerby-Arnon,L. *et al.* (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, **158**, 1199–1209.

Kim,Y.-C. *et al.* (1998) The interactive effect of Ras, HER2, P53 and Bcl-2 expression in predicting the survival of non-small cell lung cancer patients. *Lung Cancer*, **22**, 181–190.

Koziol,J.A. and Jia, Z. (2009) The concordance index C and the Mann–Whitney parameter Pr (X > Y) with randomly censored data. *Biometrical J.*, **51**, 467–474.

Krajewski,S. *et al.* (1994) Immunohistochemical determination of in vivo distribution of Bax, a dominant inhibitor of Bcl-2. *Am. J. Pathol.*, **145**, 1323.

Liu,Z. *et al.* (2014) Breast tumor subgroups reveal diverse clinical prognostic power, *Scientific Rep.*, **4**.

Metz,C.E. (1978) Basic principles of ROC analysis. *Seminars in nuclear medicine*. **8**, 283–298.

Milas,I. *et al.* (2003) Epidermal growth factor receptor, cyclooxygenase-2, and BAX expression in the primary non-small cell lung cancer and brain metastases. *Clin. Cancer Res.*, **9**, 1070–1076.

Nguyen,V.N. *et al.* (1999) CD44 and its v6 spliced variant in lung carcinomas: relation to NCAM, CEA, EMA and UP1 and prognostic significance. *Neoplasma*, **47**, 400–408.

Parry,R. *et al.* (2010) k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J.*, **10**, 292–309.

Piao,G. *et al.* (2011) Phenotype-difference oriented identification of molecular functions for diabetes progression in Goto-Kakizaki rat. In: *Systems Biology (ISB), 2011 IEEE International Conference on*. IEEE, pp. 111–116.

Shivaswamy,P.K. *et al.* (2007) A support vector approach to censored targets. In: *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, pp. 655–660.

Siegel,S. (1956) Nonparametric statistics for the behavioral sciences. International Student Edition-McGraw-Hill Series in Psychology,1.

Simon,R.M. *et al.* (2011) Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief. Bioinf.*, **12**, 203–214.

Smid,M. *et al.* (2008) Subtypes of breast cancer show preferential site of relapse. *Cancer Res.*, **68**, 3108–3114.

Steck,H. *et al.* (2008) On ranking in survival analysis: Bounds on the concordance index. In: *Advances in Neural Information Processing Systems*. pp. 1209–1216.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Van Belle,V. *et al.* (2011) Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics*, **27**, 87–94.

Wolf,P. *et al.* (2011) The use of ROC for defining the validity of the prognostic index in censored data. *Stat. Probab. Lett.*, **81**, 783–791.

Yuan,Y. *et al.* (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.*, **32**, 644–652.

Zhao,X. *et al.* (2011) AUC-based biomarker ensemble with an application on gene scores predicting low bone mineral density. *Bioinformatics*, **27**, 3050–3055.

Zhu,Z.-H. *et al.* (2009) Three immunomarker support vector machines–based prognostic classifiers for stage IB non–small-cell lung cancer. *J. Clin. Oncol.*, **27**, 1091–1099.