Biostats 202
Final Project
9/12/19

Copy Number Variant as a Predictor of Death in Clear Cell Renal Carcinoma

Background

With advancements in technology, survival for previously terminal illnesses is improving dramatically. Cancer survival, in particular has improved dramatically over time. This is, in part, due to improved treatments and supportive measures, however it is also notably related to early detection and targeted therapies. While it is known that clinical features such as age, patient co-morbidities, tumor grade and stage play a significant role in prognosis, we are only now learning the molecular mechanisms that drive the pathogenesis of cancer. With greater information about molecular mechanisms that contribute to disease severity, we may derive therapeutic interventions to specifically target these mutations thus improving outcomes. One such method for identifying patterns of molecular changes in measurement of tumor-derived copy number variants (CNV). CNV detection is a more sensitive means of measuring the genetic changes specific to tumors. In the Kidney Renal Clear Cell Carcinoma (KIRC) dataset, genomic data was collected from 243 variables including CNV for 69 chromosomal loci. These data may be used to determine if there are patterns of genomic changes specific to tumors that may be associated with higher risk for mortality with renal cancer.

Research question

Do specific genomic footprints, measured by copy number variant expression predict risk of death in adults with renal cell carcinoma? We hypothesize that certain CNV footprints will be able to predict risk of death from renal cancer. Further studies may identify CNV patterns that are more likely to be associated with specific stage or grade of cancer.

Variables/Predictors:

The predictors in this study will be copy number variant (CNV) expression. The dataset offers a copy number for each genetic loci that is measured. Each CNV is coded with a portion of a chromosome. For example, CNV 20q indicates a copy of the short arm of chromosome 20.

Outcome

The primary outcome is overall survival.

Methods

This will be a supervised machine learning algorithm since there is a defined outcome of interest with identified predictors. Originally, I had intended to use mRNA expression rather than CNV expression, however there are >25,000 mRNA results and SPSS was unable to process these with sufficient speed. The CNV data did not contain any missing data or fields that could not be analyzed by SPSS. After reading in the raw data as a csv file, a type node was applied. In reviewing the associated codebook, I made the “feature” the record ID, “age” a continuous variable, “gender” a flag variable, “grade” and “stage” were ordinal/categorical variables, and “OS_vital_status” (survival) a flag variable. All CNV variables were listed as continuous variables.

I then applied a filter node to remove all mRNA and miRNA variables from the analysis. Since I was evaluating multiple outcomes (survival, stage and grade), I subsequently applied type nodes to the filter node to select each of these as a target while keeping all other variables listed as inputs.

I will compare results using traditional partitioning of train/test (70/30), train/validate/test, and train/test with cross-validation. With this sample size, it may be beneficial to use cross validation to ensure sufficient data is input into the model to increase predictive ability.

Results

Review of the clinical data is presented in figure 1. Data is stratified by individuals who survived vs those who did not survive. As seen in the figure, there were more men than women in this data set however survival did not affect one sex over the other. Distribution of tumor rgade and stage are presented in figure 1. Of note, proportion of subjects who survive decreases as both stage and grade increase. There is a statistically significant difference in the survival by tumor grade and stage, but not by sex.

Table 1. Patient Characteristics

	All Subjects n (%)	*Survived n (%)*	*Did not survive n (%)*	*p-value**
n	243 (100)	153 (63)	90 (37)	
Mean Age (y)	61.276 (range 26-90)	58.6	65.9	
Sex				
Male	162 (33.3)	103 (63.6)	59 (36.4)	p=0.78
Female	81 (66.67)	50 (61.7)	31 (38.3)	
Tumor Grade				p=0.001
Grade 1	4 (1.7)	4 (100)	0 (0)	
Grade 2	98 (40.3)	80 (81.6)	18 (18.367)	
Grade 3	95 (39.1)	57 (60)	38 (40)	
Grade 4	45 (18.5	12 (26.6)	33 (73.3)	
Grade X	1(0.4)	0 (0)	1 (100)	
Tumor Stage				p=0.001
Stage 1	104 (42.8)	87 (83.7)	17 (16.3)	
Stage 2	20 (8.2)	15 (75)	5 (25)	
Stage 3	68 (28)	40 (58.9)	28 (41.2)	
Stage 4	51 (21)	11 (21.6)	40 (78.4)	

*p values derived from Kruskall-Wallis equality of proportions rank test

Using several classification methods, multiple models were created. In particular, I had run models for logistic regression, K-nearest neighbors and neural networks. In efforts to achieve highest predictive capacity, I adjusted each model to address feature selection as well as to test model performance depending on whether the model was tested on one testing set vs on a testing and validation set. These results are presented in table 2.

Discussion

In general, CNV expression how relatively sensitivity in predicting survival in this population with average accuracy of models in the 50-60% range. Many models had very poor sensitivity for predicting death associated with CNV status, however most models had high specificity. This indicates that in the absence of the CNVs, there is a higher likelihood that the predictive model would provide an accurate outcome. Of the models tested, K nearest neighbors models using both forward selection and cross validation provided the most accurate and specific outcomes with an AUC in the 70% range.

Though this study does include a vast number of variables, the population is not too large, thus limiting generalizability. A larger population may provide additional insight into trends that may be more true in the general population. The small patient population and large difference between changes seen in the training vs testing and validation sets also suggests that these models are at risk for overfitting. Additionally, almost all models used different predictors to create the model. This argues that the models do not necessarily reflect the research question asked (can we identify CNVs that correlate with risk of death?). Perhaps a more suitable research question would be whether CNV data can be used to predict risk of death in patients with kidney cancer. Further studies may be done to achieve more sensitive evaluation of particular CNVs that may predict poor outcomes in patients with kidney cancer. Additionally, models may be developed using these data to determine certain CNVs that may correlate with risk for a particular stage or grade of cancer. This may be more useful in determining risk of mortality as it is shown in figure 1 that both tumor grade and stage are highly predictive of mortality.

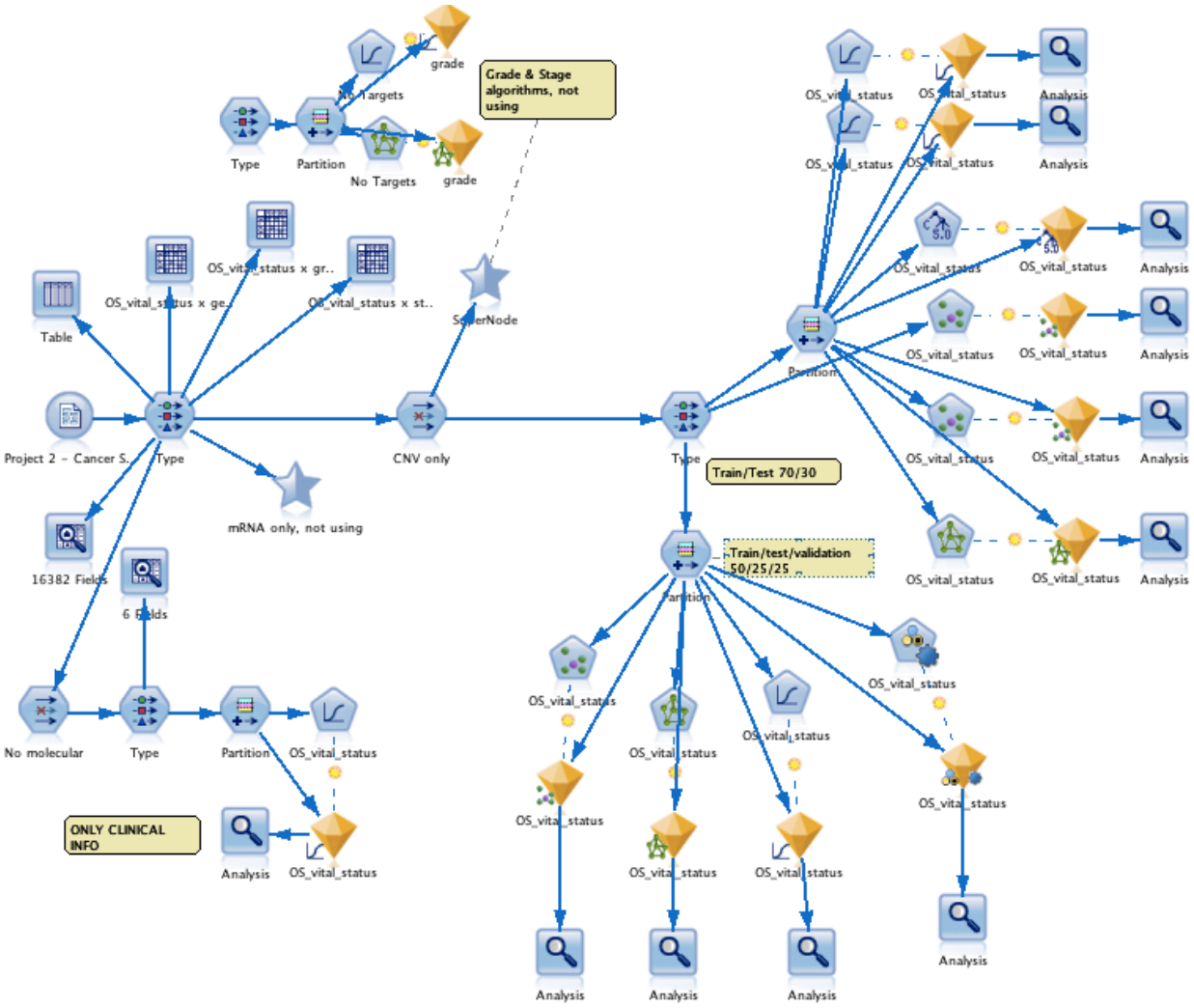
Table 2: SPSS Model Summary

Model	Model parameters	Partition*	Model Performance	Predictor Importance (up to 3)
Logistic Regression	Forward or stepwise procedure	T/T (70/30)	Testing set accuracy: 63.6% Sensitivity: 40% Specificity: 80.1 AUC: 0.624	1. CNV 3p25.3 2. CNV 9p21.3
Logistic regression	multinomial procedure	T/T (70/30)	Testing set accuracy: 63.6% Sensitivity: 44% Specificity: 73.1% AUC: 0.569	1. CNV 12p 2. CNV 8q 3. CNV 20q
C5.0	No cross validation	T/T (70/30)	Testing set accuracy: 61% Sensitivity: 32% Specificity: 75% AUC: 0.53	1. CNV 9p21.3 2. CNV 1p31.1
C5.0	with cross validation (10 fold)	T/T (70/30)	Testing set accuracy: 61% Sensitivity: 32% Specificity: 75% AUC: 0.53	1. CNV 9p21.3 2. CNV 1p31.1
K nearest neighbors	feature selection	T/T (70/30)	Testing set accuracy: 76.6% Sensitivity: 40% Specificity: 94% AUC: 0.738	1. CNV 1q43 2. CNV 9q21.3 3. CNV 3q21.32
K nearest neighbors	Cross Validation, without feature selection	T/T (70/30)	Testing set accuracy: 64.9% Sensitivity: 8% Specificity: 92.3% AUC: 0.682	n/a. K selection: lowest error rate wth K=14
Neural Network	Did not change default settings	T/T (70/30)	Testing set accuracy: 52% Sensitivity: 28% Specificity: 63% AUC: 0.485	1. CNV 11p 2. CNV 8q24.22 3. CNV 15q

Model	Model Paremeters	Partition*	Model Performance	Predictor Importance (up to 3)
Logistic	Forward Stepwise	T/V/T (50/25/25)	Validation set accuracy: 54.4% Testing set accuracy: 63.3% Validation Sensitivity: 26% Validation Specificity: 86.4% Testing Sensitivity: 22.8% Testing Specificity: 67% Validation AUC: 0.656 Testing AUC: 0.529	1. CNV 3p25.3 2. CNV 1p36.23
K nearest neighbors	With Feature Selection	T/V/T (50/25/25)	Validation set accuracy: 76.7% Testing set accuracy: 75.5% Validation Sensitivity: 48% Validation Specificity: 95% Testing Sensitivity: 36% Testing Specificity: 91% Validation AUC: 0.801 Testing AUC: 0.72	Predictor importance not reported
K nearest neighbors	Cross validation, without feature selection	T/V/T (50/25/25)	Validation set accuracy: 75% Testing set accuracy: 67.7% Validation Sensitivity: 35% Validation Specificity: 100% Testing Sensitivity: 18% Testing Specificity: 91% Validation AUC: 0.786 Testing AUC: 0.659	Predictor importance not reported
Neural Networks	Did not change default settings	T/V/T (50/25/25)	Validation set accuracy: 70% Testing set accuracy: 57.35% Validation Sensitivity: 48% Validation Specificity: 84% Testing Sensitivity: 32% Testing Specificity: 70% Validation AUC: 0.623 Testing AUC: 0.406	1. CNV 21q 2. CNV 8q24.22 3. CNV 8p

*T/T = train/test partition, T/V/T = train/validate/test partition, proportion of dataset indicated in parentheses.

Final Stream:



References:

Molparia, B., Nichani, E. and Torkamani, A. (2019). *Assessment of circulating copy number variant detection for cancer screening*.