Comprehensive evaluation of machine learning models and gene expression signatures for prostate cancer prognosis using large population cohorts

Ruidong Li[1,2*], Jianguo Zhu[3], Wei-De Zhong[4,5,6], Zhenyu Jia[1,2*]

Affiliations:

[1] Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

[2] Graduate Program in Genetics, Genomics, and Bioinformatics, University of California, Riverside, CA, USA

[3] Department of Urology, Guizhou Provincial People's Hospital, Guizhou, China

[4] Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

[5] Urology Key Laboratory of Guangdong Province, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University, Guangzhou, China

[6] Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, China

* To whom correspondence should be addressed: Ruidong Li at rli012@ucr.edu and Zhenyu Jia at arthur.jia@ucr.edu.

Abstract

Overtreatment of prostate cancer (PCa) remains the pervasive problem in PCa management due to the highly variable outcomes of the disease and the lack of accurate clinical tools for patient stratification. Many gene expression signatures have been developed to improve the prognosis of PCa and some of them have already been used in clinical practice, however, no comprehensive evaluation was performed to compare the performances of the signatures. In this study, we conducted a systematic and unbiased evaluation of 15 machine learning algorithms and 30 published PCa gene expression-based prognostic signatures leveraging 10 transcriptomics datasets with 1,754 primary PCa patients from public data repositories. The results revealed that survival analysis models always outperformed binary classification models for risk assessment, and the performances of the survival analysis methods - Cox model regularized with ridge penalty (Cox-Ridge) and partial least squares regression for Cox model (Cox-PLS) – were generally more robust than the other methods. Based on the Cox-Ridge algorithm, some top prognostic signatures that performed equally well or even better than the commercial panels have been identified. The findings from the study can facilitate the identification of existing prognostic signatures that are promising for further validations in prospective studies before the clinical use and the selection of the optimal approaches for the development of new prognostic models. Moreover, the study provided a valuable resource with 10 transcriptomics datasets from large primary PCa cohorts and a comprehensive collection of 30 published gene expression-based signatures that can be used to develop, validate, and evaluate new signatures for PCa prognosis.

Key worlds: prostate cancer, prognosis, gene expression signatures, machine learning algorithms, large population cohorts

Background

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men worldwide, which accounts for 14.1% of new cancer cases in 2020 [1]. Localized PCa is a highly heterogeneous disease which may lead to variable clinical outcomes. The majority of patients with slow-growing low-risk PCa only require active surveillance, whereas patients with aggressive PCa require immediate local treatment. Radical prostatectomy (RP) is the primary treatment for localized PCa with good oncologic outcomes [2]. However, approximately 20-40% of patients experience biochemical recurrence (BCR), i.e., escalated prostate-specific antigen (PSA) levels within 10 years after RP [3–5]. Despite of decade's effort, it remains challenging to predict the clinical outcomes at the time of diagnosis or following RP and identify patients needing additional treatments, such as chemotherapy, radiation or immunotherapy. Thus, overtreatment, which causes various side effects and impacts the quality of patients' lives, continues to be an issue in PCa management.

In the past decades, many gene expression-based signatures have been developed for PCa prognosis based on various computational approaches or statistical methodologies. Some of these prognostic models compared patients in two risk groups, i.e., aggressive and indolent, to identify signature genes to develop a binary classifier [6,7], while other models selected prognostic genes based on their association with the time to BCR for the prediction of relapse-free survival (RFS) [8–10]. Unsupervised approaches were also used for biomarker identification and patient stratification [11,12]. Among those gene expression-based prognostic signatures for PCa, three commercial tests have already been applied to clinical practice, including Decipher [6], OncotypeDX [13], and Prolaris [8]. Decipher, a random-forest classifier consisting of 22 markers, was developed for predicting the risk of early PCa metastasis. The OncotypeDX Genomic Prostate Score (GPS) test, which consists of 17 genes (including 5 reference genes) representing multiple biological pathways, such as stromal response, androgen signaling, cellular organization, and proliferation, can be used to predict adverse

pathology at the time of diagnosis. Prolaris, a prognostic signature with 31 genes involved in cell cycle progression (CCP), was developed for the prediction of outcome following RP or transurethral resection of the prostate (TURP).

Although many gene expression-based prognostic signatures have been developed and some of them have already been utilized in clinical practice, no systematic and unbiased comparisons have been performed to evaluate their performances on PCa prognosis. To fill this void, we leveraged ten public transcriptomics datasets consisting of 1,754 primary PCa cases to comprehensively evaluate the performances of 15 machine learning algorithms and 30 published gene expression signatures for PCa prognosis. Generally, the survival models considering the time to BCR as response variable outperformed the binary classification models which dichotomizes patients into high-risk (BCR) group and low-risk (BCR-free) group with 5-year follow-up time. Overall, two survival analysis methods, i.e., Cox-Ridge and Cox-PLS, had better performance than the other methods or algorithms. Most of the 30 prognostic signatures showed certain prognostic powers, while a few signatures, including Penney [14], Wu [10], Li [15], and Sinnott [16], had comparable as or even superior performance than the commercial panels. These promising prognostic signatures, once validated with prospective trials, may be included in the disease management to further boost the potential of PCa prognosis in clinical practice. In addition, our study showed that prediction models using the whole transcriptome as predictor variables had lower rankings than the top signatures, indicating that it's critical to identify signature genes associated with clinical outcomes from the transcriptome to achieve a high level of power for cancer prognostic models.

This is the first study that comprehensively evaluated the performances of machine learning models and published prognostic signatures using PCa population cohorts of large sizes. The results for the study are very useful to guide the selection of promising prognostic signatures for further validations and the adaption of the optimal machine learning algorithms for the development of improved prognostic models. Moreover, we have established a valuable data

resource, which consists of 10 transcriptomics datasets for a total of 1,745 PCa cases and a collection of 30 gene expression prognostic signatures, for the development, validation, and evaluation of new PCa prognostic models. The processed transcriptomics data and harmonized metadata have been deposited in the PCaDB database [17] and can be easily downloaded from the database.

Methods

Collection of PCa transcriptomics data from public data repositories

A comprehensive search for transcriptomics data of patients with primary PCa was performed in the public data repositories, including the National Cancer Institute (NCI) Genomic Data Commons (GDC) [18], cBioportal [19], National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [20], and ArrayExpress [21]. The following criteria were used for dataset selection: (i) the dataset must have a complete record of the time to BCR or the time to the last follow-up if no BCR is incurred after RP; (ii) the sample size should be greater than 80; and (iii) the dataset must be generated using a genome-wide gene expression profiling platform. A total of 10 datasets were selected in the study (Table 1).

Table 1. Summary of the transcriptomics datasets used in the study

Dataset	Sample Size	Platform	Citation
TCGA-PRAD	495	Illumina RNA-seq	[22]
CPC-Gene	213	Affymetrix Human Gene 2.0 ST Array	[23]
Taylor	131	Affymetrix Human Exon 1.0 ST Array	[24]
DKFZ	118	Illumina RNA-seq	[25]
GSE54460	100	Illumina RNA-seq	[9]
Cambridge	125	Illumina HumanHT-12 V4.0	[12]
Stockholm	94	Illumina HumanHT-12 V4.0	[12]
CancerMap	129	Affymetrix Human Exon 1.0 ST Array	[11]
CIT	101	Affymetrix Human Gene 2.0 ST Array	[26]
Belfast	248	Almac Diagnostics Prostate Disease Specific Array	[27]

The HTSeq-Counts data from The Cancer Genome Atlas Prostate Adenocarcinoma (TCGA-PRAD) project were downloaded and preprocessed using the R package *GDCRNATools* [28]. The raw .CEL files of the four Affymetrix microarray datasets including CPC-Gene (GSE107299), Taylor (GSE21034), CancerMap (GSE94764), and CIT (E-MTAB-6128) were downloaded from GEO/ArrayExpress and normalized with the Robust Multichip Average (RMA) method implemented in the R package *oligo* [29]. The raw sequencing data for the GSE54460 dataset was downloaded from SRA (https://www.ncbi.nlm.nih.gov/sra) under the accession number SRP036848 using *fasterq-dump* in the SRA Toolkit (version 2.10.8). *STAR* (version 2.7.2a) [30] was used for sequence alignment and *featureCounts* (version 2.0.0) [31] was used for gene expression quantification. The count data was then normalized using the Trimmed Mean of M values (TMM) method implemented in the R package *edgeR* [32]. The reads per kilobase per million mapped reads (RPKM) values for the DFKZ dataset was downloaded from cBioPortal and log2 transformation was performed. The processed intensity data of the other datasets including Cambridge (GSE70768), Stockholm (GSE70769), and Belfast (GSE116918) were downloaded directly from GEO using the R package *GEOquery* [33]. The *ExpressionSet* objects of the processed data, including the normalized gene expression data and harmonized metadata, were deposited in the PCaDB database [17]. The gene expression values in each dataset were rescaled by z-score transformation for model evaluation.

Collection of published gene expression signatures for PCa prognosis

Gene expression-based signatures for PCa prognosis were collected by an inclusive literature screening. The keywords ‘prostate cancer’, ‘prognosis, and ‘gene expression signature’ were used to search the PubMed database. Some signatures were identified in recent review papers on prostate cancer prognostic signatures or in research papers with signature comparisons [34–36]. Different types of gene identifiers may be reported in the original papers, so the gene IDs were harmonized by searching against the Ensembl [37], the HUGO Gene Nomenclature Committee (HGNC) [38], and the NCBI Entrez Gene [39] databases.

A total of 30 gene expression signatures for PCa prognosis were collected and evaluated in the study (Table 2). The gene lists of the 30 published prognostic signatures were provided in Table S1.

Table 2. Summary of the 30 gene expression signatures for PCa prognosis

Signature Name	Alternative Name	No. of Genes	Year	Citation
Agell		12	2012	[40]
Bibikova		16	2007	[41]
Bismar		12	2006	[42]
Cuzick	Prolaris	31	2011	[8]
Ding		4	2011	[43]
Erho	Decipher	19	2013	[6]
Glinsky		11	2005	[44]
Irshad		19	2013	[7]
Jia		15	2016	[45]
Kamoun	CIT36	36	2018	[26]
Klein	OncotypeDX	12	2014	[13]
Li		160	2020	[15]
Long		24	2014	[9]
Luca	DESNT	45	2018	[11]
Mo	SDMS	93	2018	[35]
Nakagawa		17	2008	[46]
Olmos	LPD1	9	2012	[47]
Penney		157	2011	[14]
Planche		36	2011	[48]
Ramaswamy		17	2003	[49]
Ramos-Montoya		222	2014	[50]
Ross		6	2012	[51]
Ross-Adams		100	2015	[12]
Sharma		16	2013	[52]
Sinnott		30	2017	[16]

Talantov	3	2010	[53]
Varambally	44	2005	[54]
Wu	29	2013	[10]
Yang	28	2018	[36]
Yu	14	2007	[55]

Binary classification algorithms

Nine binary classification algorithms were evaluated in the study, including Elastic Net, support vector machines (SVM) using the linear function (SVM-Linear), the polynomial kernel function (SVM-Poly), and the radial basis function (SVM-RBF), random forest (RF), partial least square (PLS), linear discriminant analysis (LDA), XGBoost using the linear booster (XGBoost-Linear), and the tree booster (XGBoost-Tree).

The R package *caret* [56], which has a set of functions to streamline the process for creating predictive models, was used for model training and parameter tuning. In each model, the predictor variables were the genes in a given signature and the response variable was the 5-year BCR status, where 0 represents non-BCR and 1 represents BCR incurred within 5 years after RP. The grid search with a tuning length of 10 was used for tuning parameters. The 10-fold cross validation resampling scheme and ROC metric were used to select the optimal model. A probability score for the BCR class was calculated for each patient, where a greater score indicates a higher probability to experience BCR.

Survival analysis methods

Six survival analysis methods were evaluated in the study, including Cox proportional hazards (CoxPH), Cox model regularized with ridge penalty (Cox-Ridge) and lasso penalty (Cox-Lasso), supervised principal components (SuperPC), partial least squares regression for Cox models (Cox-PLS), and random survival forest (RSF).

The CoxPH models were built using the R package *survival (*https://CRAN.R-project.org/package=survival*)*. A linear combination of the expression values and coefficients of the signature genes were computed as the risk scores for the patients. Cox-Ridge and Cox-Lasso models were built using the R package *glmnet* [57] with the penalty α=0 and α=1, respectively. The R package *superpc* [58] was used to build the SuperPC models. A subset of genes in a signature with the univariate regression coefficients exceeding the threshold of 0.3 was used to calculate the principal components to build the model. Cox-PLS models were built using the R package *plsRcox* [59] and two components were included in the models. The R package *randomForestSRC* [60] was used to build the RSF models with 100 trees.

Comparison of machine learning models and prognostic signatures

Intra-dataset (within dataset) comparison

The 10-fold cross-validation (CV) was performed for any dataset to evaluate the performance of a signature using a predictive model within the dataset. In a 10-fold CV, the patient cases were randomly partitioned into ten portions with approximately equal size. In each iteration, nine portions were used as the training set to develop the model, and the remaining one portion was retained as the test set for model evaluation. This process was repeated ten times with each of the ten portions was used exactly once as the test data. The training and test folds were exactly the same for all the predictive models.

Inter-dataset (across datasets) comparison

For the inter-dataset comparisons, the models were trained with one dataset and then tested by the other nine independent datasets. Note we have selected ten datasets that met the three selection criteria for this study.

Evaluation Metrics

We used three metrics, i.e., concordance index (C-index), time-dependent receiver operating characteristics (ROC) curve, and hazard ratio (HR) estimated by the

Kaplan Meier (KM) analysis, to evaluate the performances of the machine learning models and prognostic signatures. The C-index were calculated using the R package *survcomp* [61], the area under the ROC curves (AUC) were estimated using the R package *survivalROC* [62], and KM analyses were performed using the R package *survival* to estimate the HR and 95% confidence intervals (CIs). For the KM analyses, the median values of the risk scores were used to dichotomize the patients into low- and high-risk groups.

Differential expression analysis, CoxPH survival analysis, and functional enrichment analysis for the signature genes

The R packages *limma* [63] was used to perform differential expression (DE) analysis and *survival* was used for CoxPH survival analysis to investigate whether the genes in the prognostic signatures were differentially expressed between tumor and normal samples, and if they were significantly associated with RFS outcomes based on the TCGA-PRAD data. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Disease ontology (DO) enrichment analysis were performed using the R package *clusterProfiler* [64].

Results

Overview of the study design

A total of 50 public transcriptomics datasets of patients with primary PCa were identified from the public data repositories, including GDC, cBioportal, GEO, and ArrayExpress, while only 14 of them have the record of BCR data. Eventually, 10 transcriptomics datasets (a total of 1,754 cases) which met three data selection criteria (*see Methods*) were included in the study. We gleaned 30 published gene expression prognostic signatures for PCa through a comprehensive literature search. Nine binary classification algorithms and six survival analysis methods were used to build the predictive models based on the expression data of genes in each PCa prognostic signature. The comprehensive comparisons included: (i) the predictive abilities between binary classification models versus survival analysis models; (ii) the performances of different survival algorithms; and (iii) the

prognostic powers of the 30 gene expression signatures based on the optimal survival algorithm identified in (ii). Both intra-dataset comparison (10-fold CV within one dataset) and inter-dataset comparison (one dataset as a training set and the other independent datasets as test sets) were conducted, and three metrics including C-index, time-dependent AUC, and HR based on KM survival analysis were used to evaluate the models and signatures. The overall design of the study was presented in Figure 1.

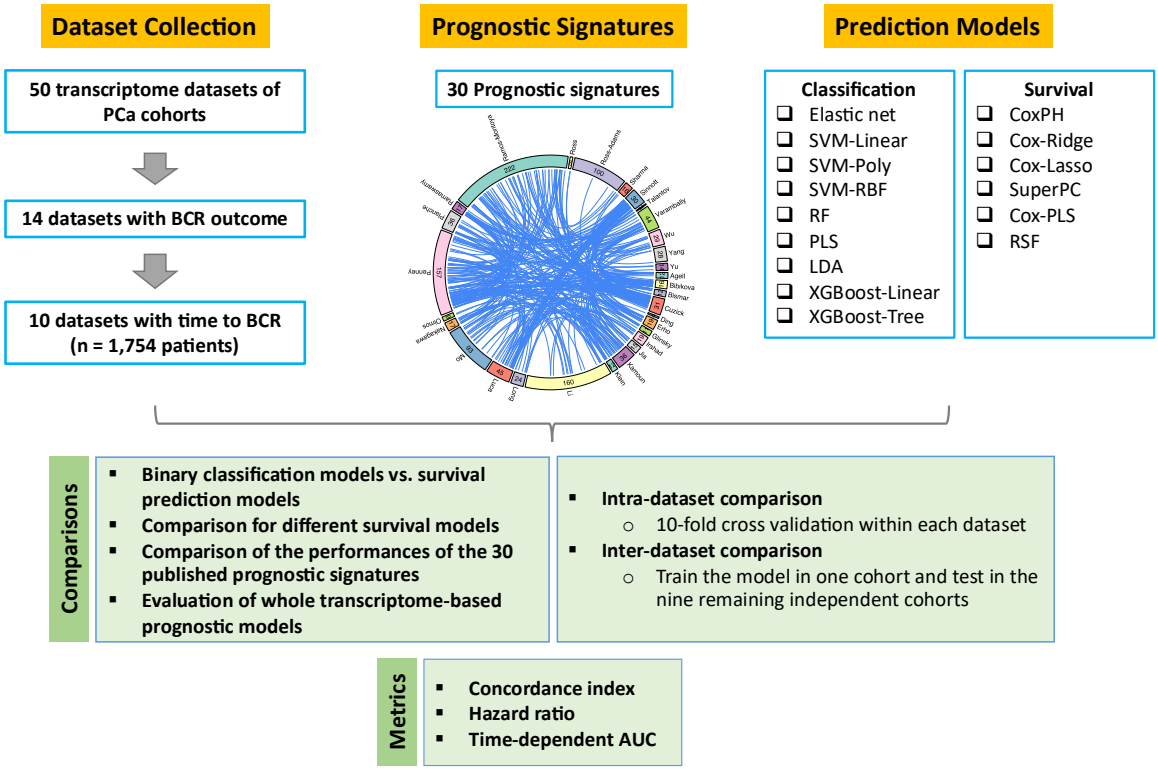

Figure 1. Overview of the study design.

Functional characterization of the signature genes

In total, there were 1,032 unique genes in the 30 published prognostic signatures for PCa. Gene expression profiles were compared between the primary tumor samples versus the tumor-adjacent normal samples by leveraging the TCGA-PRAD data. The result showed that 223 among the 1,032 genes (21%) were differentially expressed with the absolute fold change (FC) > 2 and the false discovery rate (FDR) < 0.01 (Figure 2A). The CoxPH survival analysis based on

the TCGA-PRAD data showed that 558 out of the 1,032 genes (54%) were significantly associated with RFS with P values < 0.05 (Figure 2B). The 1,032 genes were significantly enriched in many important pathways in cancer, including the cell cycle pathway, apoptosis pathway, p53 signaling pathway, FoxO signaling pathway, PI3K-Akt signaling pathway, prostate cancer pathway, bladder cancer pathway, etc. (Figure 2C), indicating the biological importance of these signature genes. The genes involved in those pathways were usually included in many signatures (Figure S1).

Analysis of the gene overlap among different signatures indicated that the prognostic signatures were largely distinct in terms of gene identity (Figure 2D). For example, a maximum of three genes were shared between the Erho signature (Decipher) and the Penney signature, while only one or two common genes were identified between the Erho signature and eight other signatures. Among the 1,032 genes, 142 genes (13.8%) were shared by more than two signatures and 40 genes (3.9%) were common in three or more signatures (Figure 2E). Based on the TCGA-PRAD data, 28 among the 40 common genes (70%) were significantly differentially expressed between the primary tumor samples versus the tumor-adjacent normal samples, and 36 out of the 40 common genes (90%) were significantly associated with RFS. The disease ontology analysis showed that these 40 genes were primarily associated with cancers, including prostate cancer (Figure 2F). Taken together, the results suggested that a lot of the signature genes were indeed relevant to PCa and disease outcomes, and the lack of overlap among these signatures indicated the potential of improvement in PCa prognosis by integrating these signature genes.

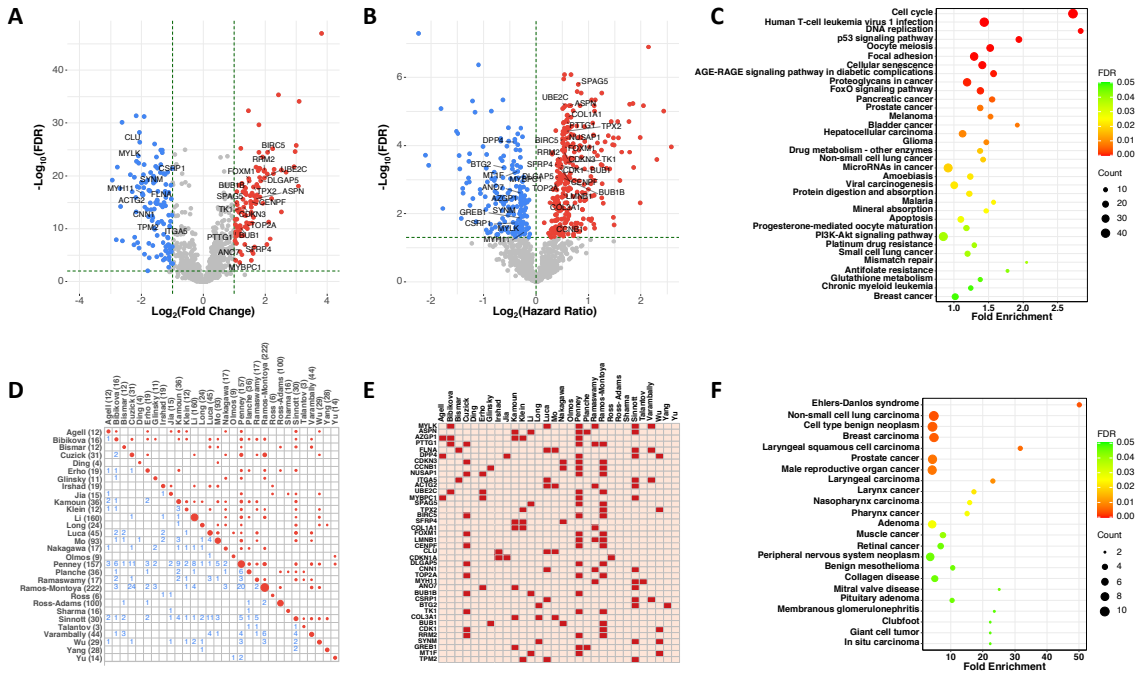

Figure 2. Functional characterization of the signature genes. (**A**) Differential expression analysis of the 1,042 genes between tumor vs normal samples based on the TCGA-PRAD data. Genes in more than 3 signatures were labeled. (**B**) CoxPH survival analysis of RFS for the 1,042 genes based on the TCGA-PRAD data. Genes in more than 3 signatures were labeled. (**C**) KEGG pathways that the 1,042 genes enriched in. **(D)** Overlap of genes among different prognostic signatures. (**E**) The 40 common genes that were used in three or more signatures. (**F**) DO enrichment analysis of the 40 common genes in three or more signatures.

Comparison of survival analysis models versus binary classification models

Two approaches were commonly used to analyze the survival data to develop cancer prognostic models, depending on how the response variable is defined: (i) binary classification (relapse vs. non-relapse): patients are dichotomized into two risk groups based on whether the outcome event is observed within a cutoff of follow-up time (eg., 5 years), and (ii) survival time: a patient either has definitive outcome (relapse and time to relapse) or the relapse-free follow-up time with right-censoring. We compared the prediction performances of nine binary classification algorithms and six survival analysis methods in risk assessment. The expression

levels of the genes in each signature were used as the predictor variables in these models, and the median values of the three metrics, i.e., C-index, AUC, and HR, were calculated across all the signatures and the datasets to evaluate these models. For the intra-dataset comparison, 10-fold CV was used for model evaluation, whereas for the inter-dataset comparison, one dataset was treated as the training set and the other datasets were used as the test datasets at a time. The results indicated that almost all the survival models were superior to binary classification models in the intra-dataset comparison (Figure 3). The median values of C-index, AUC, and HR across all the signatures and the datasets for the survival models were generally greater than those for the binary classification models. The results in the inter-dataset comparison were consistent with those in the intra-dataset comparison. Some of the top binary classification models such as PLS, elastic net, and RF performed equally well or slightly better than the CoxPH or SuperPC survival models, however, the top survival models including Cox-Ridge, RSF, and Cox-PLS always outperformed all the binary classification models in the inter-dataset comparison.

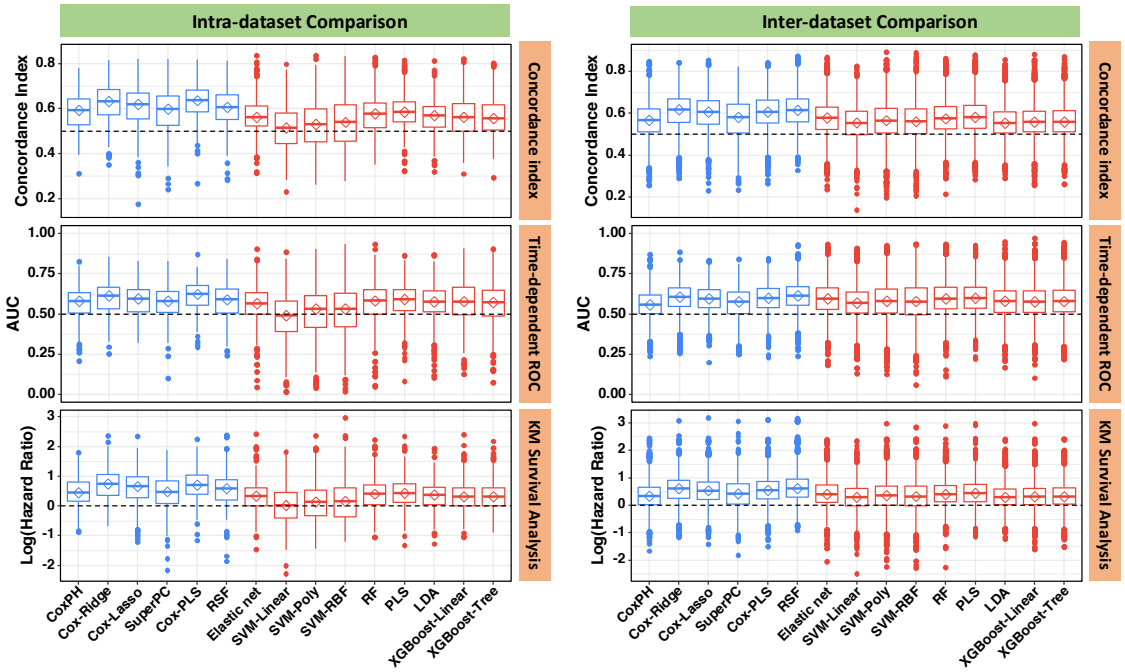

Figure 3. Comparison of survival analysis models versus binary classification models. (blue: survival models; red: classification models)

There are two possible reasons why the survival analysis models were superior to the binary classification models: (i) the sample size became smaller in binary classification models if the censored data were excluded, and (ii) the variability in survival time provides more information than binary outcome, leading to increased data resolution and therefore statistical power. To test these hypotheses, we compare the survival models with the binary classification models using the same data for which the patients with censored data were removed. The results indicated that the prediction performances for all the survival models increased as sample size (Figure S2), and the survival models generally outperformed the binary classification models (Figure S3), which supported our hypotheses.

Performance evaluation of the survival analysis methods

A comprehensive evaluation of the performances of the six algorithms for survival analysis were carried out (Figure 4). For the intra-dataset comparison based on 10-fold CV, Cox-Ridge and Cox-PLS models generally performed better than the other models. The performances of the SuperPC and RSF algorithms substantially varied and largely depended on the datasets that were used for analysis, while CoxPH usually ranked bottom among six survival analysis algorithms. For example, based on the C-index, the performances of RSF were very similar to the top 2 algorithms Cox-Ridge and Cox-PLS in the DKFZ, GSE54460, and Cambridge datasets, whereas the performances of RSF ranked the last or the second last in the CPC-Gene, Stockholm, CancerMap, CIT, and Belfast datasets. The performances of Cox-Lasso usually ranked higher than RSF, SuperPC, and CoxPH, but lower than Cox-Ridge and Cox-PLS. The result was slightly different for the inter-dataset comparison, where the performances of Cox-Ridge, RSF, and Cox-PLS were very similar and generally better than the other algorithms. The ranks of Cox-Lasso, CoxPH, and SuperPC were consistent with that in the intra-dataset comparison, where Cox-Lasso > SuperPC = CoxPH.

It was as expected that the survival analysis models usually performed better in the intra-dataset comparison than that in the inter-dataset comparison. This was

mainly because that in the intra-dataset comparison, the training and test data were generated in the same study using the same technology, whereas in the inter-dataset comparison, different technologies, different platforms, and even different bioinformatics pipelines may be used by the training set and test set. The cohorts may also be quite different between studies or datasets, yielding increased level of heterogeneity in inter-dataset analysis. These results in the study emphasized the importance of validation with independent cohorts when building and evaluating prognostic models.

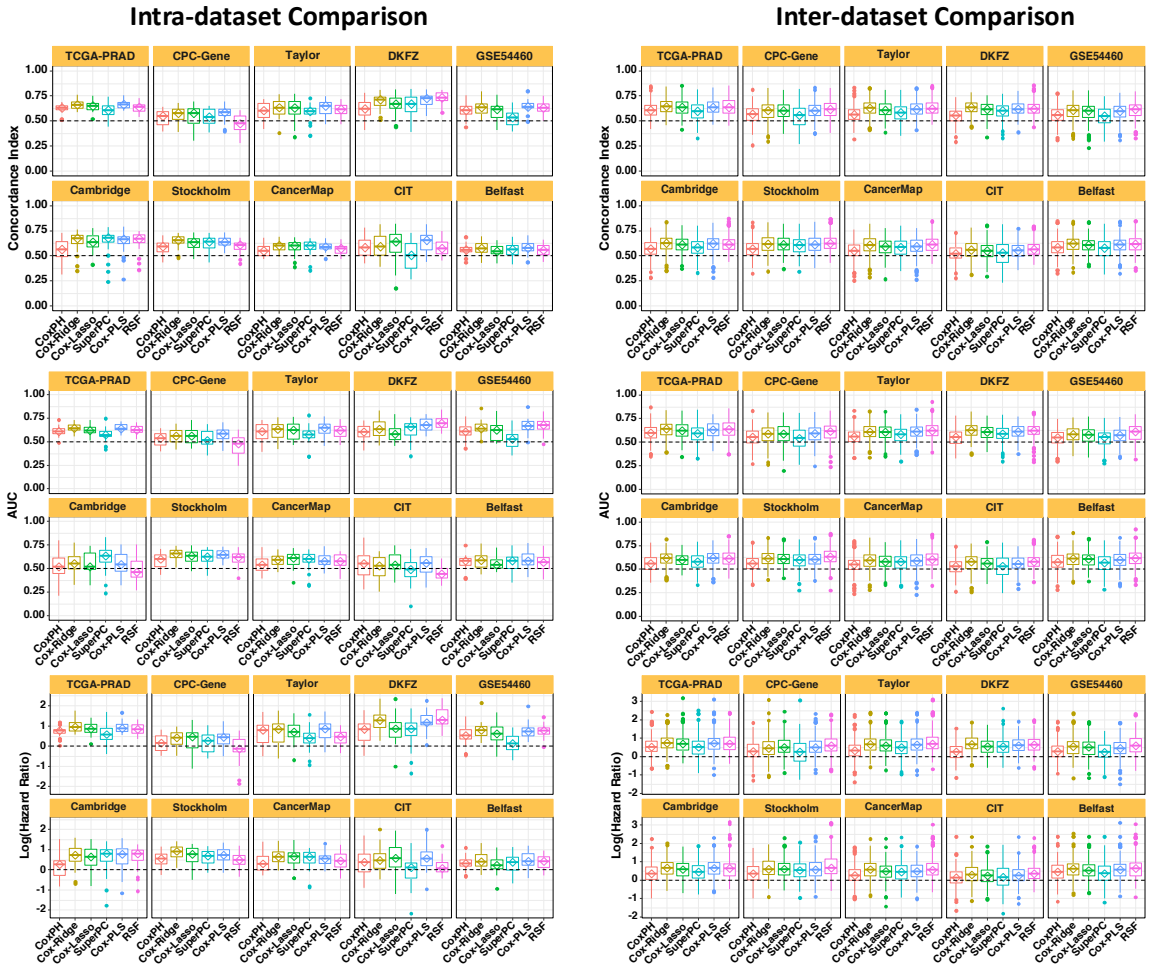

Figure 4. Comparison of the survival analysis algorithms based on the three metrics (C-index, AUC, and HR) in the intra-dataset and inter-dataset comparisons. (The title of each panel refers to the training dataset)

Evaluation of the published gene expression signatures for PCa prognosis

Based on the comprehensive comparisons for different survival analysis algorithms, Cox-Ridge, which performed consistently well across all the datasets was selected to further evaluate the performances of various prognostic signatures for PCa. The median values of the three metrics were calculated across all the datasets and were used to evaluate these prognostic signatures. The results indicated that almost all the signatures had some prognostic powers with median C-indexes and median AUCs greater than 0.5 and median HRs greater than 1 both in the intra-dataset and in the inter-dataset comparisons. Some prognostic signatures, including Penny, Li, Klein, Sinnott, Wu, Erho, Kamoun, Planche, and Long, almost always ranked in the top 10 signatures based on the three metrics. Two of the three commercially applied prognostic signatures, i.e., Klein (OncotypeDX) and Erho (Decipher), performed consistently well across the datasets, especially that Klein always ranked in the top 5 signatures in the intra-dataset comparison and ranked the second in the inter-dataset comparison. The median values of C-index, AUC, and HR for the signature Penney usually ranked the first, which was the only signature performing better than the Klein signature in most comparisons. Some of the other top signatures including Li, Sinnott, and Wu always had higher rankings than the Erho signature in the intra-dataset comparison, and Wu also outperformed the Erho signature in the inter-dataset comparison.

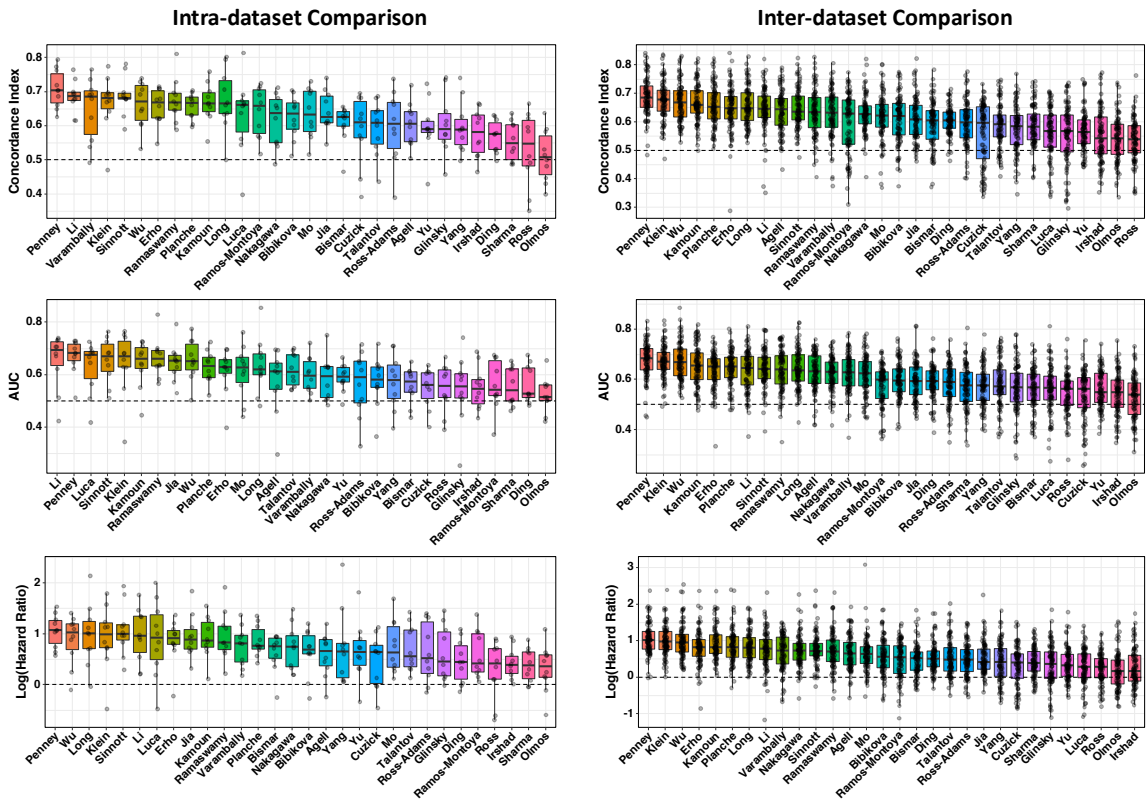

Figure 5. Performance evaluation of the prognostic signatures based on the Cox-Ridge method and the three metrics (C-index, AUC, and HR) in the intra-dataset and inter-dataset comparisons

We also did similar comparisons using the other well-performed algorithm Cox-PLS to investigate whether the performances of the signatures relied on the methods for model development. The results indicated that although the rankings may be slightly different based on different algorithms, the top signatures using the Cox-Ridge algorithm also ranked in the top list when Cox-PLS was used (Figure S4). The signature Ramaswamy performed slightly better using the Cox-PLS algorithm, which made it always ranked among the top 10 signatures.

Since the test characteristics, i.e., C-index, AUC, and HR, may vary when different datasets were used as training set and test set, we utilized the rankings of the median values of these three metrics across all possible combinations of training and test sets to reflect the overall performances of the signature being

evaluated. Figure 6 showed details about the HR-based performances of the signatures for each dataset in the intra-dataset comparison, while Figure 7 revealed such performances when each dataset was used as training set in the inter-dataset comparison. The rankings of the top signatures such as Penny, Li, Klein, Sinnott, Wu, Erho, etc. were generally high across multiple comparisons. For example, in the intra-dataset comparison, the signature Sinnott always ranked among the top five signatures in six datasets, including CPC-Gene, Taylor, DKFZ, GSE54460, CancerMap, and Belfast, whereas the signature Penney ranked among the top five signatures in five datasets, including GSE54460, Cambridge, Stockholm, CancerMap, and Belfast (Figure 6). In the inter-dataset comparison, the signature Penney and Klein can be independently validated in more than half of the nine test datasets no matter which training dataset was used when building the model (Figure 7). The signature Wu can also be validated by multiple independent cohorts when different training sets were used. We also observed that some signatures may perform well when certain datasets were used for model development and validation, but they could show very low or even no prognostic power when other datasets were tried. The results suggested the importance of leveraging data from multiple cohorts to develop and validate new prognostic models or to evaluate any existing prognostic model.

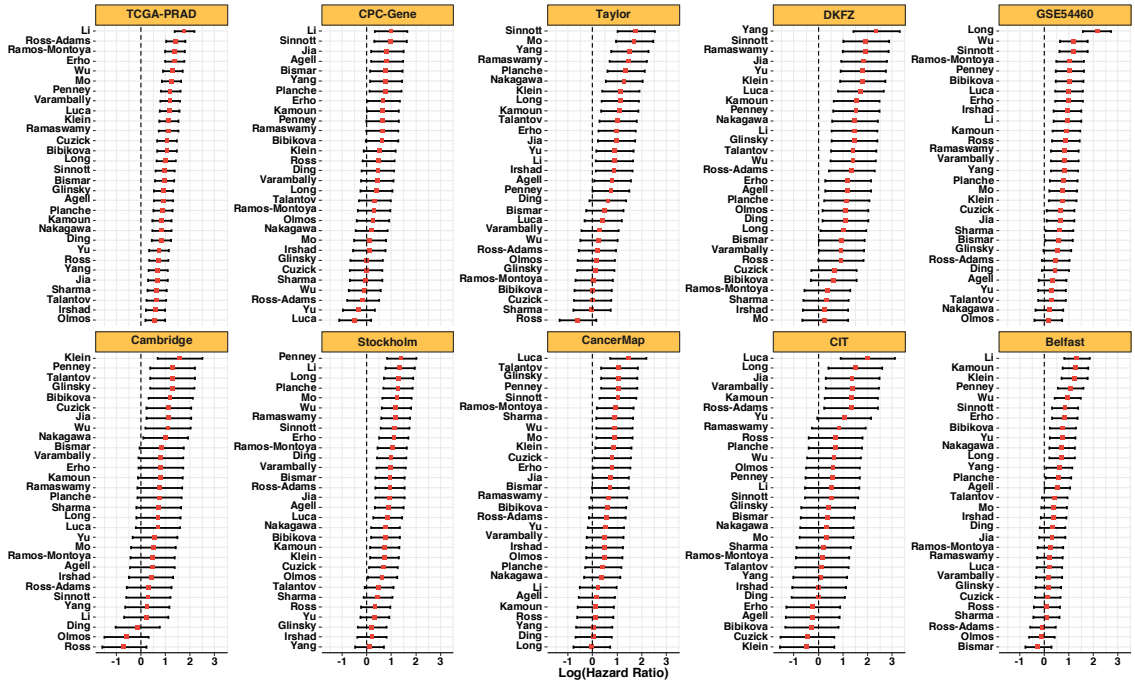

Figure 6. Performances of the signatures based on HRs from KM survival analysis for each dataset in the intra-dataset comparison. (The title of each panel refers to the training dataset)

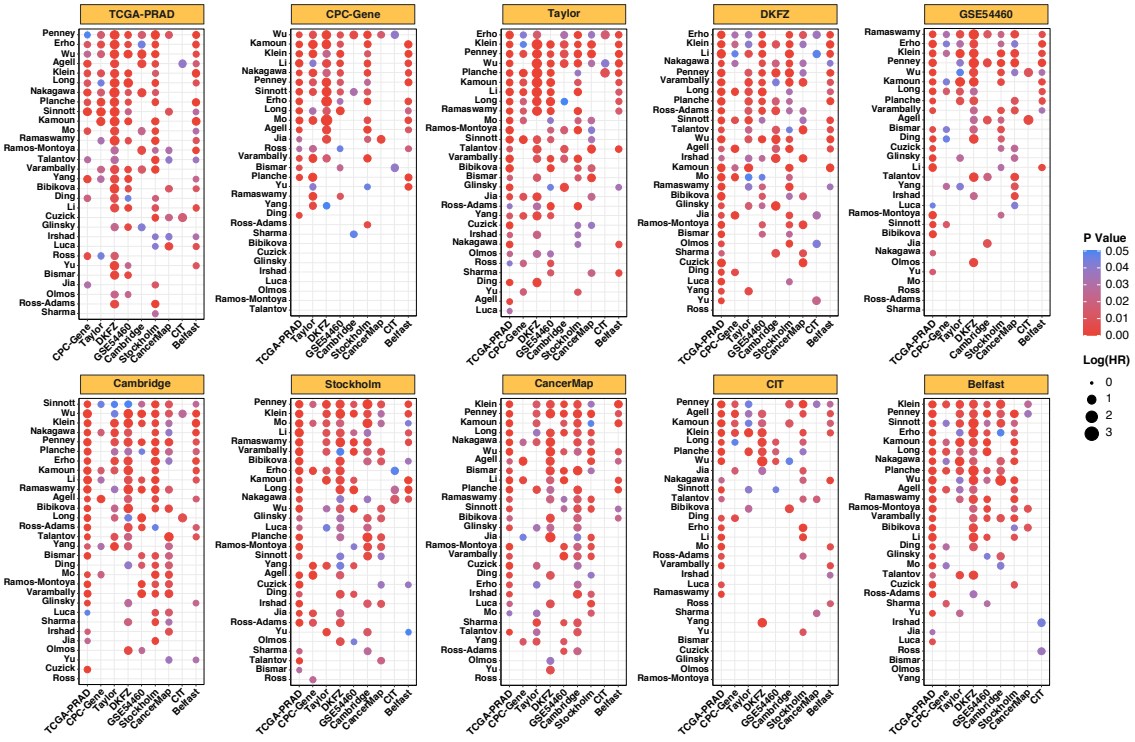

Figure 7. Performances of the signatures based on HRs from KM survival analysis for each training-test combination in the inter-dataset comparison. (The title in each panel refers to the training dataset; x-axis indicates the test datasets and the y-axis indicates the signatures; The signatures were ranked based on the number of datasets that successfully validated the predictive models)

PCa prognostic model using the whole transcriptome

In addition to the comparisons among the prognostic signatures, we also investigated the performance of prediction models leveraging the whole transcriptome for PCa prognosis. The performances of the survival models with the entire transcriptome were similar to that in the comparison of various prognostic signatures, i.e., Cox-Ridge, Cox-PLS and RSF generally performed better than Cox-Lasso and SuperPC (Figure S5). The CoxPH was not able to be evaluated because it cannot handle the data with $p >> n$, where p is the number of parameters and n is the sample size.

Similarly, the Cox-Ridge algorithm was used to evaluate the prognostic power when the whole transcriptome was factored in the models. The models using the whole transcriptome generally performed reasonably well with the median C-index, AUC, and HR ranked the 7th, 10th, and 11th, respectively, in the intra-dataset comparison, whereas these models ranked the 10th, 12th, and 13th, respectively, in the inter-dataset comparison (Figure S6). The results revealed that although the performance of the models using the whole transcriptome was better than two third of the signatures, prediction accuracy may be further boosted by well selected outcome-associated transcripts.

We also investigated the performances of the models with the whole transcriptome in each dataset and the conclusion is the same that although the models performed well in some datasets, they were generally not as good as the top signatures (Figure S7). The sum of the evidence indicated that it is desirable to develop PCa prognostic models using the most significant outcome-associated genes selected from transcriptome.

Discussion

A systematic and unbiased evaluation of the machine learning models and the gene expression signatures for cancer prognosis using large cohorts is very critical to select the optimal approaches for building the predictive models and identify the most promising signatures for further validations in prospective clinical studies before they can be applied to clinical practice. Although validations with independent cohorts and comparisons with some existing signatures were usually performed for most of the newly developed PCa signatures in the original studies, no comprehensive evaluation of these signatures and prognostic models has been conducted. In this study, we identified 30 gene expression signatures for PCa prognosis, collected the transcriptomics data for 1,754 patients (10 cohorts) with primary prostate cancer from public data repositories, compared the performances of nine binary classification algorithms and six survival analysis methods leveraging these datasets, and evaluated the performances of the 30 signatures based on the most robust machine learning algorithm.

Because different statistical approaches and different training datasets may be used to derive the signatures and develop the predictive models in the original papers, it would be very challenging to replicate these prognostic models. The focus of the study is not to reproduce the exact predictive models, but to evaluate the performances of the signatures based on the same machine learning algorithms and the same training and validation datasets in a systematic and unbiased manner. The results from our study revealed that the survival analysis algorithms which took both the status of the event (i.e., BCR) and the time to the event into consideration generally outperformed the binary classification algorithms. This is mainly because that the censored data were removed when building the binary classifiers which resulted in smaller sample size, and the binary classification algorithms themselves were not as good as the survival analysis algorithms for risk assessment. Among the six survival analysis methods, Cox-Ridge and Cox-PLS generally performed better than the other algorithms, whereas the widely used method CoxPH usually performed the worst or the second worst.

The performances of RSF were not good in the intra-dataset comparisons based on 10-fold CV, but were very robust in the inter-dataset comparison using one dataset as the training data and the other independent datasets as the test data. It is as expected that the survival analysis models usually performed worse in the inter-dataset comparison than that in the intra-dataset comparison because of the high heterogeneities between different cohorts. This emphasized the importance of clinical validations in independent cohorts for the development and evaluation of prognostic models.

Although there were not much overlaps of genes among the signatures, the comprehensive evaluation of the gene expression signatures showed that most of the signatures had some prognostic powers with median C-indexes and median AUCs greater than 0.5, and median HRs greater than 1. The two commercial tests, OncotypeDX and Deciper almost always ranked in the top signatures in the comparisons. Many other signatures such as Penney, Wu, Li, and Sinnott generally had equally well or even better performances than the commercial panels. These four signatures and some other top signatures including Kamoun, Planche, and Long could be promising for developing robust prognostic models to be incorporated into clinical practice with further prospective validations. The promising signatures were identified mainly based on the overall performances. It's possible that a signature could perform well using a given machine learning algorithm and a given training data to develop the predictive model, while could have no prognostic power when other algorithms or training datasets were used. Multiple survival analysis algorithms and large PCa population cohorts were recommended when developing new prognostic signatures. We also evaluated the performances of prognostic models using the whole transcriptome data and found that the transcriptome models could outperform some of the signatures but were not comparable with the top signatures, indicating that it's critical to derive gene signatures that are relevant to PCa outcomes from the transcriptome to further improve the PCa prognosis.

In summary, this is the first study that performed a comprehensive evaluation of the performances of the machine learning models and the published gene expression signatures using large PCa cohorts. The findings from the study can guide the selection of existing prognostic signatures for further validations and the selection of the optimal algorithms for the development of new signatures. In addition, our study also provided a valuable resource with 10 transcriptomics datasets from large primary PCa cohorts and a comprehensive collection of 30 gene expression prognostic signatures that can be used to develop, validate, and evaluate new signatures for PCa prognosis.

Data Access

The processed data including the normalized expression data and the harmonized metadata of all the 10 transcriptomics datasets were deposited in the PCaDB database (http://bioinfo.jialab-ucr.org/PCaDB/) and the *ExpressionSet* object of each dataset can be easily downloaded from this database. All the scripts used in this study, including data preprocessing, model comparison, and data visualization, are publicly available at https://github.com/rli012/PCaSignatures.

Funding

This work was supported by Z.J.'s UC Riverside Faculty Start-up Fund and UC Cancer Research Coordinating Committee Competition Award. J.Z. was supported by the Science and Technology Project of Guizhou Province in 2017 ([2017]5803), the High-level innovative talent project of Guizhou Province in 2018 ([2018]5639), and the Science and Technology Plan Project of Guiyang in 2019 ([2019]2-15). W.Z. was supported by the grants from National Natural Science Foundation of China (82072813, 8157142) and Guangzhou Municipal Science and Technology Project (201803040001).

Disclosure Declaration

The authors declare that they have no competing interests.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71:209–49.

2. Hruza M, Bermejo JL, Flinspach B, Schulze M, Teber D, Rumpelt HJ, et al. Long-term oncological outcomes after laparoscopic radical prostatectomy. BJU International. 2013;111:271–80.

3. Freedland SJ, Humphreys EB, Mangold LA, Eisenberger M, Dorey FJ, Walsh PC, et al. Risk of Prostate Cancer–Specific Mortality Following Biochemical Recurrence After Radical Prostatectomy. JAMA. 2005;294:433–9.

4. Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Lilja H, et al. Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. J Clin Oncol. 2006;24:3973–8.

5. Bhargava P, Ravizzini G, Chapin BF, Kundra V. Imaging Biochemical Recurrence After Prostatectomy: Where Are We Headed? American Journal of Roentgenology. American Roentgen Ray Society; 2020;214:1248–58.

6. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PLoS One. 2013;8:e66855.

7. Irshad S, Bansal M, Castillo-Martin M, Zheng T, Aytes A, Wenske S, et al. A Molecular Signature Predictive of Indolent Prostate Cancer. Science Translational Medicine. American Association for the Advancement of Science; 2013;5:202ra122-202ra122.

8. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol. 2011;12:245–55.

9. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. Cancer Res. 2014;74:3228–37.

10. Wu C-L, Schroeder BE, Ma X-J, Cutie CJ, Wu S, Salunga R, et al. Development and validation of a 32-gene prognostic index for prostate cancer progression. Proc Natl Acad Sci U S A. 2013;110:6121–6.

11. Luca B-A, Brewer DS, Edwards DR, Edwards S, Whitaker HC, Merson S, et al. DESNT: A Poor Prognosis Category of Human Prostate Cancer. Eur Urol Focus. 2018;4:842–50.

12. Ross-Adams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, Massie CM, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. EBioMedicine. 2015;2:1133–44.

13. Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM, Maddala T, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. Eur Urol. 2014;66:550–60.

14. Penney KL, Sinnott JA, Fall K, Pawitan Y, Hoshida Y, Kraft P, et al. mRNA Expression Signature of Gleason Grade Predicts Lethal Prostate Cancer. JCO. Wolters Kluwer; 2011;29:2391–6.

15. Li R, Wang S, Cui Y, Qu H, Chater JM, Zhang L, et al. Extended application of genomic selection to screen multiomics data for prognostic signatures of prostate cancer. Brief Bioinform. 2021;22:bbaa197.

16. Sinnott JA, Peisch SF, Tyekucheva S, Gerke T, Lis R, Rider JR, et al. Prognostic Utility of a New mRNA Expression Signature of Gleason Score. Clin Cancer Res. 2017;23:81–7.

17. Li R, Jia Z. PCaDB - a comprehensive and interactive database for transcriptomes from prostate cancer population cohorts. bioRxiv. Cold Spring Harbor Laboratory; 2021;2021.06.29.449134.

18. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood. 2017;130:453–9.

19. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci Signal. 2013;6:pl1.

20. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2013;41:D991–5.

21. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Research. 2003;31:68–71.

22. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163:1011–25.

23. Sinha A, Huang V, Livingstone J, Wang J, Fox NS, Kurganovs N, et al. The Proteogenomic Landscape of Curable Prostate Cancer. Cancer Cell. 2019;35:414-427.e6.

24. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010;18:11–22.

25. Gerhauser C, Favero F, Risch T, Simon R, Feuerbach L, Assenov Y, et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. Cancer Cell. 2018;34:996-1011.e8.

26. Kamoun A, Cancel-Tassin G, Fromont G, Elarouci N, Armenoult L, Ayadi M, et al. Comprehensive molecular classification of localized prostate adenocarcinoma reveals a tumour subtype predictive of non-aggressive disease. Annals of Oncology. Elsevier; 2018;29:1814–21.

27. Jain S, Lyons CA, Walker SM, McQuaid S, Hynes SO, Mitchell DM, et al. Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy. Annals of Oncology. 2018;29:215–22.

28. Li R, Qu H, Wang S, Wei J, Zhang L, Ma R, et al. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinformatics. 2018;34:2515–7.

29. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26:2363–7.

30. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

31. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

33. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846–7.

34. Luca B-A, Moulton V, Ellis C, Connell SP, Brewer DS, Cooper CS. Convergence of Prognostic Gene Signatures Suggests Underlying Mechanisms of Human Prostate Cancer Progression. Genes (Basel). 2020;11:E802.

35. Mo F, Lin D, Takhar M, Ramnarine VR, Dong X, Bell RH, et al. Stromal Gene Expression is Predictive for Metastatic Primary Prostate Cancer. Eur Urol. 2018;73:524–32.

36. Yang L, Roberts D, Takhar M, Erho N, Bibby BAS, Thiruthaneeswaran N, et al. Development and Validation of a 28-gene Hypoxia-related Prognostic Signature for Localized Prostate Cancer. EBioMedicine. 2018;31:182–9.

37. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Research. 2020;48:D682–8.

38. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Research. 2021;49:D939–46.

39. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research. 2005;33:D54–8.

40. Agell L, Hernández S, Nonell L, Lorenzo M, Puigdecanet E, de Muga S, et al. A 12-gene expression signature is associated with aggressive histological in prostate cancer: SEC14L1 and TCEB1 genes are potential markers of progression. Am J Pathol. 2012;181:1585–94.

41. Bibikova M, Chudin E, Arsanjani A, Zhou L, Garcia EW, Modder J, et al. Expression signatures that correlated with Gleason score and relapse in prostate cancer. Genomics. 2007;89:666–72.

42. Bismar TA, Demichelis F, Riva A, Kim R, Varambally S, He L, et al. Defining Aggressive Prostate Cancer Using a 12-Gene Model. Neoplasia. 2006;8:59–68.

43. Ding Z, Wu C-J, Chu GC, Xiao Y, Ho D, Zhang J, et al. SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. Nature. 2011;470:269–73.

44. Glinsky GV, Berezovska O, Glinskii AB. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of

cancer. J Clin Invest. American Society for Clinical Investigation; 2005;115:1503–21.

45. Jia Z, Rahmatpanah FB, Chen X, Lernhardt W, Wang Y, Xia X-Q, et al. Expression Changes in the Stroma of Prostate Cancer Predict Subsequent Relapse. PLOS ONE. Public Library of Science; 2012;7:e41371.

46. Nakagawa T, Kollmeyer TM, Morlan BW, Anderson SK, Bergstralh EJ, Davis BJ, et al. A Tissue Biomarker Panel Predicting Systemic Progression after PSA Recurrence Post-Definitive Prostate Cancer Therapy. PLOS ONE. Public Library of Science; 2008;3:e2318.

47. Olmos D, Brewer D, Clark J, Danila DC, Parker C, Attard G, et al. Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. Lancet Oncol. 2012;13:1114–24.

48. Planche A, Bacac M, Provero P, Fusco C, Delorenzi M, Stehle J-C, et al. Identification of Prognostic Molecular Features in the Reactive Stroma of Human Breast and Prostate Cancer. PLOS ONE. Public Library of Science; 2011;6:e18640.

49. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet. 2003;33:49–54.

50. Ramos-Montoya A, Lamb AD, Russell R, Carroll T, Jurmeister S, Galeano-Dalmau N, et al. HES6 drives a critical AR transcriptional programme to induce castration-resistant prostate cancer through activation of an E2F1-mediated cell cycle network. EMBO Molecular Medicine. John Wiley & Sons, Ltd; 2014;6:651–61.

51. Ross RW, Galsky MD, Scher HI, Magidson J, Wassmann K, Lee G-SM, et al. A whole-blood RNA transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study. Lancet Oncol. 2012;13:1105–13.

52. Sharma NL, Massie CE, Ramos-Montoya A, Zecchini V, Scott HE, Lamb AD, et al. The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. Cancer Cell. 2013;23:35–47.

53. Talantov D, Jatkoe TA, Böhm M, Zhang Y, Ferguson AM, Stricker PD, et al. Gene Based Prediction of Clinically Localized Prostate Cancer Progression After Radical Prostatectomy. The Journal of Urology. 2010;184:1521–8.

54. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell. 2005;8:393–406.

55. Yu J, Yu J, Rhodes DR, Tomlins SA, Cao X, Chen G, et al. A Polycomb Repression Signature in Metastatic Prostate Cancer Predicts Cancer Outcome. Cancer Res. 2007;67:10657–63.

56. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28:1–26.

57. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw. 2011;39:1–13.

58. Bair E, Tibshirani R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLoS Biol. 2004;2:e108.

59. Bastien P, Bertrand F, Meyer N, Maumy-Bertrand M. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. Bioinformatics. 2015;31:397–404.

60. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics. Institute of Mathematical Statistics; 2008;2:841–60.

61. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011;27:3206–8.

62. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56:337–44.

63. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43:e47–e47.

64. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;100141.

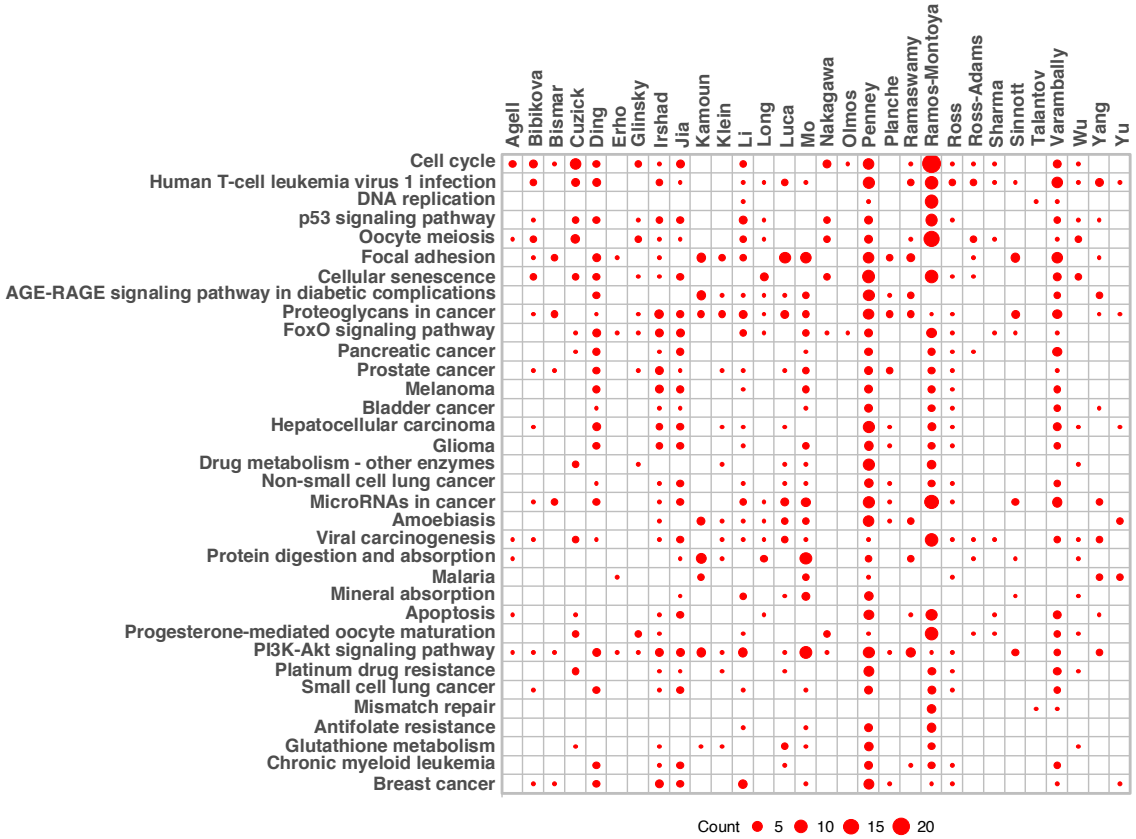

Figure S1. The number of genes in the signatures that are invovled in the KEGG pathways

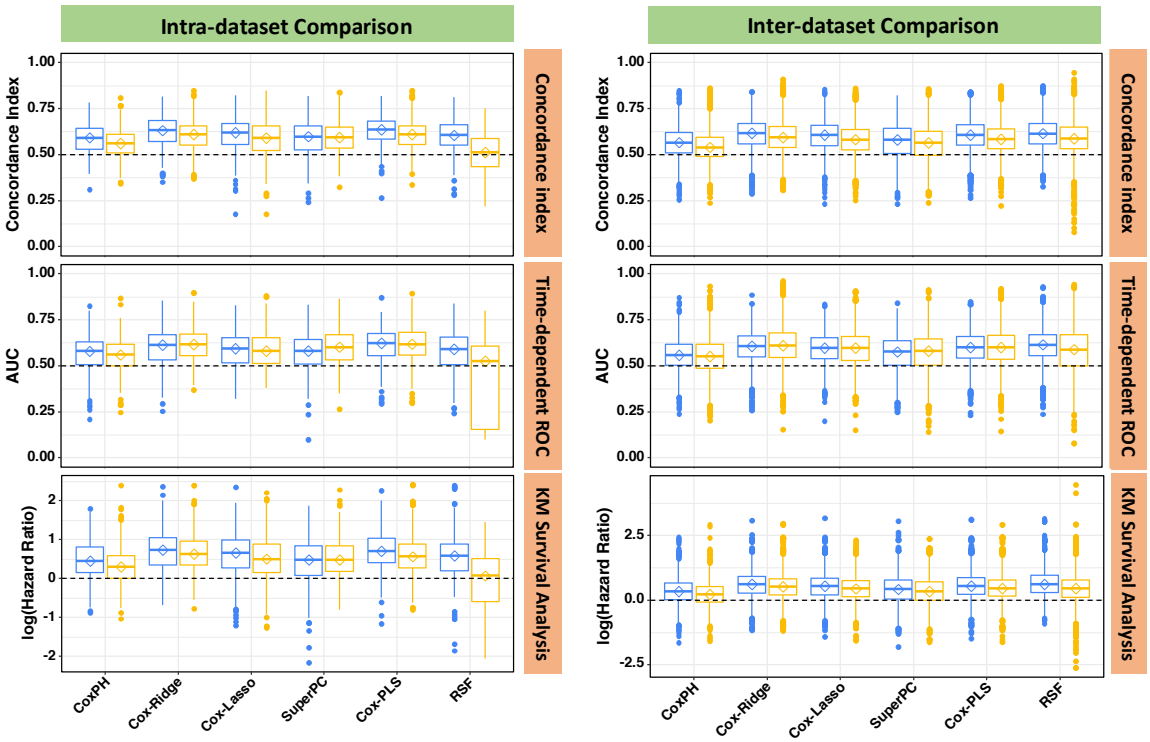

Figure S2. Comparison of the survival analysis models based on different sample sizes. (blue: all the samples were used; yellow: censored data were removed)

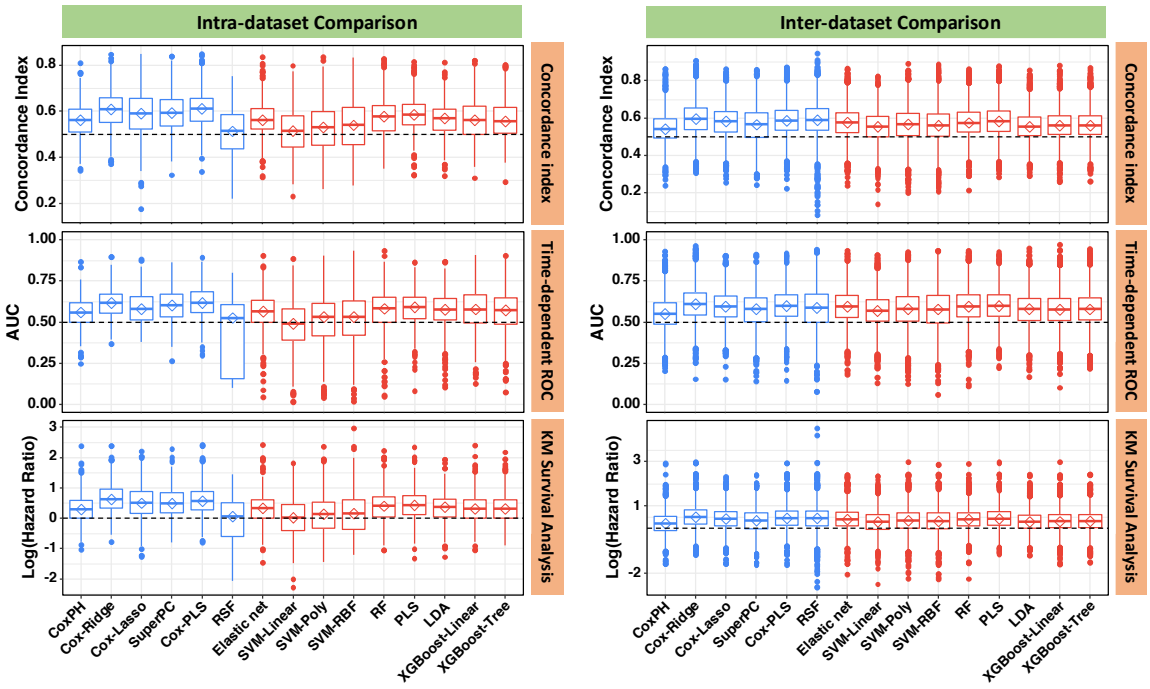

Figure S3. Comparison of the survival analysis models versus the binary classification models using the same data for which the patients with censored data were removed. (blue: survival models; red: classification models)

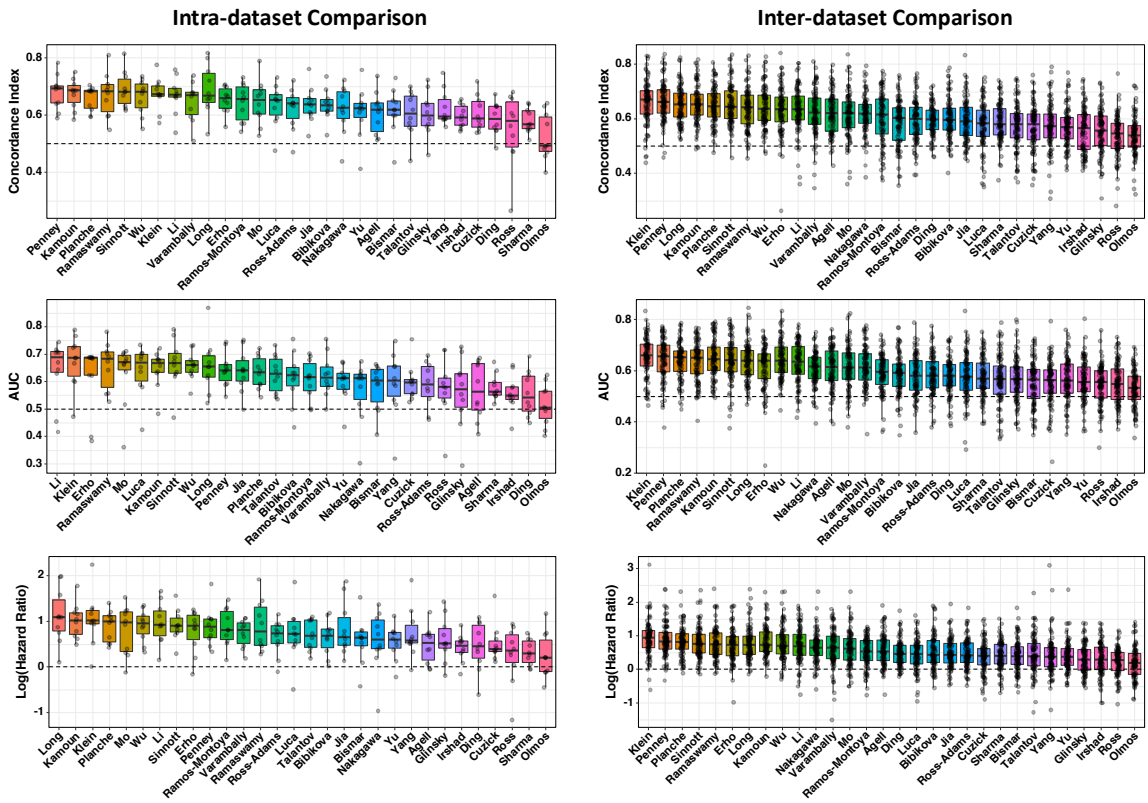

Figure S4. Performance evaluation of the prognostic signatures based on the Cox-PLS method and the three metrics (C-index, AUC, and HR) in the intra-dataset and inter-dataset comparisons

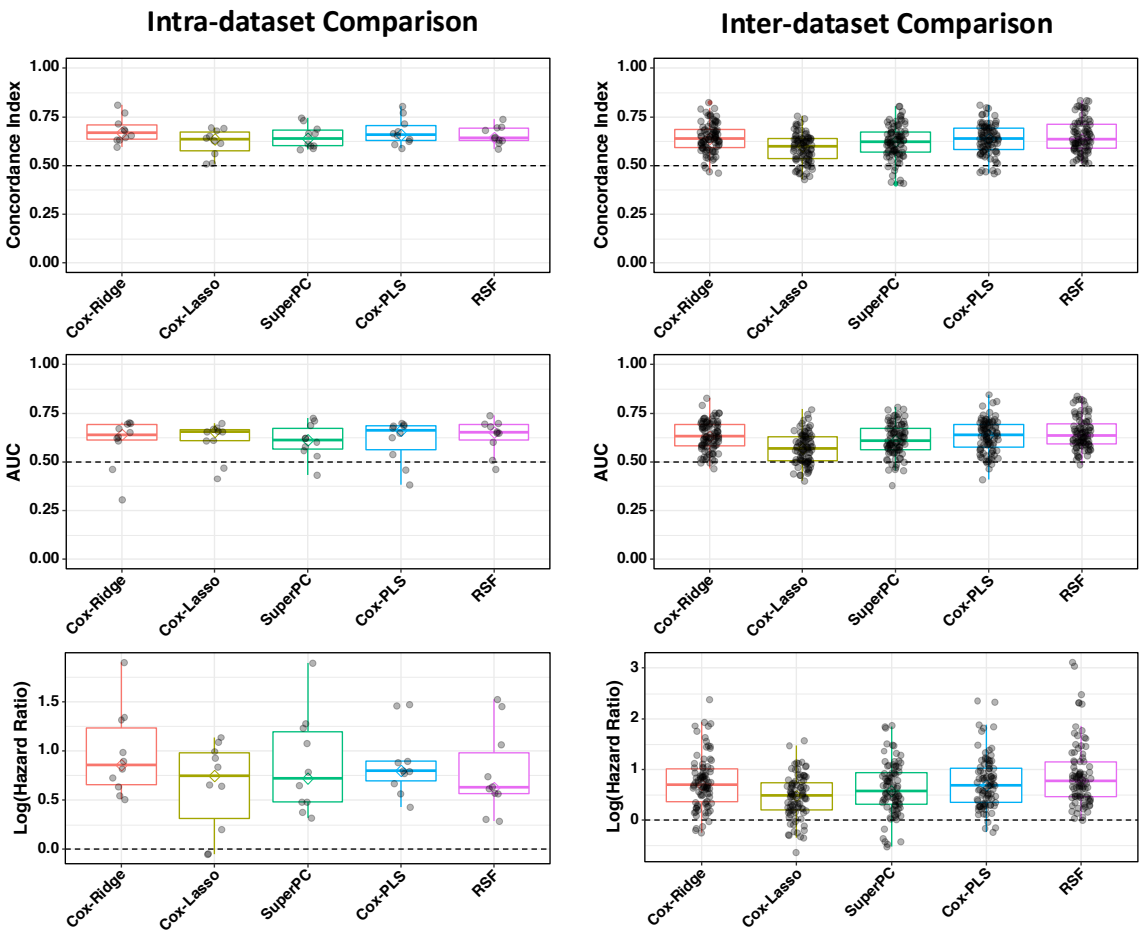

Figure S5. Comparison of the survival analysis algorithms based on the whole transcriptome data

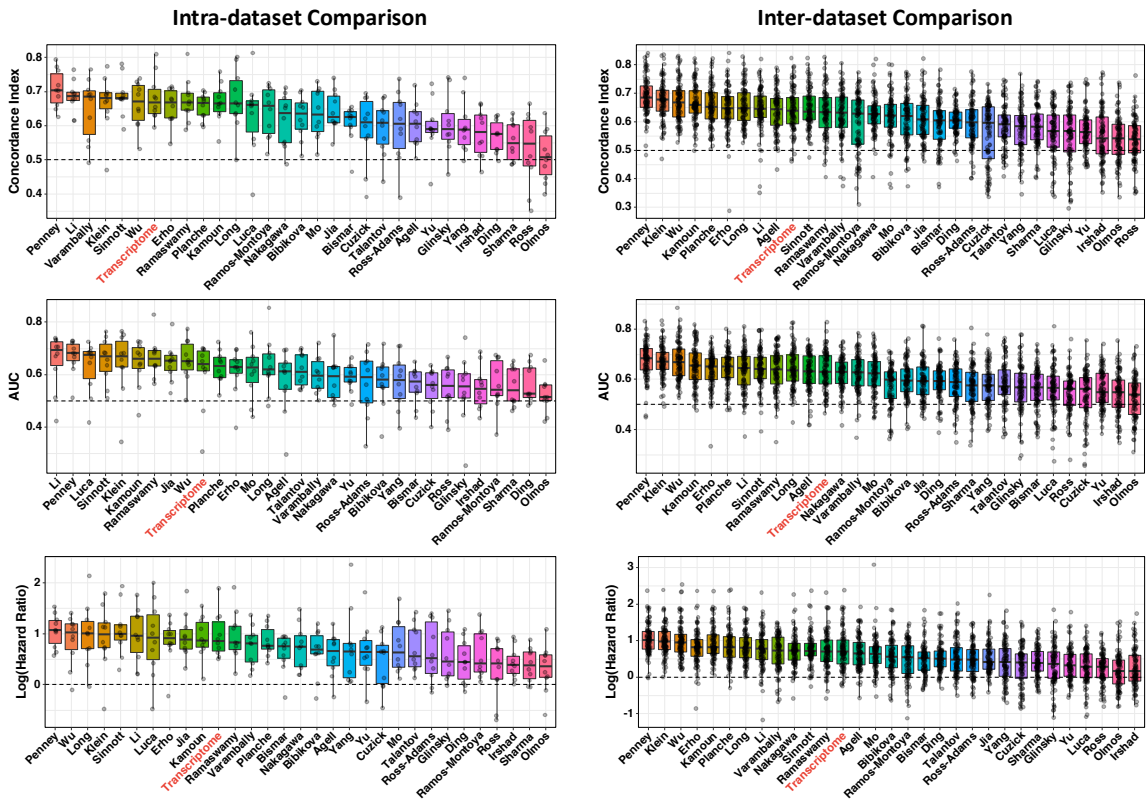

Figure S6. Comparison of the whole transcriptome-based prognostic models with gene expression signature-based prognostic models

Test Set

Training Set	TCGA-PRAD	CPC-Gene	Taylor	DKFZ	GSE54460	Cambridge	Stockholm	CancerMap	CIT	Belfast
Concordance Index										
TCGA-PRAD	3	15	17	19	10	4	5	15	5	6
CPC-Gene	7	15	4	10	10	11	10	1	21	7
Taylor	10	3	16	7	19	12	4	5	12	6
DKFZ	2	20	19	5	18	2	3	14	8	8
GSE54460	8	15	23	8	6	13	15	24	11	5
Cambridge	4	16	19	5	8	20	6	7	22	11
Stockholm	2	6	6	3	8	7	11	21	7	6
CancerMap	20	16	20	18	22	6	25	7	14	25
CIT	9	22	12	13	19	15	21	19	2	12
Belfast	15	9	12	9	24	19	17	23	10	5
Time-dependent ROC										
TCGA-PRAD	2	13	18	12	9	12	5	26	5	9
CPC-Gene	7	9	10	6	10	13	12	12	11	5
Taylor	25	7	15	1	16	23	7	5	16	3
DKFZ	2	13	24	11	20	14	3	16	1	10
GSE54460	12	11	26	8	10	7	20	19	3	8
Cambridge	6	16	22	11	2	31	5	7	21	8
Stockholm	2	6	4	1	7	12	8	20	5	12
CancerMap	17	23	25	14	19	12	19	12	13	22
CIT	3	22	7	18	13	9	22	20	20	8
Belfast	11	4	8	16	24	10	16	30	26	11
KM Survival Analysis										
TCGA-PRAD	5	15	30	8	7	1	7	20	3	4
CPC-Gene	9	11	12	4	4	16	9	9	20	10
Taylor	7	4	19	3	17	19	10	5	18	10
DKFZ	2	16	18	4	10	9	2	9	23	6
GSE54460	12	16	17	20	9	17	20	18	7	6
Cambridge	6	17	27	4	7	20	8	2	26	15
Stockholm	2	13	5	1	6	4	17	15	9	8
CancerMap	24	20	24	23	14	3	25	14	21	30
CIT	9	23	20	12	13	12	16	24	7	15
Belfast	19	8	10	12	25	18	9	28	6	8

Figure S7. The rankings of the whole transcriptome-based models comparing with the gene expression signature-based models. (the rows are the training datasets and the columns are the test datasets; the rankings on the diagonals are from the intra-dataset comparison based on 10-fold CV, others are from the inter-dataset comparison)