1. 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]. There are two types of lymphoid cells, B and T, therefore exist two types of ALL: B-cell (Early pre-B ALL prevalence of 10%, Common ALL 50%, Pre-B ALL 10%, Mature B-cell ALL –Burkitt leukemia- 4%) and T-cell (Pre-T ALL 5-10%, and Mature T-cell ALL 15-20%) [2]. The symptoms experienced by the patients depends on the cell line that is affected by the rapid cell grow, if the red line is affect it will lead to anemia and the patient will be pale, if the megacariocit is affected the patient will experience easy bruising, and because of the number of white cells in the body the patient can experience fever and enlarged live and or spleen [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.

2. 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 cross-validation techniques. These methods were applied to build a 42-probe-set (38-gene) 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 end-induction MRD. To evaluate the model, they applied the likelihood-ratio test (LRT) score and the prediction error rate[2, 5].

3. 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.

As we progress which our project we had to come up with novel methods to be able to reproduce the results of the Kang et al [8] paper, which we got our data set from. Two major innovations that

we came up with are: novel stratification method and using the open source R to normalize our Arry-array data. The first of the major innovations our group has come up with is the novel stratification method where we divide the data set into 8 stratums which is based on the combination of 3 key clinical features. For each stratum, we random separate it into 5 subgroups. And then we pick one subgroup from each stratum and combine the data as one test data set. In this way, we partition the data into 5 folds and also balance the data to preserve the key clinical features. The second major innovation that our group came up with; was using R to normalize the Affy-array data. In the Kang et al [8] paper, where our data set was generated did not go into details about the method used to normalized the Affy-array data. In order to reproduce the same results that were produced in the Kang et al [8] paper we need to find a way to normalize the arrays in way where that would be similar in the mythology used. With R we were able to normalized the Affy-array using the open source bioLite pakage.

4. 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 ith gene record it as r_i and then we rank the genes by ordering $|r_i|$. The bigger the value of $|r_i|$ the highest association with the relapse free survival.
The dataset consists of four conditions: relapse, death, censored and SMN. we consider relapse as events and all the other three cases as censored. (death, censored and SMN).

Step 2. τ is a threshold, and we only consider the genes that has $|r_i|> \tau$.
Suppose τ is given, there are p genes satisfy $|r_i|> \tau$, for the jth patient, the standardized gene expression level is $(x_{1j,}x_{2j,} \dots, x_{pj})$
(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 $(\phi_1, \phi_2, \dots, \phi_p)$, then we can get the PCA score for the jth patient as a linear combination:
$w_j = \phi_1 * x_{1j} + \phi_2 * x_{2j} \dots + \phi_p * x_{pj}$ (1)
Denote the predicted PCA score for 207 patients as $w = (w_1, w_2, \dots, w_{207})$.
(3) Cox proportional hazard model assumes that the instantaneous event rate at time $t\ is\ \lambda(t|Z)$ is of the following form:

$$\lim_{\Delta t \to 0} \frac{P(T < t + \Delta t | T > t)}{\Delta t} = \lambda(t|Z) = \lambda_0(\mathrm{t}) \exp(\beta^T Z) \qquad (2)$$

When covariate is at different level, say Z_1, Z_2 , the ratio of hazard rate is proportional to the difference of the covariate level:
$\frac{\lambda(t|Z_1)}{\lambda(t|Z_2)} = \exp(\beta^T(Z_1 - Z_2))$
We assume the hazard rate for the relapse of leukemia follows the Cox model, the covariate Z here is the PCA score w (see equation (1)), and the hazard rate is $\lambda(t|Z) = \lambda_0(\mathrm{t}) \exp(\beta w)$. Since PCA score w 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)~ w). 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 $\lambda(t|Z) = \lambda_0(\mathrm{t}) \exp(\beta w)$, we fit the coefficient with $\hat{\beta}$. For the jth patient, the linear combination of the genes
$w_j = \left(\phi_1 * x_{1j} + \phi_2 * x_{2j} \dots + \phi_p x_{pj}\right)$ (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 j', and get $w_{j'} = \left(\phi_1 * x_{1j'} + \phi_2 * x_{2j'} \dots + \phi_p x_{pj'}\right)$. 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 τ

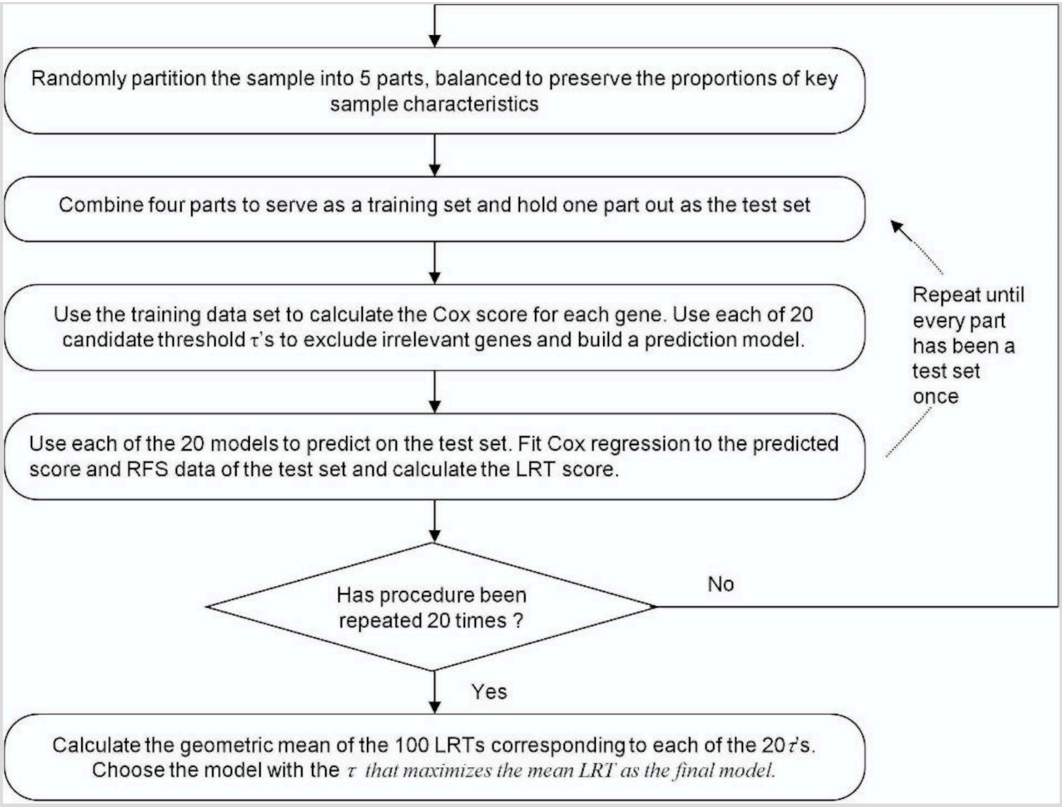
In part I step 2 is under the condition that τ is given. To select τ, an easier way might be we just assign a value to τ, so that it only includes the top, let's say, 10% or 5% highly expressed genes. However, this kind of assignment is not that convincing, and one possible way to improve it is to use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

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 Figure 1 [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 τ, an example of 5-fold CV for selecting τ combined with part I is the Figure 2 [8].

Figure 1. K-fold cross validation.

Figure 2. Cross-validation procedure of determine the best model for predicting RFS



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 x_1, x_2 and x_3 for each key clinical feature as following:

$$x_1 = \begin{cases} 1 & age < 1 \; or > 10 \\ 0 & if \; 1 < age < 10 \end{cases} \qquad x_2 = \begin{cases} 1 & WBC > 5k \\ 0 & WBC \leq 5k \end{cases} \qquad x_3 = \begin{cases} 1 & MRD \; day \; 29 \; postive \\ 0 & MRD \; day \; 29 \; nonpostive \end{cases}$$

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.
    1) First, we shuffle the remaining patients in the first stratum to achieve randomness. And we distribute the remaining 4 patients into 4 folds.
    2) Then, we sort the length of each fold by the length from the smallest to the largest.
    3) 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.
    4) Repeat steps 1) to 3) for each stratum until all the remaining patients are distrusted.

Part III Supplementary[8]:

1. Calculating cox-score:

Let i denote genes ($i = 1,2,3, \ldots., 21000$) and j denote patients ($j = 1,2,3, \ldots, 207$). The cox score for gene i is h_i :

$$h_i = \frac{r_i}{s_i + s_0}; i = 1,2, \cdots, p .$$

sample j as $y_j = (t_j, \Delta_j)$, where t_j is time and $\Delta_j = 1$ if the observation is relapse, 0 if censored. Let D be the indices of the K unique death times $z_1, z_2, \cdots z_K$. Let $R_1, R_2, \cdots, R_K$ denote the sets of indices of the observations at risk at these unique relapse times, that is $R_k = \{i : t_i \geq z_k\}$. Let $m_k =$ the number of indices in R_k. Let d_k be the number of deaths at time z_k and $x_{ik}^* = \sum_{t_j = z_k} x_{ij}$ and $\bar{x}_{ik} = \sum_{j \in R_k} x_{ij} / m_k$. Then

$$r_i = \sum_{k=1}^{K} (x_{ij}^* - d_k \bar{x}_{ik})$$

and

$$s_i = \left[\sum_{k=1}^{K} (d_k / m_k) \sum_{j \in R} (x_{ij} - \bar{x}_{ik})^2 \right]^{\frac{1}{2}}.$$

s_0 is the median of all s_i.

5. Result

1. Data Structure

The data consist of 207 CELL files (207 patients). After RMA, we obtained for each patient 54,675 probes (data frame 54,675 rows and 207 columns). After filtering process, we ended up with 21,148 probes per patient. (Kang et al 23,775). The data structure is summarized in Table 1.
And the following figure describes the distribution of the filtered gene expression.

Table 1. Data strcture

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

2. Stratified Random Sampling

To partition the data into five groups to prepare for the cross validation and also preserve the key features, we adopted stratified random sampling. Based on literature review and our data structure, we decide to partition the sample data into 8 strata based on the combination of three key features: Age, WBC count at diagnosis and MRD (minimal residual disease) day 29. After adopting the algorithm proposed earlier, we get the five stratified samples with sample sizes as following: 41,41,41,42,42. We can tell from the result that the algorithm is effective in balancing sample size between stratified samples. And the corresponding training sets are 166, 166, 166, 165 and 165.

3. Calculate cox score and determine the candidate thresholds

We calculated the cox score for genes on each training set and get the candidate thresholds based on the cox-score distribution on each training set. Then we further select the genes with cox score

greater than the thresholds and keep them for approaching PCA. These genes are considered more associated with relapse free survival. The genes sorted by the cox-score value is saved as sorted_cox_score.csv. By checking the genes sorted by the cox-score, we spotted some genetic meaning behind it.

Biological aspect

Probe-ID	Gene	COX-Score	Function
221349_at	VPREB1	2.193905	-Immunoglobulin superfamily
215925_s_at	CD72	2.088864	-Chronic Lymphocytic Leukemia
1554733_at	LINC02363	2.082338	-Immunoglobulin superfamily
218829_s_at	CHD7	-2.05561	-CHARGE syndrome
205081_at	CRIP1	2.012747	-Intracellular zinc transport protein
226123_at	CHD7	-1.879853	-CHARGE syndrome
208302_at	HMHB1	1.866091	-cytotoxic T lymphocyte (CTL)
218469_at	GREM1	1.839573	-BMP (bone morphogenic protein) antagonist family.
227611_at	TARSL2	1.831222	- microbicidal activity of neutrophils
203949_at	MPO	-1.8129	

To accommodate the variation, we get the maximum cox-score and five minimum cox-score from each training set. The upper bound 2.178759627 is the minimum among the five maximum cox scores, and the lower bound 0.10893798 is the maximum value from the five minimum cox-scores Candidate thresholds are values evenly distributed between the lower and upper bound. The candidate thresholds are summarized in Table 2.

Table 2. Candidate thresholds selected.

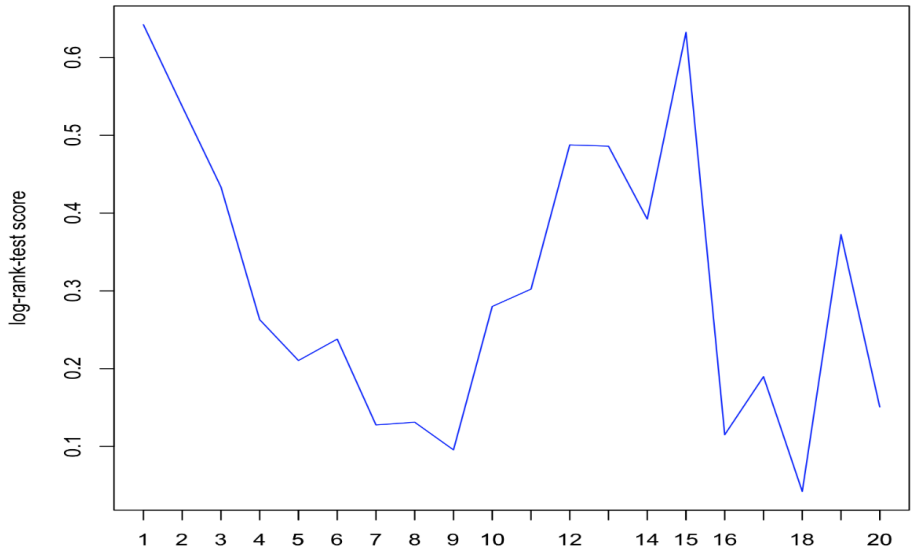
Candidate thresholds	# of genes
0.10893798135046058, 0.21787596270092116, 0.32681394405138176, 0.43575192540184232, 0.54468990675230289, 0.65362788810276351, 0.76256586945322402, 0.87150385080368464, 0.98044183215414527, 1.0893798135046058, 1.1983177948550663, 1.307255776205527, 1.4161937575559875, 1.525131738906448, 1.6340697202569088, 1.7430077016073693, 1.8519456829578298, 1.9608836643082905, 2.069821645658751, 2.1787596270092116	16885 12930 9522 6799 4668 3140 2052 1260 753 446 267 152 85 48 30 19 15 5 3 2

4. **Perform PCA and build cox proportional hazard model.**

We computed the first principal component score w for each patient in the test dataset, which is based on the PCA model in the train dataset. We get the survival data by combine event_free_suvival_time, relapse, w together. In total we have 20(#thresholds) *5(#cv_folds) = 100 files for building cox proportional hazard model in R by using coxph() packages.

5. **Calculate geometric mean the LRT score for each threshold and determine the final model.**

We calculated the geometric mean of log rank test score in each test dataset and then we get the averaged log rank test score for each threshold. We choose the model with the threshold that maximize the mean LRT score as our final model. The relationship between LRT score and candidate thresholds is summarized in Figure 3 below.



. Figure 3. Likelihood function as a function of candidate thresholds

This figure explains the relationship between log rank test score and candidate thresholds. The vertical axis is value of log rank test score and horizontal axis labels the twenty candidate thresholds in Table 2 in resending order. We can tell from the figure that the LRT score reaches its peak at the 15th candidate threshold, which corresponds to value 1.634, and the corresponding number of probe-sets is 30 (highlighted in Table 2). As a result, the final model chosen is the one with threshold value 1.634 and 30 probe-sets.

6. Fit cox proportional hazard model

Based on the 30 probe-sets selected, PCA is conducted on the whole dataset, and the First Principle Component Score (w) is found for each patient. w – a linear combination of the important probe-sets, can be interpreted a covariate to the relapse rate. The cox-proportional hazard model concerning relapse as event and w as covariate is fitted, result shows w is significant for predicting the relapse rate(see Figure 4). Under the assumption of cox model, the exponential value of coefficient stands for the on unit change in w, 1.1852 unit increase in the hazard of relapse. In other words, w is concordance with relapse rate- a big value w of indicates a high risk of relapse.

```
n=207, number of events=75

    coef  exp(coef)  se(coef)      z      p  lower 0.95  upper 0.95
w 0.1699     1.1852    0.0414 4.1026 0.0000      0.0887      0.2511  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance = 0.622
```

Figure 4. Output of Cox Proportional Hazard model

7. Classify high and low risk group based on Risk Factor

To define the risk group, we introduce Risk Factor, which is w multiply its coefficient in the cox model. Zero value of Risk Factor stands for baseline hazard, positive Risk Factor indicates enlarged relapse chances compared with the baseline hazard, and negative Risk Factor means reduced hazard. Among the 207 patients, 105 were classified in high-risk group, with 50% chance of relapse, while the low-risk group has only 25% relapse rate(see Figure 5).

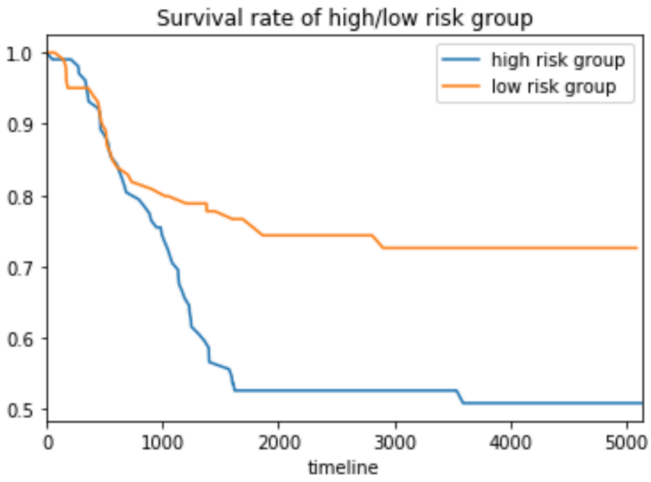


Figure 5. Relapse rate based on classified groups

Conclusions

We applied supervised PCA successfully distinguished the high/low risk group based on the Relapse Free Survival Rate. By looking at the value of 30 probe-sets, we can predict the risk of a new patient, which is pretty convenient.
However, we should also be cautious about this convenience. Theoretical statistical inference should be set up before using the 30 probe- sets as predictors in reality. In addition, the method is very data driven. The dataset we used is collected up to date July 2016, while Kang's paper is based on the data around 2008, applying the same method, we got quite different results.

To evaluate the method, we need always keep in mind to check the biological meaning behind the result, and try to interpret the code of life behind the data. Working with Affy-array is a great way to see gene expression because Affy-array have 36,000 transcripts and variants per sample. This allows us to see the key genes that are acting during the phase of the disease. Also, Affy-arrays get accurate and reproducible gene expression data by using multiple independent measurements for each transcript. With Affy-arrays we saw that the top genes playing a major role in our prediction model are genes that highly tied to immune function in the body. All these genes release immune kinase enzymes that modifies other proteins by chemically adding phosphate groups to them (phosphorylation) when the body starts to become ill. Phosphorylation usually results in a functional change of the target protein (substrate) by changing enzyme activity, cellular location, or association with other proteins. These results are in line with our prediction model because if the patient relapses, we see the high levels of

these kinases being release. If the patient does not relapse, than there are no levels of activity of these kinases.

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.
24. *Childhood Acute Lymphoblastic Leukemia Treatment (PDQ®)—Health Professional Version - National Cancer Institute*. 2017 2017-10-25]; Available from: https://www.cancer.gov/types/leukemia/hp/child-all-treatment-pdq - link/_580_toc.