Development of Molecular Profiles to Predict Treatment Outcomes in Lymphoma Patients

JOE MOEN
ELIZABETH WOLF
SARA BURNS

MENTOR = “DR. BRIAN SMITH”
Outline

- Background information
- Introduce data set
- Univariate screening
- Clustering
- Dimension reduction (scoring)
- Multivariate Cox regression model
- Conclusion
What is Lymphoma?

- Lymphoma is a type of cancerous cell that develops in the immune system
- 5th most common cancer in North America
- Treatments:
  - chemotherapy
  - radiotherapy
  - bone marrow transplantation
### What is Lymphoma?

#### Stage Distribution and 5-year Relative Survival by Stage at Diagnosis for 2002-2008, All Races, Both Sexes (NCI)

<table>
<thead>
<tr>
<th>Stage at Diagnosis</th>
<th>Stage Distribution (%)</th>
<th>5-year Relative Survival (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized (confined to primary site)</td>
<td>27</td>
<td>82.0</td>
</tr>
<tr>
<td>Regional (spread to regional lymphnodes)</td>
<td>19</td>
<td>77.8</td>
</tr>
<tr>
<td>Distant (cancer has metastasized)</td>
<td>47</td>
<td>61.7</td>
</tr>
<tr>
<td>Unknown (unstaged)</td>
<td>8</td>
<td>66.5</td>
</tr>
</tbody>
</table>
Previous Study

- **G. LENZ STUDY ON DIFFUSE LARGE B CELL LYMPHOMA**

- **A PREVIOUS STUDY IN 2008 BY THE NCI WAS PUBLISHED IN THE NEW ENGLAND JOURNAL OF MEDICINE**

- **IT MEASURED THE SURVIVAL RATES OF LYMPHOMA PATIENTS**

- **THE TWO TREATMENT GROUPS WERE R-CHOP AND CHOP**
Previous Study

- **OBJECTIVE**: Predict survival as a function of gene expression variables
- **OUTCOME** = Time to death
- **PREDICTORS** = Gene expression level obtained by microarray testing
Microarray Testing

Most Affordable and Commonly Used Form of Testing Gene Expression

- Results are quantitative
- 54,000 numeric variables
Summary Statistics of R-CHOP Data

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>60.13</td>
<td>61</td>
</tr>
<tr>
<td>Status</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>Follow up time</td>
<td>2.14</td>
<td>2.41</td>
</tr>
</tbody>
</table>

This is a retrospective study, therefore our data is censored (time to death is not always measurable)
Our Study

Start:
- 54,000 genes
- 412 patients

Reduction:
- 232 R-CHOP

Screening:
- 178 genes
Our Study

- Clustering: 7 clusters
- Scoring: 7 clusters
- Cox Regression: 2 clusters
DISMISSED 180 CHOP PATIENTS

- Analysis is performed in R-Studio
- Parsed through full dataset
- Created new matrix that contains only patients treated with R-CHOP
- New matrix contained 232 patients and 54,000 genes
- We focused on the R-CHOP data because it is the newest and most effective form of chemotherapy treatment for lymphoma
Univariate Screening

ASSOCIATION BETWEEN GENE AND OVERALL SURVIVAL

For a given gene, and a randomly selected subject i and j:

\[ H_0: \Pr(g_i > g_j \mid t_i < t_j) = 0.5 \]
\[ H_A: \Pr(g_i > g_j \mid t_i < t_j) \neq 0.5 \]

Where \( g_i \) and \( g_j \) are the gene expressions for a randomly selected subject i and j; and \( t_i \) and \( t_j \) are their time to death.

```r
install.packages("Hmisc")
library(Hmisc)
genes = t(exprs(etrain))
rcorrcens1 <- function(e) {
  t = e$futime
d = e$fustatus
  p = rrow(e)
c = rep(NA, p)
pvalue = rep(NA, p)
  for(j in 1:p) {
    x = genes[,j]
r = rcorrcens(Surv(t, d) ~ x)
c[j] = r["x","C"]
pvalue[j] = r["x","P"]
  }
  list(cvalues = c, pvalues = pvalue)
  #list(pval = 1 - pchisq(x2, 1), hr = exp(beta), betas = sqrt(hr))
}
```

\[ r = rcorrcens1(etrain) \]
\[ pvals = r$pvalues \]
Univariate Screening

- Measure of association between time of death and level of gene expression.
  - 0.5 indicates no association

- Calculated p-values to test the hypothesis based on the C-index of Harrell

- P-values were converted to Q-values

- Genes selected to maintain 10% FDR
False Discovery Rate

- False discovery rate: among those selected, the average number of genes thought to be significant that proved NOT to be significant.
- FDR =10%: manageable number and benchmark for FDR.
- Lenz study used p values instead of using a false discovery rate to identify significant genes in the screening process.
K-Means Clustering

- The partitioning of genes into groups with similar expression levels
- $K$- indicates the number of clusters into which the genes are partitioned
- Squared Euclidean distance
- $C(i)$- represents the cluster assignment for cluster $i$ estimated by the algorithm
- $x_i$- represents the set of expression values for expression $i$
- $\mu_k$- mean of cluster $k$

$$WSS = \sum_{k=1}^{K} \sum_{C(i)=k} ||x_i - \mu_k||.$$
<table>
<thead>
<tr>
<th>Cluster 1: 8 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;1552531_a_at&quot; &quot;1553499_s_at&quot; &quot;203434_s_at&quot; &quot;203435_s_at&quot; &quot;206310_at&quot; &quot;219874_at&quot;</td>
</tr>
<tr>
<td>&quot;231887_s_at&quot; &quot;244467_at&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2: 34 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;1555275_a_at&quot; &quot;1560397_s_at&quot; &quot;201866_s_at&quot; &quot;202022_at&quot; &quot;202172_at&quot; &quot;202740_at&quot; &quot;203285_s_at&quot;</td>
</tr>
<tr>
<td>&quot;203524_s_at&quot; &quot;203633_at&quot; &quot;203723_at&quot; &quot;204012_s_at&quot; &quot;204866_at&quot; &quot;206003_at&quot; &quot;206181_at&quot;</td>
</tr>
<tr>
<td>&quot;208456_s_at&quot; &quot;209621_s_at&quot; &quot;209825_s_at&quot; &quot;210461_s_at&quot; &quot;212133_at&quot; &quot;213534_s_at&quot; &quot;218324_s_at&quot;</td>
</tr>
<tr>
<td>&quot;221036_s_at&quot; &quot;221912_s_at&quot; &quot;222482_at&quot;</td>
</tr>
<tr>
<td>&quot;223159_s_at&quot; &quot;225207_at&quot; &quot;226930_at&quot; &quot;227220_at&quot; &quot;227684_at&quot; &quot;227904_at&quot; &quot;230509_at&quot;</td>
</tr>
<tr>
<td>&quot;235213_at&quot; &quot;235692_at&quot; &quot;235743_at&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 3: 35 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;1554306_at&quot; &quot;1559867_at&quot; &quot;1568600_at&quot; &quot;1570156_s_at&quot; &quot;202751_at&quot; &quot;203516_at&quot; &quot;203634_s_at&quot;</td>
</tr>
<tr>
<td>&quot;204530_s_at&quot; &quot;204584_at&quot; &quot;206653_at&quot; &quot;206698_at&quot; &quot;206756_at&quot; &quot;207949_s_at&quot; &quot;207954_at&quot;</td>
</tr>
<tr>
<td>&quot;209938_at&quot; &quot;213116_at&quot; &quot;215011_at&quot; &quot;218296_x_at&quot; &quot;219101_x_at&quot; &quot;219232_s_at&quot; &quot;219241_x_at&quot;</td>
</tr>
<tr>
<td>&quot;219420_at&quot; &quot;219815_s_at&quot; &quot;224357_s_at&quot;</td>
</tr>
<tr>
<td>&quot;227055_at&quot; &quot;228000_at&quot; &quot;228977_at&quot; &quot;229849_at&quot; &quot;230640_at&quot; &quot;230888_at&quot; &quot;239427_at&quot;</td>
</tr>
<tr>
<td>&quot;239973_at&quot; &quot;240616_at&quot; &quot;241599_at&quot; &quot;242240_at&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 4: 26 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;1553979_at&quot; &quot;200644_at&quot; &quot;200788_s_at&quot; &quot;201160_s_at&quot; &quot;201865_x_at&quot; &quot;203140_at&quot; &quot;204249_s_at&quot;</td>
</tr>
<tr>
<td>&quot;205668_at&quot; &quot;209306_s_at&quot; &quot;209337_at&quot; &quot;209397_at&quot; &quot;209924_at&quot; &quot;211275_s_at&quot; &quot;211671_s_at&quot;</td>
</tr>
<tr>
<td>&quot;212129_at&quot; &quot;212589_at&quot; &quot;212646_at&quot; &quot;213168_at&quot; &quot;213708_s_at&quot; &quot;216321_s_at&quot; &quot;218331_at&quot;</td>
</tr>
<tr>
<td>&quot;225331_at&quot; &quot;226496_at&quot; &quot;228167_at&quot; &quot;228812_at&quot; &quot;32128_at&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 5: 27 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;1555209_at&quot; &quot;1555729_a_at&quot; &quot;1563621_at&quot; &quot;205450_at&quot; &quot;205960_at&quot; &quot;209840_s_at&quot; &quot;210192_at&quot;</td>
</tr>
<tr>
<td>&quot;210330_at&quot; &quot;210688_s_at&quot; &quot;214071_at&quot; &quot;214597_at&quot; &quot;215056_at&quot; &quot;215784_at&quot; &quot;215828_at&quot;</td>
</tr>
<tr>
<td>&quot;217455_s_at&quot; &quot;219491_at&quot; &quot;224417_at&quot; &quot;231049_at&quot; &quot;232664_at&quot; &quot;233310_at&quot; &quot;233458_at&quot;</td>
</tr>
<tr>
<td>&quot;234284_at&quot; &quot;236231_at&quot; &quot;240921_at&quot;</td>
</tr>
<tr>
<td>&quot;241453_at&quot; &quot;242934_at&quot; &quot;244367_at&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 6: 17 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;204428_s_at&quot; &quot;211870_s_at&quot; &quot;213544_at&quot; &quot;216617_s_at&quot; &quot;220983_s_at&quot; &quot;229276_at&quot; &quot;229361_at&quot;</td>
</tr>
<tr>
<td>&quot;231367_s_at&quot; &quot;231391_at&quot; &quot;232534_at&quot; &quot;234871_at&quot; &quot;237241_at&quot; &quot;238232_at&quot; &quot;243392_at&quot;</td>
</tr>
<tr>
<td>&quot;243733_at&quot; &quot;243762_at&quot; &quot;243905_at&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 7: 31 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;1555728_a_at&quot; &quot;201161_s_at&quot; &quot;201512_s_at&quot; &quot;201554_x_at&quot; &quot;202020_s_at&quot; &quot;202171_at&quot;</td>
</tr>
<tr>
<td>&quot;203645_s_at&quot; &quot;205255_x_at&quot; &quot;209100_at&quot; &quot;212685_s_at&quot; &quot;213106_at&quot; &quot;213189_at&quot; &quot;213327_s_at&quot;</td>
</tr>
<tr>
<td>&quot;215049_x_at&quot; &quot;216945_x_at&quot; &quot;218134_s_at&quot; &quot;218862_at&quot; &quot;219061_s_at&quot; &quot;219607_s_at&quot; &quot;221675_s_at&quot;</td>
</tr>
<tr>
<td>&quot;222592_s_at&quot; &quot;222593_s_at&quot; &quot;223158_s_at&quot; &quot;223414_s_at&quot; &quot;224523_s_at&quot; &quot;225537_at&quot; &quot;226001_at&quot;</td>
</tr>
</tbody>
</table>
| "226426_at" "226452_at" "226874_at" "229594_at"
Dimension Reduction (Scoring)

- **PRINCIPLE COMPONENTS OR “EIGENGENES”**

- **DIMENSION REDUCTION AIMS TO CREATE ONE SCORE FOR EACH CLUSTER**

- **FORMS A LINEAR COMBINATION OF CLUSTER GENES**
Multivariate Cox Regression

\[ \lambda(t,x) = \lambda_0(t) \exp \{\beta_1 X_1 + ... + \beta_p X_p\} \]

- \( \lambda \)- represents death rate (over time)
- \( \lambda_0(t) \) -represents base line of death rate
- \( \beta \)-represent estimated variable effect
- Rule of thumb: no more than one variable for every ten events (n/10)
- Outcome: time to event
- Backward variable selection
  - Variables are taken out one by one starting with the least significant
Final Model Estimates

<table>
<thead>
<tr>
<th>Eigengene</th>
<th>$\beta$</th>
<th>SE ($\beta$)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2</td>
<td>0.116</td>
<td>0.040</td>
<td>0.0041</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>-0.209</td>
<td>0.061</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

**Goodness of Fit:**

$R^2=0.24$ (modest)

C-Index=0.79 (good)
Fitted Model Profile

\[ \lambda(t,x) = \lambda_0(t) \exp\{0.116Y_2 - 0.209Y_4\} \]

Gene Profile = 0.116Y_2 - 0.209Y_4
ROC Analysis

- Assess performance of gene profile in predicting survival
- Illustrates sensitivity vs specificity over the range of possible cut off values for the gene profiling score
- AUC - Area Under the ROC Curve
  - 1 = Perfect prediction
  - .5 = No predictive ability
- All of the possible cut off values for having a positive or negative test
ROC Curves

6-Month Survival

AUC = 0.844
ROC Curves

One Year Survival

Specificity

1 - Sensitivity

AUC = 0.806
ROC Curves

Five Year Survival

AUC = 0.778
AUC Curve By Year
Conclusion

**Future Studies:** apply our model to independent data sets (typically, these models work best with the population on which they are built)

**Questions?**
References

1. FRANK E. HARRELL R-STUDIO PACKAGES: ‘SURVAUC’ AND ‘RCORR.CENS {HMISC}’
2. NATIONAL CANCER INSTITUTE