CAMDA 2017 Challenge of Neuroblastoma Data Integration

Integration of CNV and RNA-seq data can increase the predictive power of Neuroblastoma endpoint

Yimin Ma, Jiajun Chen and Tieliu Shi*

The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China

Neuroblastoma (NB) is the most common extracranial solid tumor in children. To compare the predictive power between data integration and the original expression-only study, we first built two risk-score models based on RNA-seq data and CNV data respectively, we then combined them with two different strategies, last we evaluated the predictive power of these four models.

In this study, there are total 498 RNA-seq data[1] and 145 corresponding CNV data generated with comparative genomic hybridization (aCGH) microarray[2-4], we successfully matched 496 RNA-seq data and 138 CNV data to the clinical information. There are 175 clinical defined high-risk patients among the samples. The endpoint we worked on was overall survival time.

We randomly selected half of the RNA-seq data (including 248 samples) as the training set to build a risk-score model and used the remaining data as the testing set. Firstly, based on log-rank test, we selected 6382 genes which can individually divide NB patients into two groups by the training set, overall survival was significantly different between these two groups, indicating that these genes can be useful features for NB patients' stratification. Secondly, 58 genes with FDR < 0.05 were selected from those candidate genes by correlation test between genes and overall survival times. Using the Cox regression method and stepwise regression method, we built the first risk-score model with five genes selected from the 58 survival correlated genes. NB patients could be classified into a high-risk group or a low-risk group with half and half based on the median risk-score calculated by this model either in the training set or in the testing set[5, 6]. Overall survival between these two groups was significantly different (P = 0.00473 in the matched testing set, Figure 1 A-D). We compared different models based on the P-value of the testing set which is more suitable for comparing the repeatability and predict power for the model.

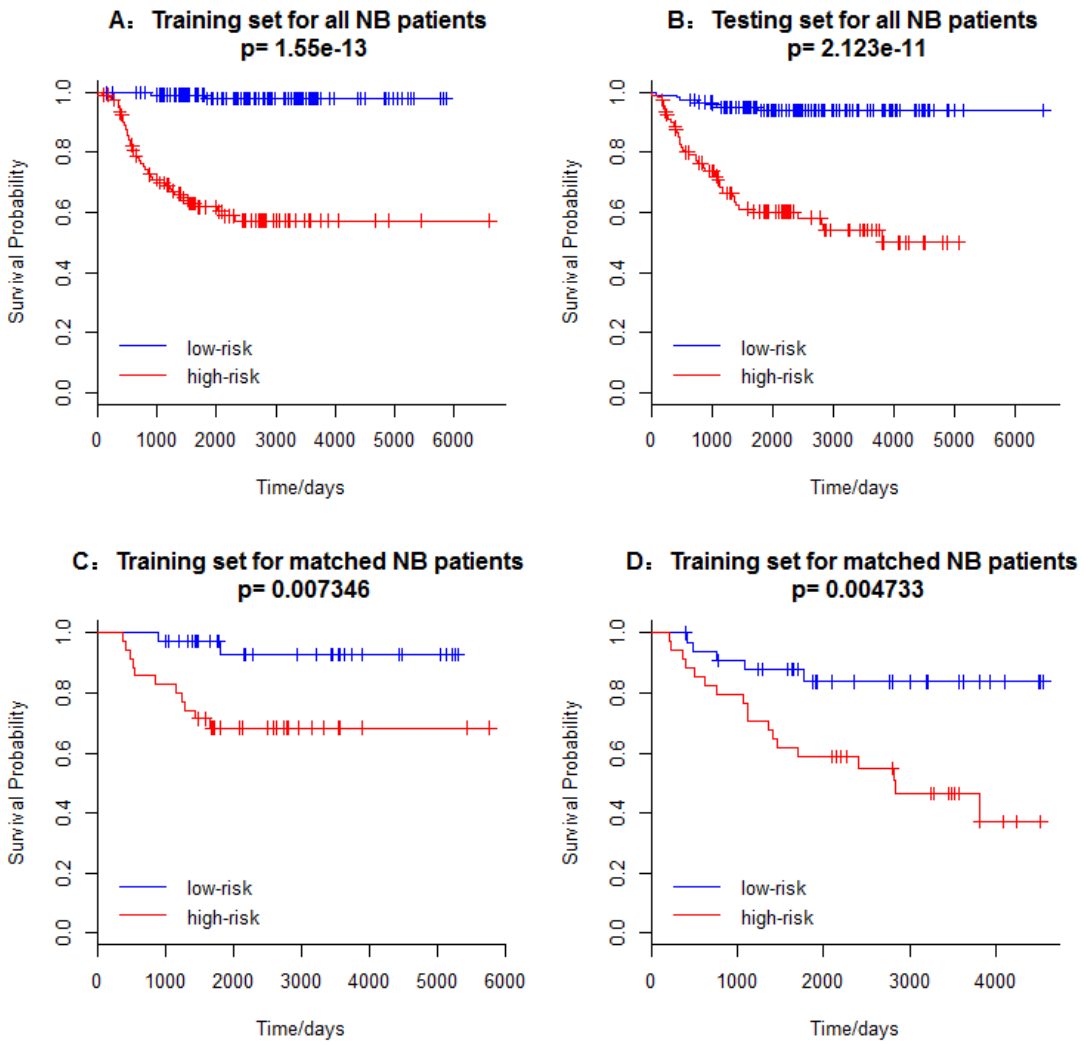

Figure 1 Kaplan-Meier curves and log-rank test of overall survival for patients in the training set and testing set.

A: The RNA-seq based model for the training set with 248 samples (124 high-risk and 124 low-risk) in all NB patients

B: The RNA-seq based model for the testing set with 248 samples (124 high-risk and 124 low-risk) in all NB patients

C: The RNA-seq based model for the training set with 70 (35 high-risk and 35 low-risk) samples having matched CNV data

D: The RNA-seq based model for the testing set with 68 (34 high-risk and 34 low-risk) samples having matched CNV data

Next, by matching the classification result of RNA-seq based model and the clinical defined classification, we found that our risk-score model showed high consistency with the clinical definition of NB patients (> 80%, Table 1, Figure 2 A-B).

Table 1 The consistency of clinical defined high/low risk and the classified results of our RNA-seq based model

	Clinical defined high-risk	Clinical defined low-risk	Sum
RNA-seq model defined high-risk	162	86	248
RNA-seq model defined low-risk	13	235	248
Sum	175	321	496

Then, we applied the 5 gene risk-score model to clinical defined high/low group and the median risk-score based on the model can further subdivide each of the clinical group into two subgroups (Figure 2 C-D).

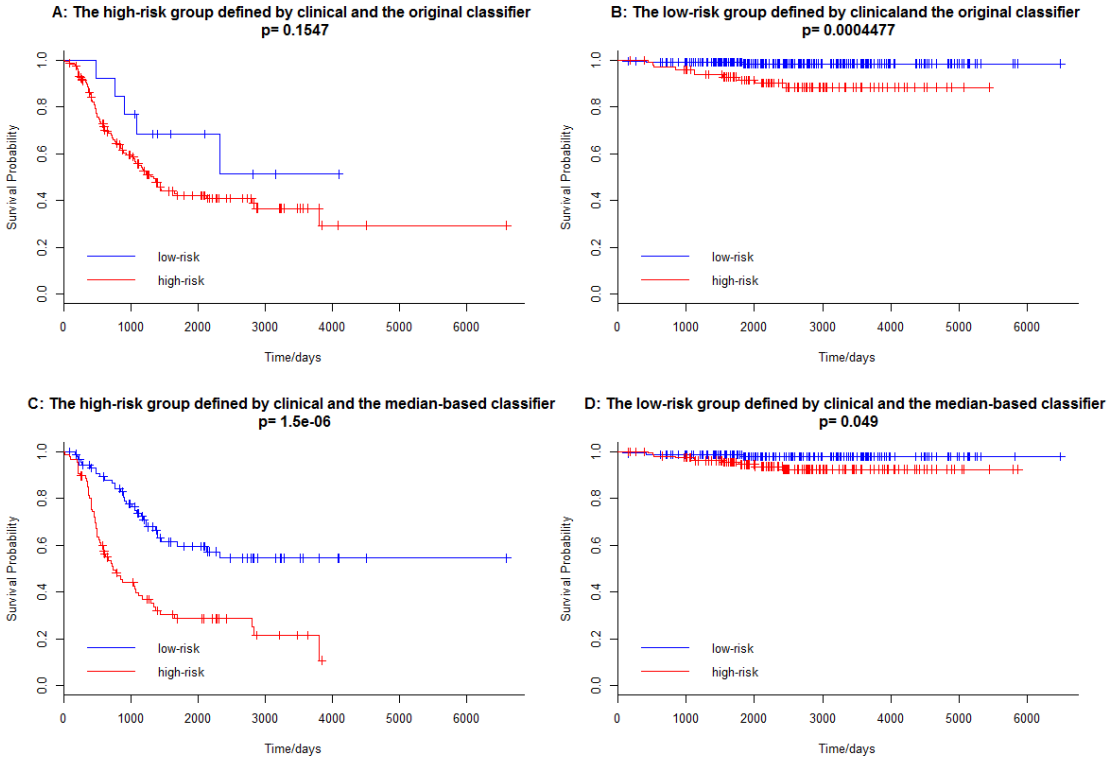

Figure 2 Kaplan-Meier curves and log-rank test of overall survival for patients in the clinical defined high-risk group and low-risk group.

A: Matched the classification result of RNA-seq based model for the high-risk group with 175 samples (162 high-risk and 13 low-risk) defined by clinical

B: Matched the classification result of RNA-seq based model for the low-risk group with 321 samples (86 high-risk and 235 low-risk) defined by clinical

C: Applied the RNA-seq based model for the high-risk group with 175 samples (87 high-risk and 88 low-risk) defined by clinical

D: Applied the RNA-seq based model for the low-risk group with 321 samples (160 high-risk and 161 low-risk) defined by clinical

According to the clinical information, patients were defined as “favorable”, “unfavorable” or unlabeled based on the patients’ response to chemotherapy. There were 86 patients defined as low-risk by the clinical but classified into high-risk group by our RNA-seq based model. We found the ratio for “unfavorable” or unlabeled of these 86 divergent patients (0.5) was significant larger than the other 235 patients (0.41) (P = 0.0001587 by chi-square test, Table 2), suggesting that the unlabeled patients in the “clinical defined low-risk but model defined high-risk” group were more likely to be unfavorable patients.

Table 2 The clinical labels among the 321 clinical defined low-risk group

	favorable	unfavorable	unlabeled	Sum
Clinical low-risk & model low-risk	137	3	95	235
Clinical low-risk & model high risk	43	10	33	86
Sum	180	13	128	321

By applying similar procedures for CNV data, we first selected 45 CNV loci which were correlated with overall survival by correlation test and finally selected three CNV loci by Cox regression method and stepwise regression method. Using these three CNV loci, we built the second risk-score model. This model can also classify those matched NB patients but the predictive power was weaker than the RNA-seq based model (P = 0.04752 in the testing set, Figure 3 A-B).

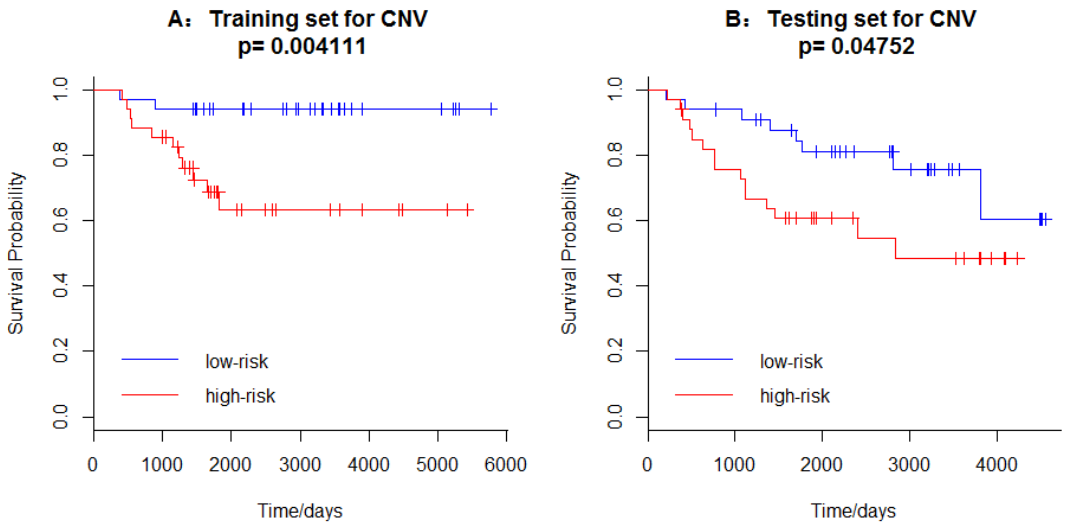

Figure 3 Kaplan-Meier curves and log-rank test of overall survival for patients in the training

set and testing set.

A: The CNV based model for the training set with 70 samples (35 high-risk and 35 low-risk) in the matched NB patients

B: The CNV based model for the testing set with 68 samples (34 high-risk and 34 low-risk) in the matched NB patients

To test whether integration of two different data (CNV and RNA-seq) can increase the predictive power or not, we combined the two individual models with two strategies. The first strategy was to define patients who were classified into high-risk group in both of previous two individual models as a new high-risk group (the intersection set), the predictive power of the new model can be significantly improved (P = 0.0001625 in the testing set, Figure 4 A-B). The second strategy was to combine the high-risk samples defined by two individual models into a new high-risk group (the union set), but the predictive power of this strategy wasn't improved (P = 0.07925 in the testing set, Figure 4 C-D). However, the clinical defined high-risk samples were entirely included in the expanded high-risk group.

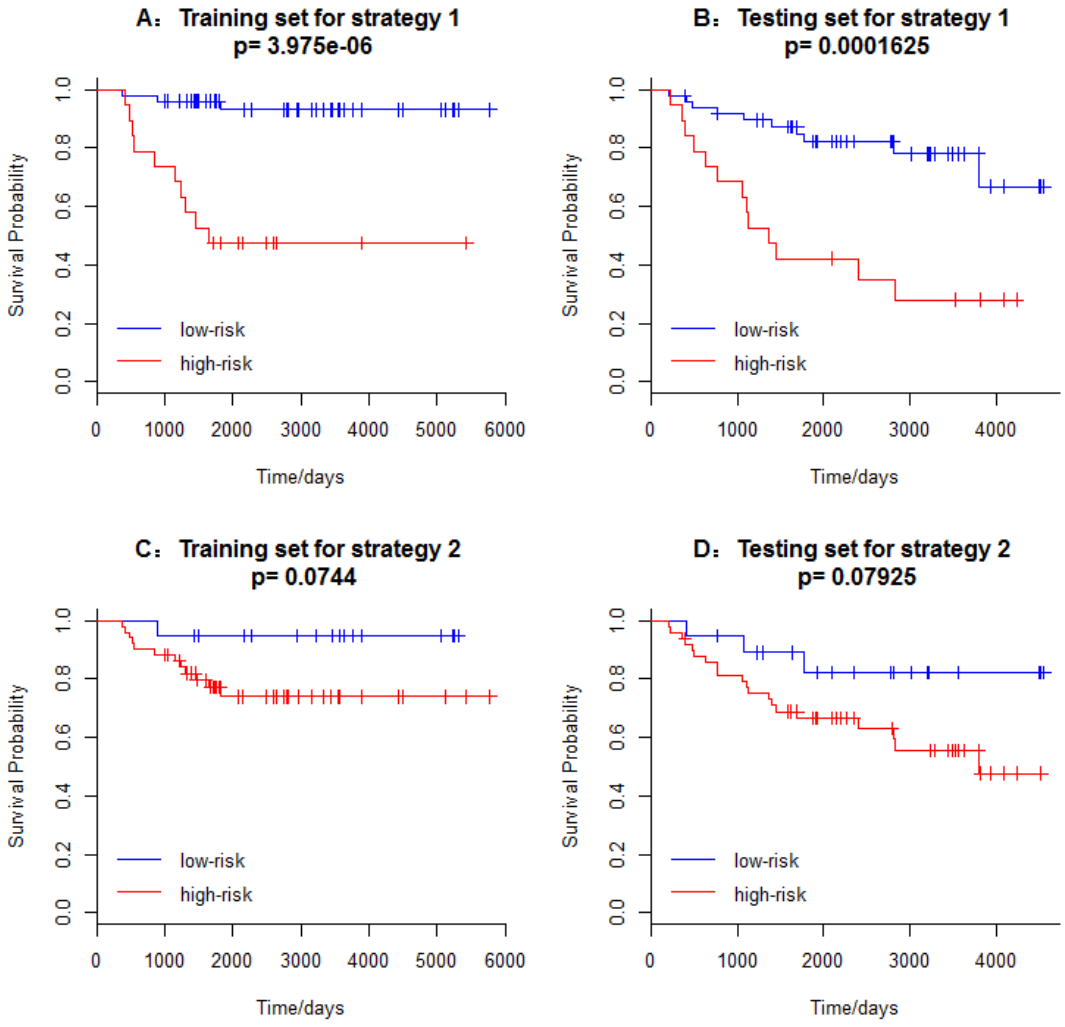

Figure 4 Kaplan-Meier curves and log-rank test of overall survival for patients in the training

set and testing set by two combining strategies.

A: The combined model for the training set with 70 samples (19 high-risk and 51 low-risk) by strategy 1

B: The combined model for the testing set with 68 samples (19 high-risk and 49 low-risk) by strategy 1

C: The combined model for the training set with 70 samples (51 high-risk and 19 low-risk) by strategy 2

D: The combined model for the testing set with 68 samples (49 high-risk and 19 low-risk) by strategy 2

According to clinical information, about three quarters of the NB patients were alive and most patients were defined as low-risk, about half of the patients have not been labeled for treatment response endpoint. When we redefined the high-risk group by overlapping the classification results of both of the models, which make the samples more consist with the risk distribution of this disease, the new model was significantly improved in predictive power, suggesting that different integration strategies for different purposes with different data should be chosen to improve the predictive performance. For example, if we want to select a refined high-risk group with high accuracy, we could choose strategy 1 to combine different models and refine the results. Otherwise, if we want to include as many high-risk patients as possible for specific study and do not want to miss any clinical defined high-risk patients, we could choose strategy 2 for this purpose.

Reference:

1. Zhang WQ, Yu Y, Hertwig F*, et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biology 2015;16.

2. Stigliani S, Coco S, Moretti S*, et al.* High genomic instability predicts survival in metastatic high-risk neuroblastoma. Neoplasia 2012;14(9):823-32.

3. Coco S, Theissen J, Scaruffi P*, et al.* Age-dependent accumulation of genomic aberrations and deregulation of cell cycle and telomerase genes in metastatic neuroblastoma. Int J Cancer 2012;131(7):1591-600.

4. Theissen J, Oberthuer A, Hombach A*, et al.* Chromosome 17/17q gain and unaltered profiles in high resolution array-CGH are prognostically informative in neuroblastoma. Genes Chromosomes Cancer 2014;53(8):639-49.

5. Zhang JX, Song W, Chen ZH*, et al.* Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. Lancet Oncology 2013;14(13):1295–1306.

6. Yu SL, Chen HY, Chang GC*, et al.* MicroRNA Signature Predicts Survival and Relapse in Lung Cancer. Cancer Cell 2008;13(1):48-57.

Correspondence: tielilushi@yahoo.com