Deep Survival Models for High-Dimension Data

April 23, 2018

- Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models (SurvivalNet)
- A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme

SurvivalNet

Problems

- Traditional Cox proportional hazards models require enormous cohorts for training models on high-dimensional datasets containing large numbers of features.
- A small set of features is selected in a subjective process that is prone to bias and limited by imperfect understanding of disease biology.

SurvivalNet

Possible Model

- Cox proportional hazards models with elastic net regularization.
- Random survival forests.
- Neural network based approaches(SurvivalNet)

SurvivalNet

Cox proportional hazards models with elastic net regularization

$$h(t) = h_0(t) \exp\{\beta^T X\},$$

$$L(\beta) = \Pi_{i=1}^{n} \frac{\exp\left(\sum_{k=1}^{p} \beta_k x_{ik}\right)}{\sum_{j \in R_i} \exp\left(\sum_{k=1}^{p} \beta_k x_{jk}\right)}$$

$$l(\beta) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{p} \beta_k x_{ik} - \ln \left[\sum_{j \in R_i} \exp(\sum_{k=1}^{p} \beta_k x_{jk}) \right] \right\}$$

$$\hat{\beta}_{(EN)} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ -\beta^T X_i + \ln \left[\sum_{j \in R_i} \exp\left(\beta^T X_j\right) \right] \right\} + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \right\},$$

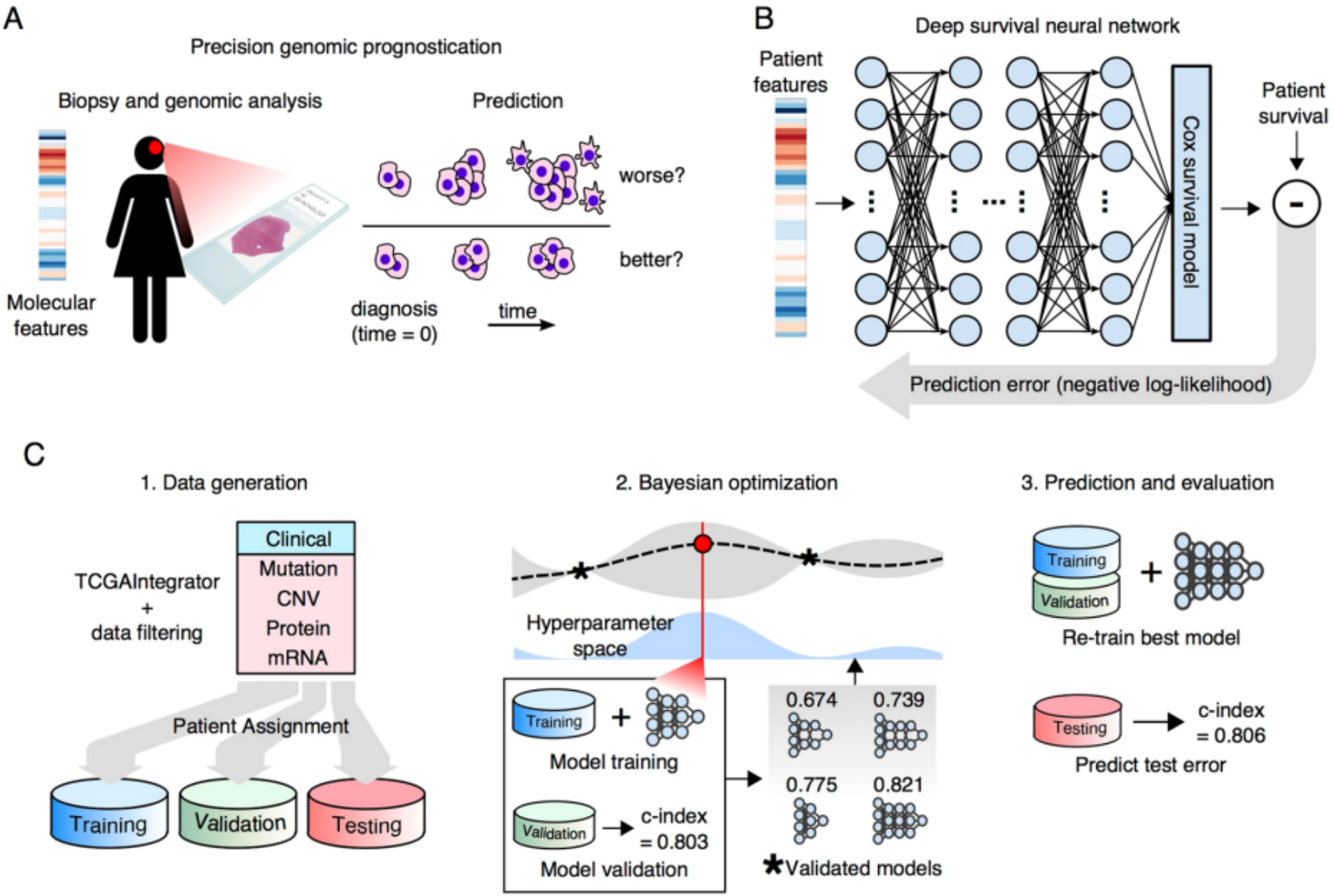
SurvivalNet

Random survival forests

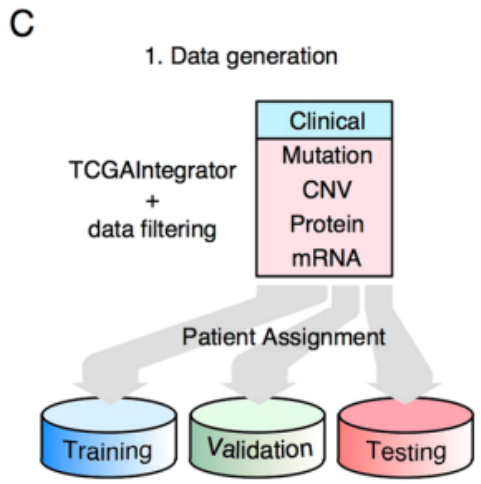
1. Draw B bootstrap samples from the original data. Note that each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data).
2. Grow a survival tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split using the candidate variable that maximizes survival difference between daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique deaths.
4. Calculate a CHF for each tree. Average to obtain the ensemble CHF.

SurvivalNet

Overview

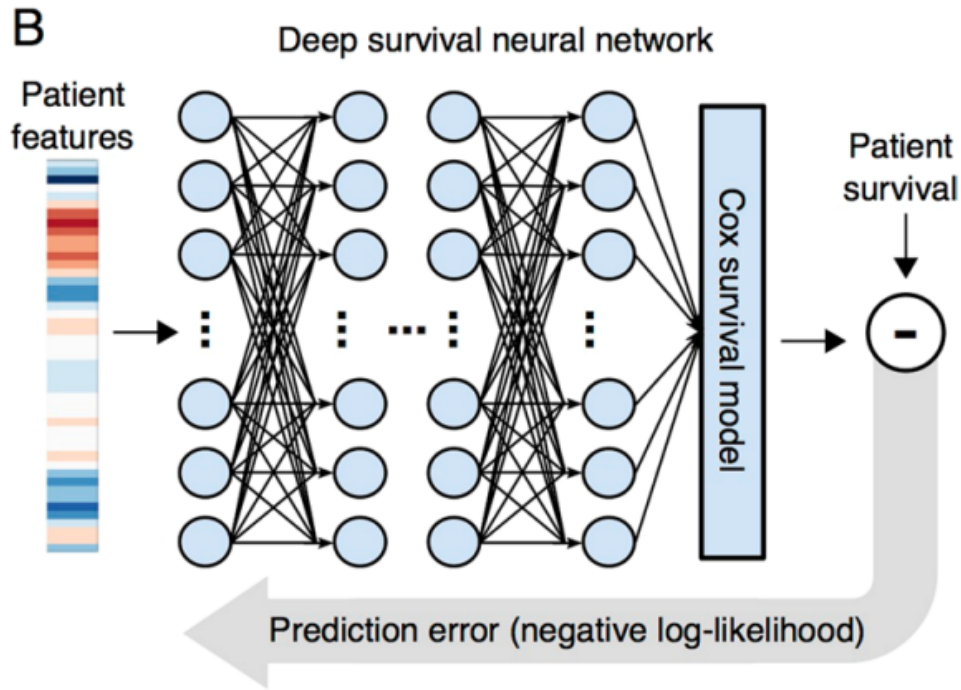


Data Splitting



- All datasets were created and filtered using TCGAIntegrator.
- Data is first split into training (60%), validation (20%), and testing (20%) sets
- Performance was evaluated with two feature-sets: 1) a "transcription" feature set containing 17,000 + gene expression features obtained by RNA-sequencing, and 2) an "integrated" feature set containing 3-400 features describing clinical features, mutations, gene and chromosome arm-level copy number variations, and protein expression features.

Deep survival neural network

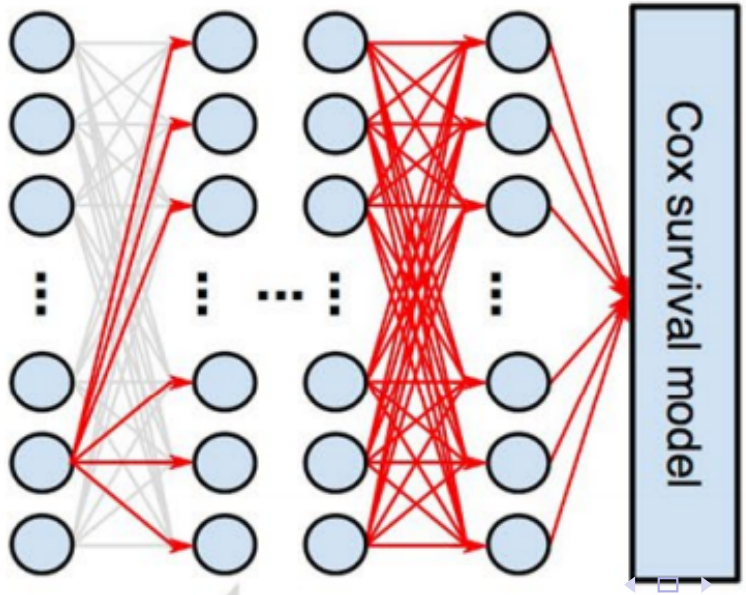


- Fully connected hidden layer is used.
- input is the feature, output is log-likelihood of Cox proportional hazards model.

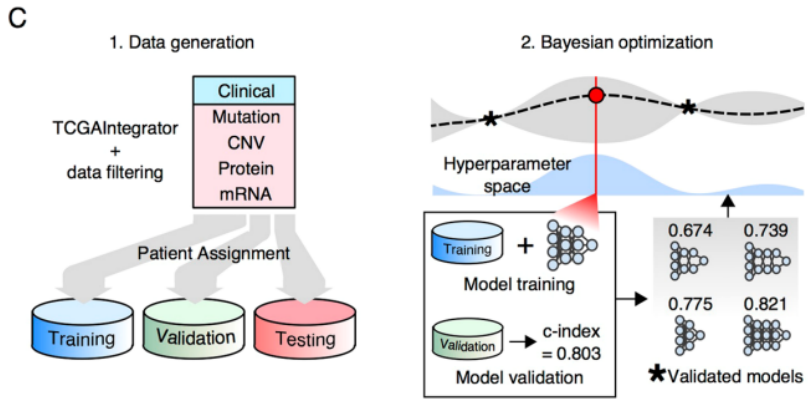
Deep survival neural network

$$l(\beta,\ X) = -\sum_{i \in U}\left(X_i\beta - log\sum_{j \in R_i} e^{X_j\beta}\right)$$

- X_i are the inputs to output layer
- β are the Cox model parameters
- U is the set of events
- R_i is the set of "at risk" samples.



Bayesian optimization for hyper-parameter tuning

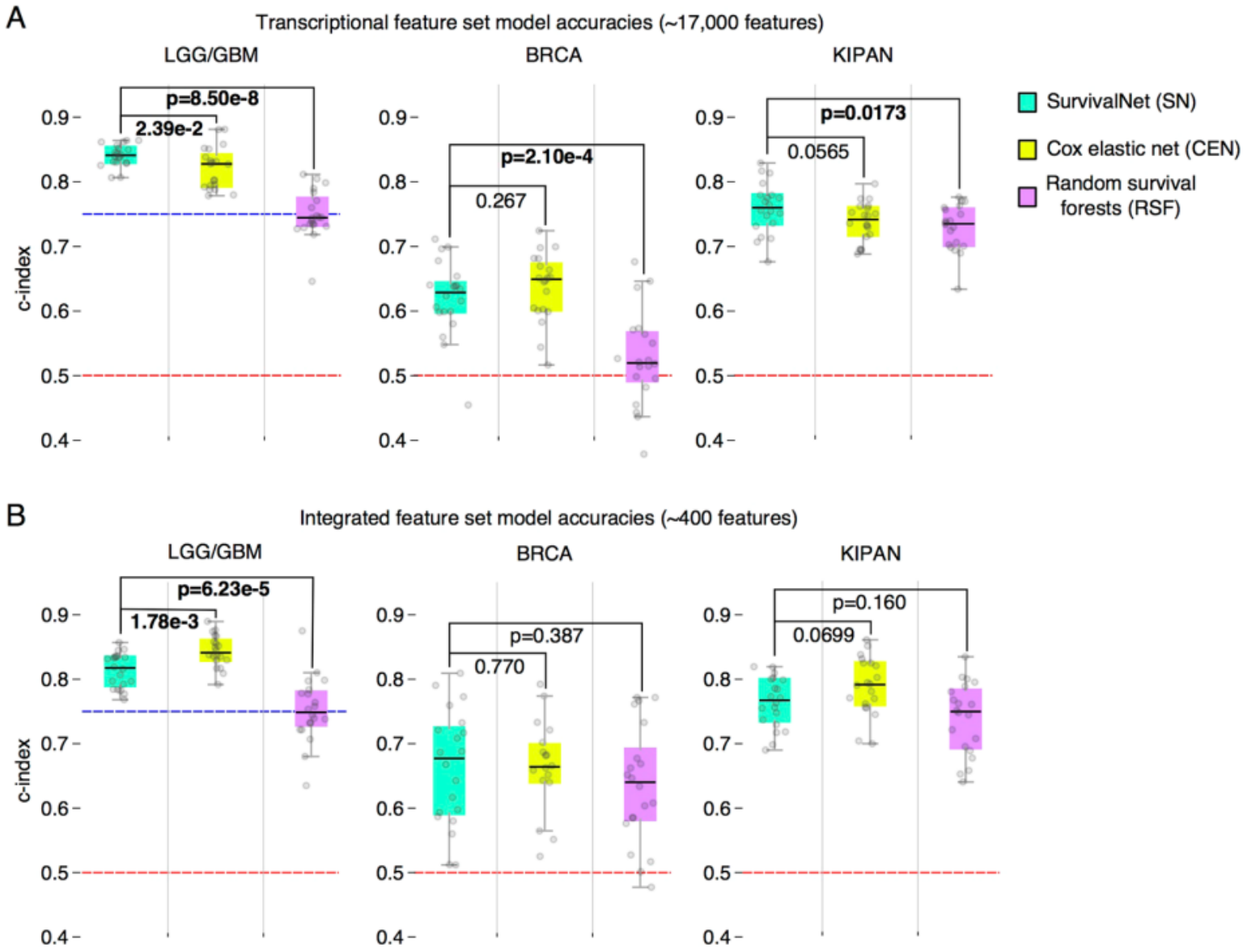


- Hyper-parameters of the deep learning structure includes number of layers, number of elements in each layer, number and type of activation functions in each layer, choices for optimization/regularization procedures, dropout fraction...
- Bayesian optimization enables users who lack experience tuning neural networks to optimize model designs automatically, and results in considerable savings in time and effort as previously reported.

Bayesian optimization for hyper-parameter tuning

- Training samples are used to train the model weights with backpropagation using the network design suggested by Bayesian optimization.
- The prediction accuracy of the trained deep survival model is then estimated using the validation samples, and is used to maintain a probabilistic model of performance as a function of hyperparameters.
- Based on the probabilistic model, the design with the best expected accuracy is inferred as the next design to test.

Model comparison results



- Both SN and CEN significantly outperform RSF models in most experiments.
- Performance is generally better on the integrated feature set than the transcriptional feature set for all methods.
- RSF models have the worst performance generally.
- Comparing performance across diseases, we noticed that prediction accuracy generally decreases as the proportion of right-censored samples in a dataset increases. This pattern holds for all prediction methods. Glioma had the highest overall prediction accuracy, being a uniformly fatal disease that has relatively fewer long-term survivors and incomplete follow-up (62-64%). Breast carcinoma had the lowest overall prediction accuracy with more than 86-91% of BRCA samples being right-censored.

Interpreting deep survival models with risk backpropagation

Linear survival model

- Linear survival models weight individual features based on their contribution to overall risk, providing a clear interpretation of the prognostic significance of individual features, and insights into the biology of disease progression.

$$h(t) = h_0(t) \exp\{\beta^T X\}.$$

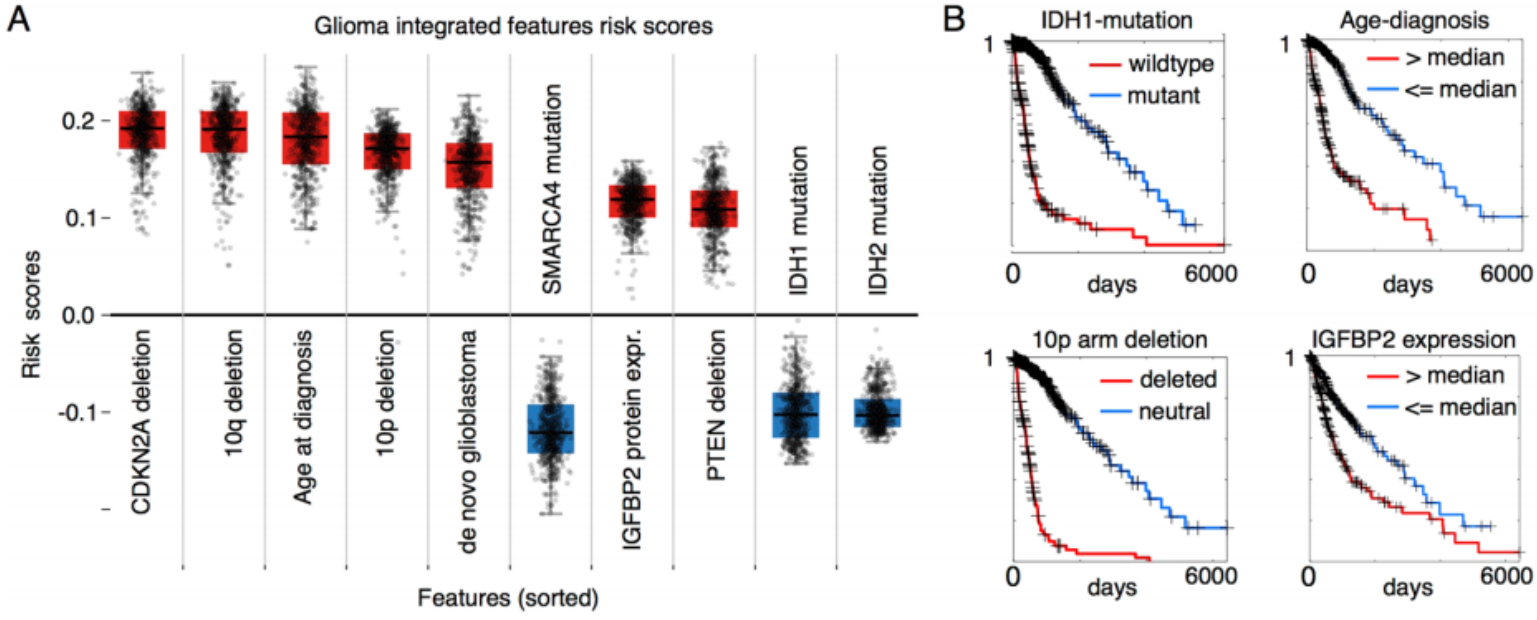
Interpreting deep survival models with risk backpropagation

Risk backpropagation

- The prediction can be conceptualized instead as a nonlinear surface where the risk gradients change depending on a patient's feature values, and so these feature weights are calculated separately for each patient.
- The models used for risk backpropagation and interpretation were created by identifying the best performing model configuration from the randomized experiments.

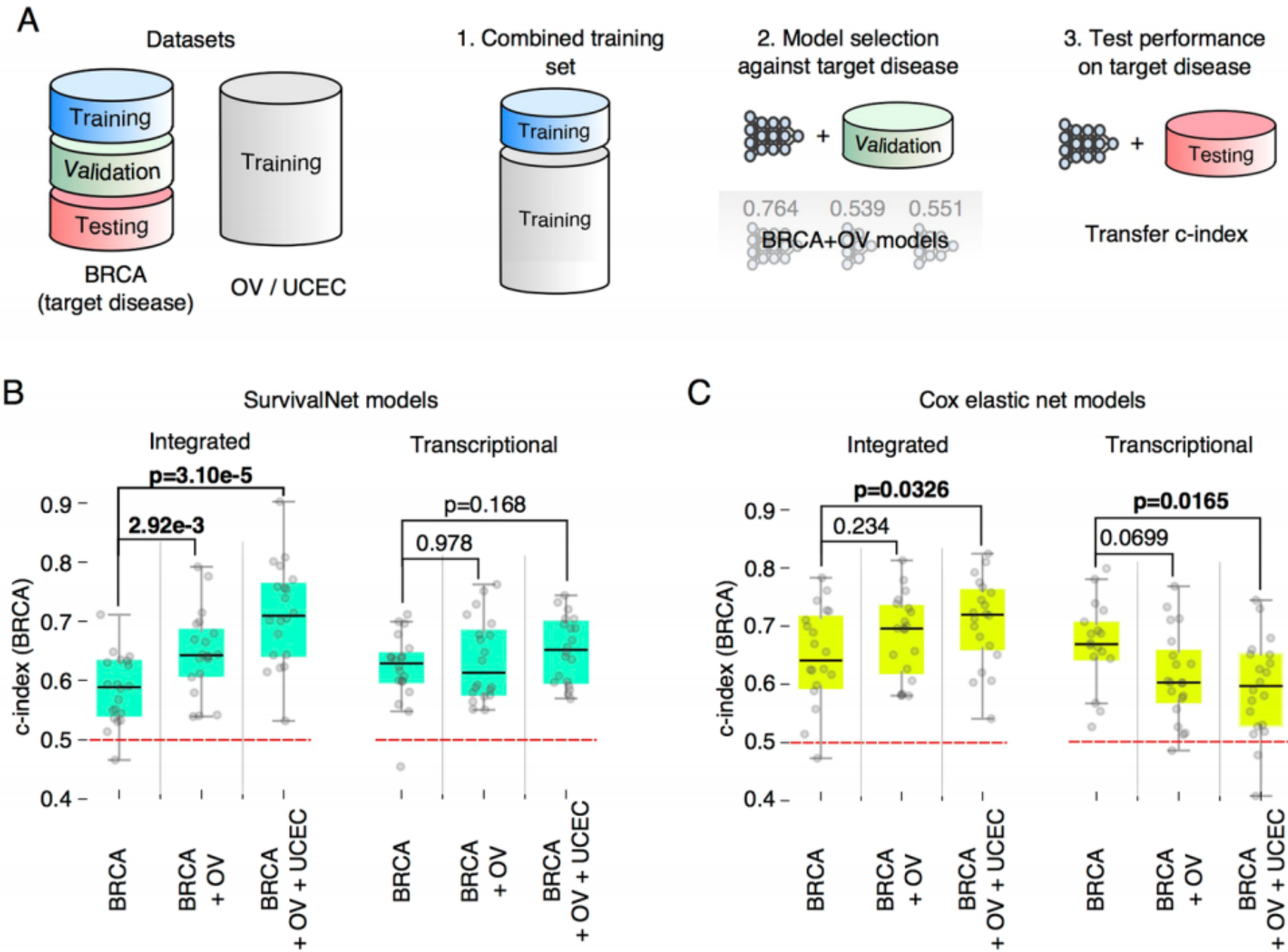
$$\frac{\partial R}{\partial f} = \beta \times \prod_{h=1}^{H} J_h$$

- H number of hidden layers, R risk, J_h Jacobian matrix of the h-th hidden layer with respect to its inputs, β is the vector of parameters of the final layer that is a linear transformation


A
Glioma integrated features risk scores
Risk scores
0.2
0.1
0.0
-0.1
CDKN2A deletion
10q deletion
Age at diagnosis
10p deletion
de novo glioblastoma
SMARCA4 mutation
IGFBP2 protein expr.
PTEN deletion
IDH1 mutation
IDH2 mutation
Features (sorted)
B
IDH1-mutation
wildtype
mutant
Age-diagnosis
> median
<= median
10p arm deletion
deleted
neutral
IGFBP2 expression
> median
<= median
1
0
days
6000

Transfer learning with multi-cancer datasets

- Perform a series of transfer learning experiments to evaluate the ability of deep survival models to benefit from training with data from multiple cancer types
- Survival models were trained using three different datasets: BRCA-only, BRCA+OV (ovarian serous carcinoma), and BRCA+OV+UCEC (corpus endometrial carcinoma), and were evaluated for their accuracy in predicting BRCA outcomes
- The large proportion of right-censored cases in the BRCA dataset (90%) makes training accurate models difficult, and so we hypothesized that augmenting BRCA training data with samples from other hormone-driven cancers could improve BRCA prognostication
- Datasets were combined using their shared features.

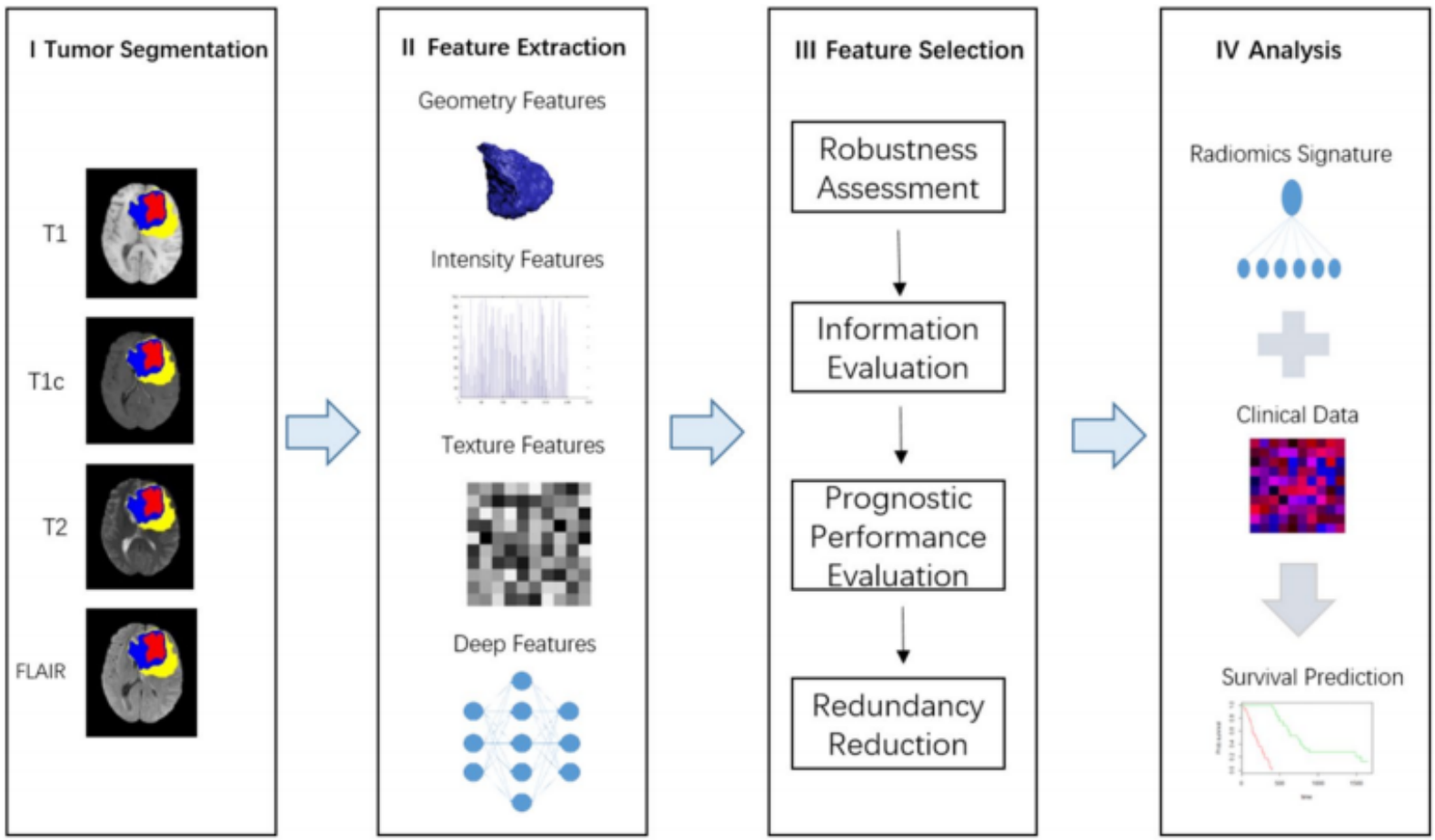

A
Datasets
Training
Validation
Testing
BRCA
(target disease)
Training
OV / UCEC
1. Combined training set
Training
Training
2. Model selection against target disease
Validation
0.764
0.539
0.551
BRCA+OV models
3. Test performance on target disease
Testing
Transfer c-index
B
SurvivalNet models
Integrated
Transcriptional
p=3.10e-5
2.92e-3
p=0.168
0.978
c-index (BRCA)
0.9
0.8
0.7
0.6
0.5
0.4
BRCA
BRCA + OV
BRCA + OV + UCEC
C
Cox elastic net models
Integrated
Transcriptional
p=0.0326
0.234
p=0.0165
0.0699

Discussion

We created a software framework for Bayesian optimization and interpretation of deep survival models, and evaluated the ability of optimized models to learn from high-dimensional and multi-cancer datasets. Our software enables investigators to efficiently construct deep survival models for their own applications without the need for expensive manual tuning of design hyperparameters, a process that is time consuming and that requires considerable technical expertise. We also provide methods for model interpretation, using the backpropagation of risk to assess the prognostic significance of features and to gain insights into disease biology. Our analysis shows the ability of deep learning to extract important prognostic features from high-dimensional genomic data, and to effectively leverage multi-cancer datasets to improve prognostication. It also reveals limitations in deep learning for survival analysis and the value of complex and deeply layered survival models that need to be further investigated.

A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme

Overview



Datasets

- A total of 112 patients (62 men and 50 women; mean age, 54.640 years $\pm$44.040; range 10-84 years) with pathologically confirmed GBM were included.
- The patient cohorts consisted of two groups: a discovery cohort comprising 75 patients from the Cancer Genome Archive (TCGA) database, and an independent validation cohort comprising 37 patients
- The inclusion criteria were that patients with newly diagnosed and treatment-naive GBM and survival information and pre-treatment MR imaging including T1-weighted, T1-weighted Gadolinium contrast-enhanced, T2-weighted, and T2-weighted FLAIR images (short for T1, T1C, T2, and FLAIR). The exclusion criteria are patients with a history of surgery or chemoradiation therapy and patients missing survival information.

Datasets

Characteristic	Discovery Data Set	Validation Data Set
No. of patients*	75 (67%)	37 (33%)
Sex^{+}($P = 0.553$)		
Male*	43 (57%)	32 (43%)
Female^{+}	19 (51%)	18 (49%)
Age^{+}($P = 0.909$)		
Ranges	19–84	10–78
Median†	57 (52–59)	55 (49–62)
Mean†	54.990 (51.710–58.260)	53.950 (48.240–59.650)
OS^{+}($P = 0.978$)		
Ranges	30–1642	77–1870
Median†	441 (381–530)	377 (332–584)
Mean†	495.160 (412.520–577.800)	494.220 (364.250–624.180)

Image Preprocessing and Tumor Segmentation

- Images were re-sampled to 1mm×1mm×1mm
- Three tumor subregions were segmented, including the necrosis area, the enhancement area and the edema area. The necrosis area was the low intensity necrotic structures within the enhancing rim in T1C and had hyper-intense signal in T2 and FLAIR. The enhancement area was confirmed as the Gadolinium enhancing rim excluding the necrotic center and hemorrhage with both T1C and T1 images. The edema area was identified as abnormality visible in T2 and FLAIR excluding ventricles and cerebrospinal fuid. The edema area may include both peritumoral edema and any non-enhancing tumor.

Feature Extraction

Handcrafed Features

- The handcrafed features were extracted from five subregions and four MR modalities. The feature extraction subregions include necrosis, enhancement, edema, tumor core (the whole tumor except edema) and whole tumor (necrosis, enhancement and edema).
- The handcrafed features can be divided into three groups: (I) geometry, (II) intensity and (III) texture.
- A total of 1403 handcrafed features are extracted, including 23 geometry features, 340 intensity features, and 1040 texture features.

Feature Extraction

Deep Features

- Deep features were extracted from pre-trained CNN via transfer learning. In this study, CNN_S was chosen as the pre-trained CNN model.
- For each patient, the necrosis, tumor core and whole tumor subregions were chosen as input of CNN_S.

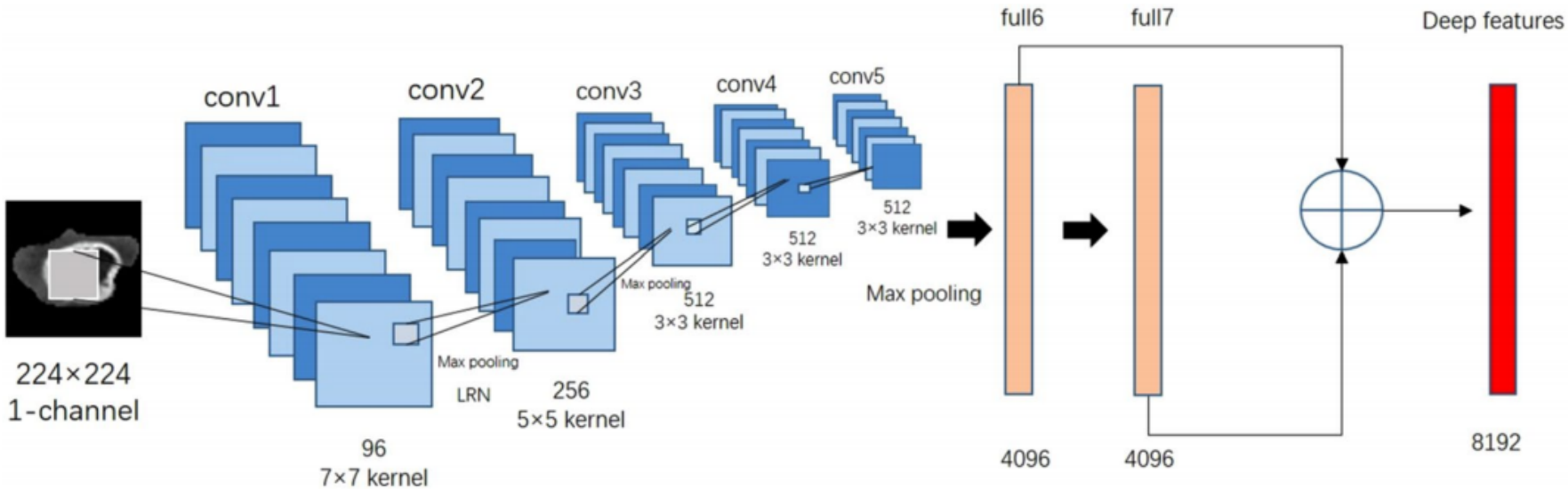
Feature Extraction

Deep Features

First, from multiple transverse slices in the segmentation volume, we picked out the slice which had the largest tumor area. Then, the gray values were normalized to range $[0, 255]$ using linear transformation. Based on the segmentation results, the three tumor subregions were cropped from the selected slices in all four MR modalities. Next, each cropped subregion image was resized to 224×224 with bicubic interpolation. The obtained images can be used as the model input. The G and B channels of CNN_S were turned of so only grayscale images were allowed to enter the model. Finally, the deep features can be computed by only forward propagation and were extracted from fully-connected layer 6 and fully-connected layer 7. In total, 98304 deep features can be extracted for each patients.(all weights of CNN_S were predetermined)

Feature Extraction

Deep Features



Feature Selection

- After features extraction, all 1403 handcrafed features and 98304 deep features for each patients were normalized as z-scores. A four-step method is used for feature selection. All calculations are performed on the discovery data set.
- Test-retest analysis and inter-rater analysis were used to determine the feature robustness. Based on 30 patients randomly chosen from the discovery data set, the test-retest analysis was performed where for each patient the tumor subregions were segmented twice by one rater.
- The data set used for inter-rater analysis included another 30 randomly chosen patients, where for each patient the tumor subregions were segmented by two raters independently. The features extracted from these multiple-segmented subregions were assessed using intraclass correlation coefficient
- Feature with ICC$\leq$ 0.85 were considered as robust against intra- and inter-rater uncertainties.

Feature Selection

- Then, the median absolute deviations (MAD) was calculated for each remained feature. Features with MAD equal to zero were discarded.
- Next, we further selected features with prognostic value. Here the prognostic performance is assessed using the concordance index (C-index). Features with C-index$\geq$0.580 are considered. After prognostic performance analysis, 1581 features remained as predictive factors.
- Then, we further reduced the data dimension by removing highly correlated features. Here the correlation coefficient between each pair of features is calculated. For feature pair with correlated coefficient âĽě0.90, the more prognostic feature is retained and the other feature is removed. Finally, the remained 150 image features are selected and regarded as robust, predictive and non-redundant.

Statistical Analysis

Signature Construction

- Based on the selected 150 features, we aimed to construct a radiomics signature using multivariate Cox regression model for prediction of survival in GBM patients. Because there were more image features than patients, strong feature selection and shrinkage were still required to prevent overprinting as well as increase interpretation. To address this problem, the least absolute shrinkage and selection operator (LASSO) Cox regression model was used on the discovery data set for signature construction.
- The retained features with nonzero coefficients were used for regression model fitting and combined into a radiomics signature. Subsequently, we obtained a radiomics score for each patient by a linear combination of retained features weighed by their model coefficients

Statistical Analysis

Signature Validation

- The association of the constructed signature with survival was assessed on the discovery data set and validated on the validation data set by using Kaplan-Meier survival analysis. Based on a threshold calculated using the radiomics score, all patients were stratified into high-risk and low-risk groups. The threshold was estimated based on the discovery data set by using an optimal cutpoint analysis with X-tile software, and tested on the validation data set. A weighted log-rank test (G-rho rank test, rho=1) was used to test the significant difference between the high-risk and low-risk groups. The C-Index was used to assess the performance of the signature.

Statistical Analysis

Signature Validation

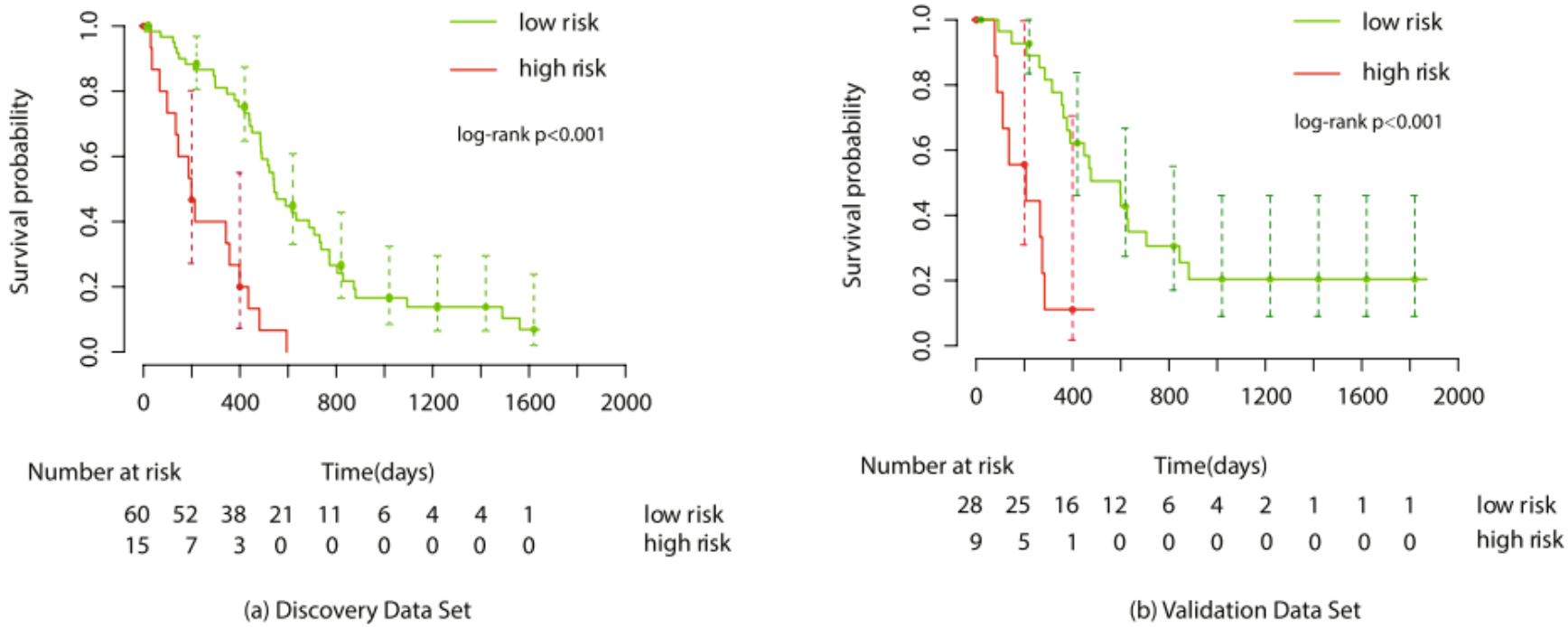


Figure: Illustration of Kaplan-Meier survival curve. The Kaplan-Meier survival curve show OS risk stratification for patients in Discovery data set (a) and Validation data set (b). Patients were classified as low risk and high risk according to radimics signature. The vertical dashed line is 95% confdence interval.

Statistical Analysis

Signature Validation

- To assess the univariate predictive performance of each feature with non-zero LASSO coefficient, the univariate analysis was performed based on both discovery and validation data sets. To assess the univariate association with risk, each non-zero feature was used for patient stratification into high-risk and low-risk groups.
- To compare the built radiomics signature with other clinical risk factors such as age and KPS, the C-indices of these clinical risk factors were calculated based on both discovery and validation data sets. To assess the combination prognostic value of the signature with clinical factors, we put the radiomics signature together with clinical parameters into the Cox regression model. The model was fitted based on the discovery data set and validated on the validation data set.

Results

Clinical Characteristics and OS

The median and mean of OS were 441 days and 495.160 days for the discovery set, 377 days and 494.220 days for the validation set. The median and mean of age were 57 years and 54.990 years for the discovery set, 55 years and 53.950 years for the validation set. The discovery set had 43 males and 32 females, while the validation set had 19 males and 18 females. There was no significant difference in clinical and follow-up data between the discovery and validation data sets (P=0.553 for sex test, 0.748 for KPS test, 0.909 for age test, 0.302 for tumor volume test and 0.978 for OS test).

Results

Signature Construction

$$
\begin{aligned}
Radiomics\_signature\_score \;=\; & FLAIR\_ST\_F7\_870 \times 0.06867720 \\
& + FLAIR\_SN\_F7\_2297 \times (-0.05909112) \\
& + T1C\_SNE\_F6\_806 \times 0.05420293 \\
& + T2\_SNE\_F7\_772 \times (-0.03454031) \\
& + T1C\_SNE\_F7\_1508 \times (-0.02240571) \\
& + FLAIR\_SNE\_F6\_2981 \times (-0.00958802)
\end{aligned}
$$

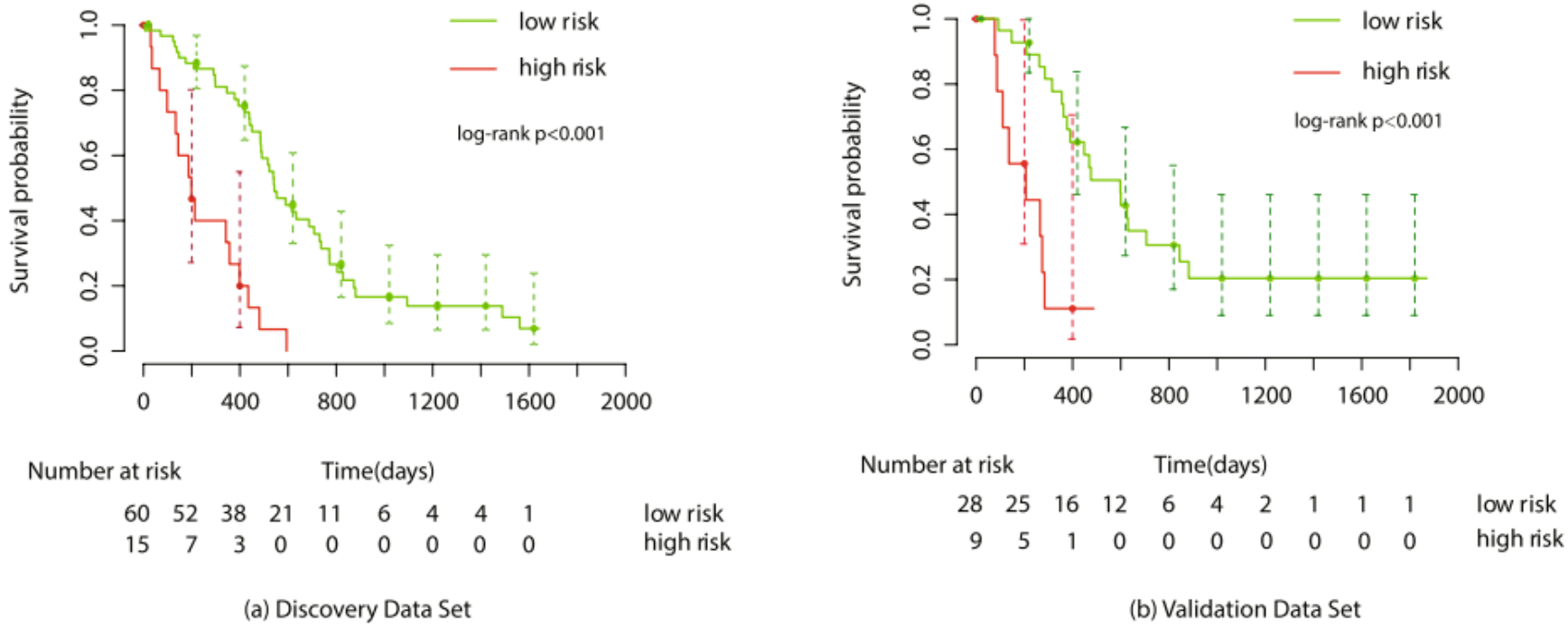


Figure: Illustration of Kaplan-Meier survival curve. The Kaplan-Meier survival curve show OS risk stratification for patients in Discovery data set (a) and Validation data set (b). Patients were classified as low risk and high risk according to radiomics signature. The vertical dashed line is 95% confidence interval.