**Introduction**

Estimations of cancer expenditure project that on the year 2020 medical spending on cancer can reach over 200 billion dollars[1]. Although pediatric cancer is less prevalent than in the adult [2], the economics and social cost of a child with cancer are tremendous. It can impact profoundly on the kid’s health and development, in the family’s psychological health and economics, and in the society by increasing disability-adjusted life years [3, 4]. The most frequent pediatric cancer is acute leukemia, being the lymphoblastic (ALL) form the most common [2]. On the US, every year 2500 to 3500 new cases of ALL are diagnosed in children [5, 6]. According to the needs, NIH has allocated over the past years a significant amount of resources on Cancer investigation [7]. This research produced advances in the therapeutics that has increase the event-free survival times [8], also provided new insights about the mutations related to ALL that were associated with a variety of outcomes and clinical characteristics.

The research output changed the clinical practices, the clinicians started to classify into groups of risk and treat according to their patients’ clinical phenotypes. The clinical characteristics used to group the patients can differ depending on the research group. For example, Berlin-Frankfurt-Münster (BFM) team base their risk groups solely on treatment response criteria, such as the prednisone prophase response, the minimal residual disease (MRD) at the end of induction phase (week five), and the minimal residual disease at the end of consolidation phase (week 12) [9]. The Children’s Oncology Group (COG) propose to stratify children with ALL according to a subset of prognostic factors like age (aged 1 to <10 years), white blood cell count at diagnosis (<50,000 cells/µL), MRD at the end of induction phase (day 29). COG also considers that genetic findings are important, for example the presence of intrachromosomal amplification or Extreme hyperploidy (59 to 84 chromosomes) or hypodiploidy (fewer than 45 chromosomes), among other factors [10]. Despite the great efforts in research some of the clinical phenotype groups obtained a projected five-year event-free survival of only 75-80%[11-14], this means that probably we are classifying different types of patients together. This implies that we are providing the same treatment to different types of cancer, thus we over treat in some cases increasing the risk of adverse event and in other patients we under treat lessen the chances to obtain a remission.

The research in the field has shown that JAK mutation in high-risk patients had a gene expression signature similar to BCR-ABL1, the last a common mutation in Philadelphia chromosome-positive ALL [15]. Also, gene expression profiling revealed that a rearrangement of cytokine receptor-like factor 2 (CRLF2) is associated with mutations of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome [16]. Research using gene expression profiling and the correlation with genome-wide DNA copy number abnormalities, has been able to establish novel cluster groups that may serve as new targets for diagnosis, risk classification, and therapy [17]. The use of deep whole-exome sequencing has provided insights into the genetics of ALL that could drive the patient to a relapse [18]. Genome-wide DNA copy number and deep whole-exome sequencing are not routinely used in clinical practice, an exception is the gene expression technology. The development of novel methodologies, using the available gene expression technology, to further classify high risk patient with ALL is required.

This project aims to reproduce the proposed by Kang et al [8] methodology to analyze gene expression data. We will develop a COX-regression model based on the principal component analysis (PCA) of the gene expression’s COX-score. Our analysis is limited to open source software. The study findings will be compared against the findings of Kang’s group [8] each step of the process.

**Previous work**

Since Acute lymphoblastic leukemia (ALL) is the most common form of pediatric cancer, it is crucial to classify children with high-risk ALL into different risk groups and provide them with the corresponding treatment[2, 11]. A recent work has shown that the gene expression classifier and flow cytometric measures of minimal residual disease (MRD; P = 0.00Q) each provided independent prognostic information[8, 11]. The combination of these two classifiers improves the risk classification. Methods that were used were supervised learning algorithms and cross­validation techniques. These methods were applied to build a 42­probe­set (38­gene) expression classifier predictive of RFS for 207 uniformly treated children with high-risk ALL[5, 11]. To test the predictive power of gene expression classifier for RFS relative to flow cytometric measures of MRD and to other clinical and genetic variables, they applied a multivariate proportional Cox hazards regression analysis. They used diagonal linear discriminant analysis to build a prediction model between gene expression classifier and end­induction MRD. To evaluate the model, they applied the likelihood­ratio test (LRT) score and the prediction error rate[2, 5].

**Improvements**

By looking at the variables in our dataset, both clinical data and Affy-array data, we could try to find out which variables are the ones with the most significant classification power of high-risk ALL. By including clinical characteristics based on demographics, such as age, race, gender, and ethnicity; clinical data, central nervous system or testicular involvement; laboratory data as white cell count at the time of the diagnosis we are making a prediction model more dynamic and robust. We then would couple with these variables with the Affy-array data to further enhance the prediction power of high-risk ALL. Having a multifaceted prediction model would help us better classify high-risk ALL children, and improve current treatment time. With the current treatment of ALL, time is the key factor in survival rate. If the leukemia is caught at an early stage the survival rate of this high-risk ALL group improves dramatically. For this main reason, we think that this prediction model will save the lives of this high-risk group. We are still looking at our data to see which variables are the best performers. Once we have found which variables are the best performers from our data, we will incorporate them in our prediction model.

**Methodology**

Affy-Array Data:

We obtained the CEL files from the ftp site of TARGET ALL phase 1 project, then we read each file using R affy package and then to normalize the data, we apply a robust multi-array average (RMA) [19] over probes on all the patients. We filter out the probes with features exhibiting little variation, or consistently low signal across the samples [20] over the normalized expression set. The normalization and the filtering process differs from Kang et at [8] methodology, they processed the CELL files using the commercial software Affymetrix GeneChip® Operating Software 1.4.0 Statistical Algorithm package; they filtered out the probe set that were present in less than 50% of the samples. Then we downloaded the clinical data set from the ftp site of TARGET ALL phase 1 project and join it with the filtered expression set. The probe annotations where obtained from the hgy133plus2.db using the annotate package [21].

Statistical Analysis:

We mainly follow the Kang et al. paper[8] and use supervised PCA [22] to detect genes highly associated with survival rate, also classify the risk groups based on their genes instead of pretreatment clinical data (for example: age, ethnicity).

**Part I:** Building prediction model and Cox-regression model

Step 1. Calculate the cox-score (see algorithm next page) for the gene record it as and then we rank the genes by ordering . The bigger the value of the highest association with the survival rate.

We calculate the cox score for the whole data set and get the following result

Step 2. is a threshold, and we only consider the genes that has > .

Suppose is given, there are genes satisfy > , for the patient, the standardized gene expression level is ()

(1) Principal component analysis is performed on the standardized expression values of the remaining genes. Principal component analysis is a popular approach of dimension reduction and

it helps produce a smaller number of linear combinations of variables. By adopting PCA method, we could extract important variables while still retain the important information in the data. The first principal component is the variable that contributes most to the variance of the data set and so on.

(2) For this data set, we take the first principal component of PCA (most relevant components), which gives loading value of selected genes , then we can get the PCA score for the patient as a linear combination:

(1)

Denote the predicted PCA score for 207 patients as .

(3) Cox proportional hazard model assumes that the instantaneous event rate at time is of the following form:

When covariate is at different level, say , the ratio of hazard rate is proportional to the difference of the covariate level:

)

We assume the hazard rate for the relapse of leukemia follows the Cox model, the covariate here is the PCA score (see equation (1)), and the hazard rate is . Since PCA score is a linear combination of the highly associated genes, we evaluate the genes related to the hazard rate of leukemia.

Cox-regression can be done with the coxph(surv(time,status)~). The LRT (likelihood ratio test) for this model will be reported as a measure of goodness of fitting.

Step 3: Test the performance of the cox-regression model:

Follow from the Cox model , we fit the coefficient with . For the patient, the linear combination of the genes

(3)

is used as a prediction model.

To exam the power of the prediction, we can predict the PCA score on new samples, let’s say patient , and get . Then cox-regression can be fitted to the survival time and the gene expression level of the new samples. The performance of the cox-regression can be evaluated by LRT. The larger LRT, the better the performance.

**Part II:** Using CV (cross validation) to select threshold

In part I step 2 is under the condition that is given. To select , an easier way might be we just assign a value to , so that it only includes the top, let’s say, 10% or 5% highly expressed genes. However, this kind of assignment is not that convincing, and one possible way to improve it is to use CV to select.

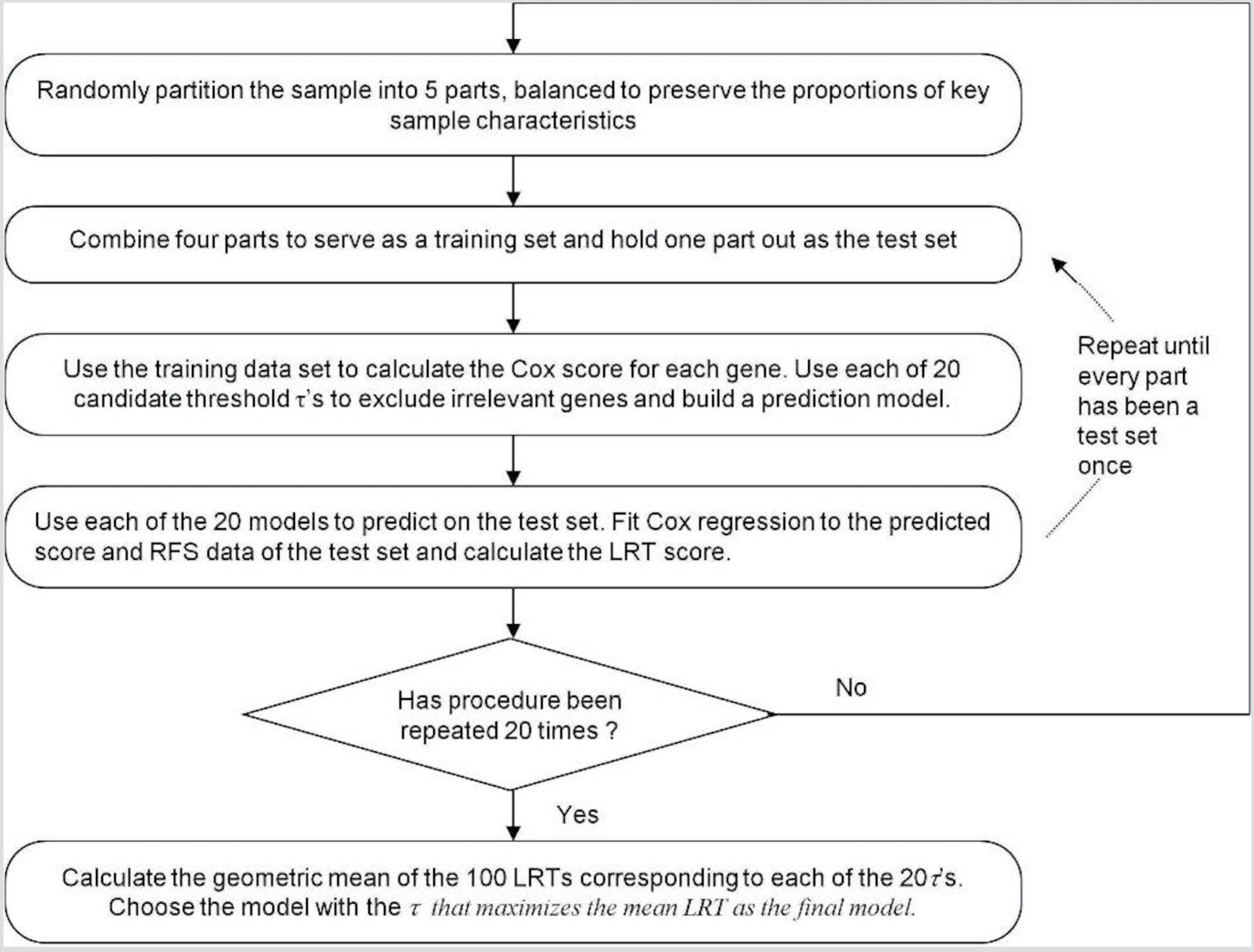
Cross-validation is a popular method in statistical analysis nowadays. It is quite powerful in measuring the prediction power of a statistic model. One way to measure the predictive ability of

a model is to test it on a set of data not used in estimation. In many real problems, we don't have enough data to set aside a large test set and this motivates the CV. The idea for using CV is to avoid over fitting and make model more robust.

The general ideal of k-fold CV is interpreted in the following graph [23]. We train model at the training folds, and test the model in the test fold. Each test fold will give a measurement of error or performance of the model built by the training folds. The model that has the smallest averaged error or best average performance would be the selected model.



To select , an example of 5-fold CV for selecting combined with part I is the following graph [8]:



After the filtering, our sample size of the data reduces to 207. In order to approach the 5-fold cross validation, we need to separate the data into 5 parts. Apart from that, we expect the stratified data to be balanced with the key sample characteristics in order to achieve better classification results. As a result, we proposed the stratified random sampling method.

The first step is to decide the key features of the data that are crucial to classification. Based on the information provided by National Cancer Institute (NCI), the Patient and clinical disease characteristics affecting prognosis include the following [24]: age at diagnosis, white blood cell (WBC) count at diagnosis, central nervous system (CNS) involvement at diagnosis, testicular involvement at diagnosis, down syndrome (trisomy 21), sex, race and ethnicity, and weight at diagnosis and during treatment. The key clinical features used to group the patients vary from data to data. By cross-checking the keys clinical features with our gene data set, we find out that some key features have lots of missing data, which makes them unusable. All things considered, we decide to base our sampling on the following three features: age, WBC count at diagnosis and MRD (minimal residual disease) day 29.

The next step is to approach the stratified random sampling. We expect each test data set to be balanced with all the key clinical features. First, we partition the data into 8 strata based on the three key features mentioned above. We set three dummy variables and for each key clinical feature as following:

As a result, we can partition the sample data into 8 stratums based on the combination of three key features. We get the sample size for these eight stratums are 4,14, 47, 60, 0, 7, 24, 51, which yields the total sum same as sample size 207. Noticed that there is one stratum that actually has no patient, thus we eliminate that particular stratum. In the following, we will randomly partition each stratum into 5 subgroups and then each subgroup from all 7 stratums are pooled as a random sample which functions as a fold in the 5-fold cross-validation. If we just do the random sampling without any restriction, we may observe the result that the smallest sample size is 38 and largest sample size is 43, which differs a lot. Since our total sample size 207 is not a large number, we want to try to balance the sample size for each fold to make them as closed as possible in order to avoid the bias. Thus, we put some effort into achieving this goal.

The algorithm of controlling the sample size in stratified random sampling is summarized below (check the sampling\_the\_data.py for codes):

1. Index the patient in the sample stratum by stratum. For instance, the index for the first stratum is 0, 1, 2, 3 and the index for the second stratum is from 4 to 13.
2. Create an empty list that consists of 5 lists as the 5 folds. For each of the 8 strata, we randomly distribute equal number of patients into 5 folds and leave out the remaining. For example, there are 14 patients in the second stratum, we randomly distribute 2 patients into 5 folds and leave out 4 patients as the remaining.
3. After each loop of distribution, we remove those patients from the original data set to avoid recurrent selection.
4. We distribute the remaining patients stratum by stratum.
5. First, we shuffle the remaining patients in the first stratum to achieve randomness. And we distribute the remaining 4 patients into 4 folds.
6. Then, we sort the length of each fold by the length from the smallest to the largest.
7. Next, we shuffle the patients in the second stratum and then distribute them into the 5 folds by the order of length. In this way, we make sure the small size fold is always assigned first.
8. Repeat steps 1) to 3) for each stratum until all the remaining patients are distrusted.

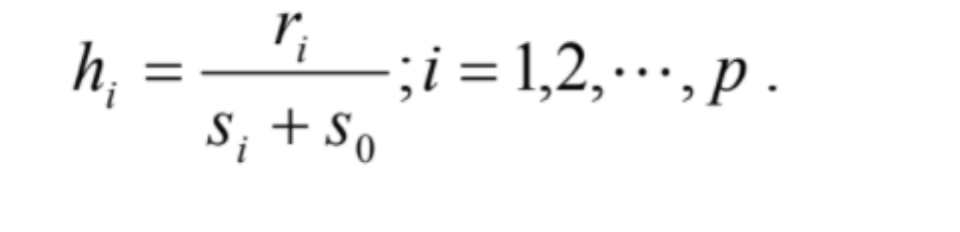
**Part III** Supplementary[8]:

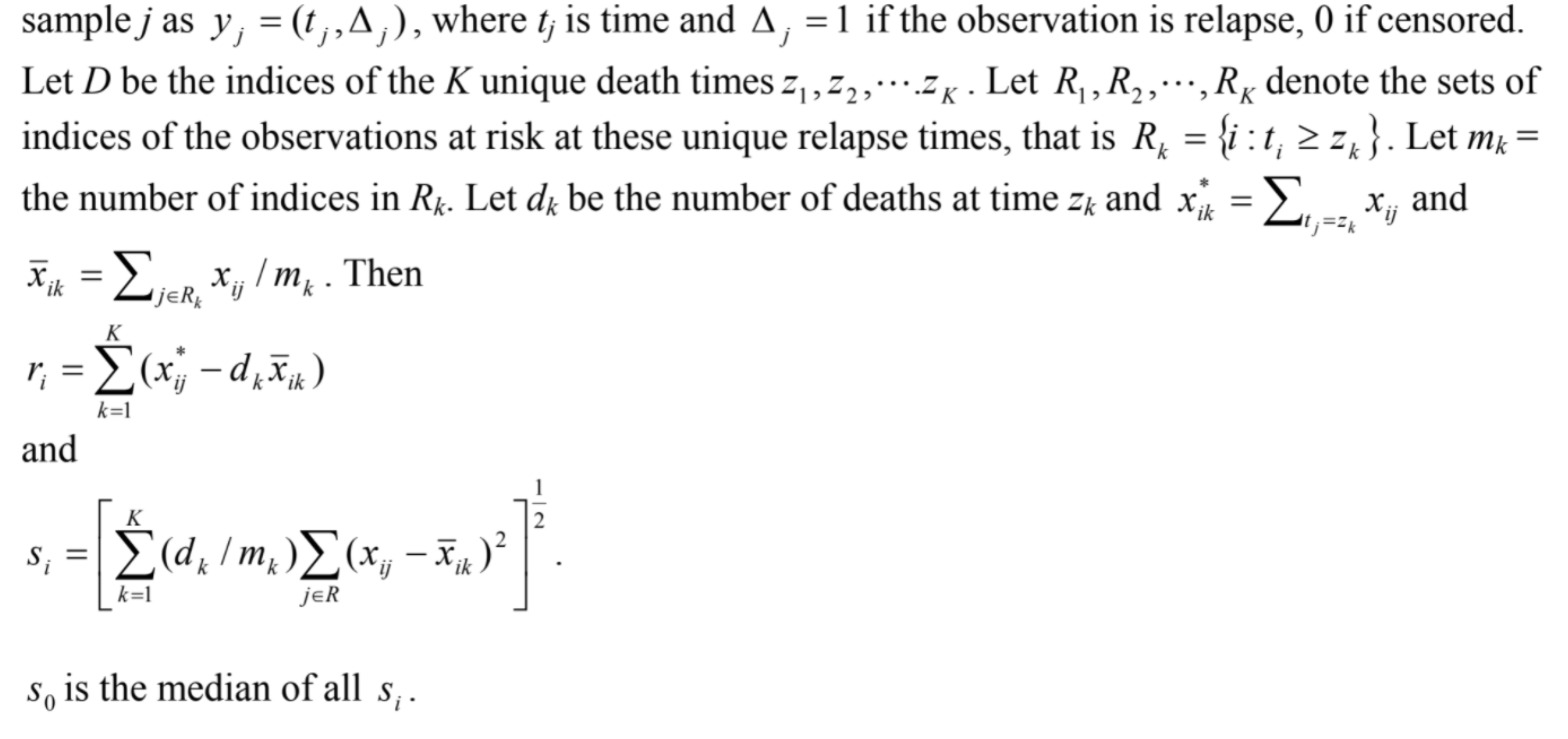
1.

2. Calculating cox-score:

Let denote genes and denote patients.

The cox score for gene is :





**Expected results**

(1) After applying the algorithm in the supplementary part, we will have cox score for each gene.

(2) Based on the cox score of genes, we can set possible candidates threshold.

(3) Randomly partitioning the 207 patients to test set and training set.

(4) Perform CV on the training and test dataset. we need to recalculate the cox-score for each gene based on the training dataset, then select genes based on different threshold, do PCA and cox-regression. Predict the PCA score and perform cox-regression in the test dataset. Each threshold will give us a LRT score. We can select the threshold base on the highest LRT score. (or based on the principle of parsimony, select the second or third highest LRT which corresponding to a set of fewer genes).

(5) The risk of each patient can be predicted by PCA score and further classify the patients on high/low risk group.

(6) Use Kaplan-Meier estimator to plot the survival rate of the high/low risk group, which would be a fine classification of the 207 patients.

**Results**

By reading the 207 downloaded CELL files and after the RMA, we obtained for each patient 54,675 probes (data frame of 54,675 rows and 207 columns). After filtering the probes using the described criteria we ended up with 21,148 probes per patient.

Table1. Comparison of microarrays data processing

|  |  |  |
| --- | --- | --- |
| Step | Kang et al. [8] | Group project |
| Cell Files | 207 | 207 |
| Probes after RMA | 54,675 | 54,675 |
| Probes after filtering | 23,775 | 21,148 |

**References**

1. *Cancer costs projected to reach at least $158 billion in 2020*. 2015 2015-07-23 2017-09-10]; Available from: <https://www.ncbi.nlm.nih.gov/pubmed/>.

2. Dores, G.M., et al., *Acute leukemia incidence and patient survival among children and adults in the United States, 2001-2007.* Blood, 2012. **119**(1): p. 34-43.

3. *WHO | Disability-adjusted life years (DALYs)*. WHO 2017 2017-01-27 15:23:13 2017-09-10]; Available from: <http://www.who.int/gho/mortality_burden_disease/daly_rates/text/en/>.

4. *WHO | Metrics: Disability-Adjusted Life Year (DALY)*. WHO 2014 2014-03-11 14:56:00 2017-09-10]; Available from: <http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/>.

5. Ward, E., et al., *Childhood and adolescent cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(2): p. 83-103.

6. Svendsen, A.L., et al., *Time trends in the incidence of acute lymphoblastic leukemia among children 1976-2002: a population-based Nordic study.* J Pediatr, 2007. **151**(5): p. 548-50.

7. *NIH Categorical Spending -NIH Research Portfolio Online Reporting Tools (RePORT)*. 2017 2017-09-10]; Available from: <https://report.nih.gov/categorical_spending.aspx>.

8. Kang, H., et al., *Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia.* Blood, 2010. **115**(7): p. 1394-405.

9. Conter, V., et al., *Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study.* Blood, 2010. **115**(16): p. 3206-14.

10. Schultz, K.R., et al., *Risk- and response-based classification of childhood B-precursor acute lymphoblastic leukemia: a combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG).* Blood, 2007. **109**(3): p. 926-35.

11. Koo, H.H., *Philadelphia chromosome-positive acute lymphoblastic leukemia in childhood.* Korean J Pediatr, 2011. **54**(3): p. 106-10.

12. Moricke, A., et al., *Risk-adjusted therapy of acute lymphoblastic leukemia can decrease treatment burden and improve survival: treatment results of 2169 unselected pediatric and adolescent patients enrolled in the trial ALL-BFM 95.* Blood, 2008. **111**(9): p. 4477-89.

13. Moghrabi, A., et al., *Results of the Dana-Farber Cancer Institute ALL Consortium Protocol 95-01 for children with acute lymphoblastic leukemia.* Blood, 2007. **109**(3): p. 896-904.

14. Veerman, A.J., et al., *Dexamethasone-based therapy for childhood acute lymphoblastic leukaemia: results of the prospective Dutch Childhood Oncology Group (DCOG) protocol ALL-9 (1997-2004).* Lancet Oncol, 2009. **10**(10): p. 957-66.

15. Mullighan, C.G., et al., *JAK mutations in high-risk childhood acute lymphoblastic leukemia.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9414-8.

16. Harvey, R.C., et al., *Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia.* Blood, 2010. **115**(26): p. 5312-21.

17. Harvey, R.C., et al., *Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome.* Blood, 2010. **116**(23): p. 4874-84.

18. Zhang, J., et al., *Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group.* Blood, 2011. **118**(11): p. 3080-7.

19. Gautier, L., et al., *affy--analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-15.

20. Hahne, R.G.a.V.C.a.W.H.a.F., *genefilter: genefilter: methods for filtering genes from high-throughput experiments*. 2017.

21. Gentleman, R., *annotate: Annotation for microarrays*. 2017.

22. Bair, E., et al., *Prediction by Supervised Principal Components.* Journal of the American Statistical Association, 2006. **101**(473): p. 119-137.

23. *k-fold croos-validation*. 2017 2017-09-27]; Available from: <https://sebastianraschka.com/images/faq/evaluate-a-model/k-fold.png>.