Development of Molecular Profiles to Predict Treatment Outcomes in Lymphoma Patients

JOE MOEN
ELIZABETH WOLF
SARA BURNS

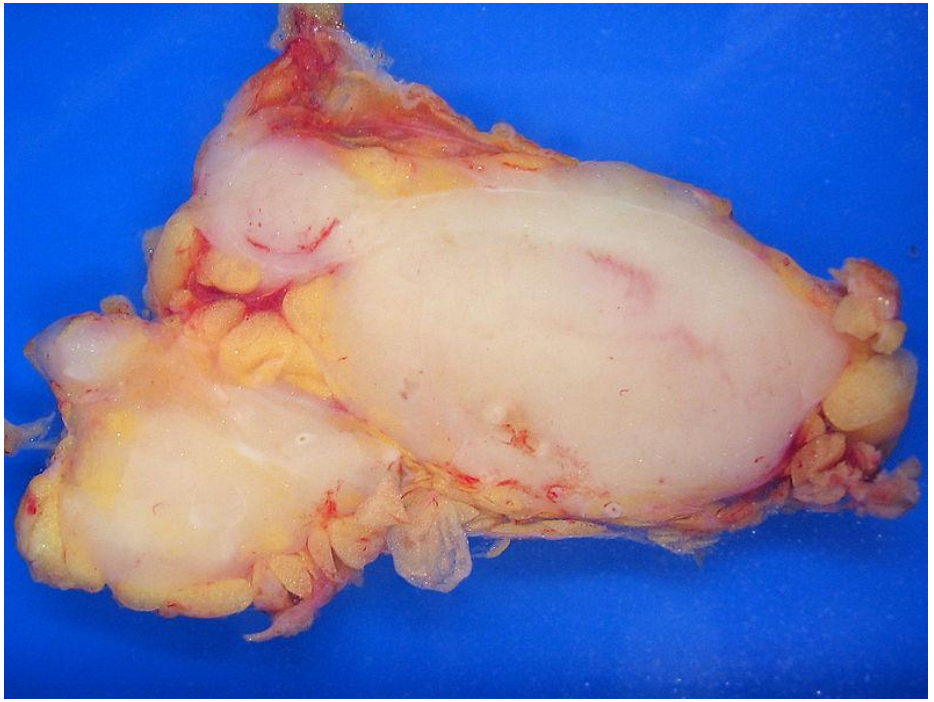
MENTOR== "DR. BRIAN SMITH"

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)

Stage at Diagnosis	Stage Distribution (%)	5-year Relative Survival (%)
Localized (confined to primary site)	27	82.0
Regional (spread to regional lymphnodes)	19	77.8
Distant (cancer has metastasized)	47	61.7
Unknown (unstaged)	8	66.5

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

	Mean	Median
Age	60.13	61
Status	0.25	0
Follow up time	2.14	2.41

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

Reduction

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_O: $Pr(g_i > g_j \mid t_i < t_j) = 0.5$
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.

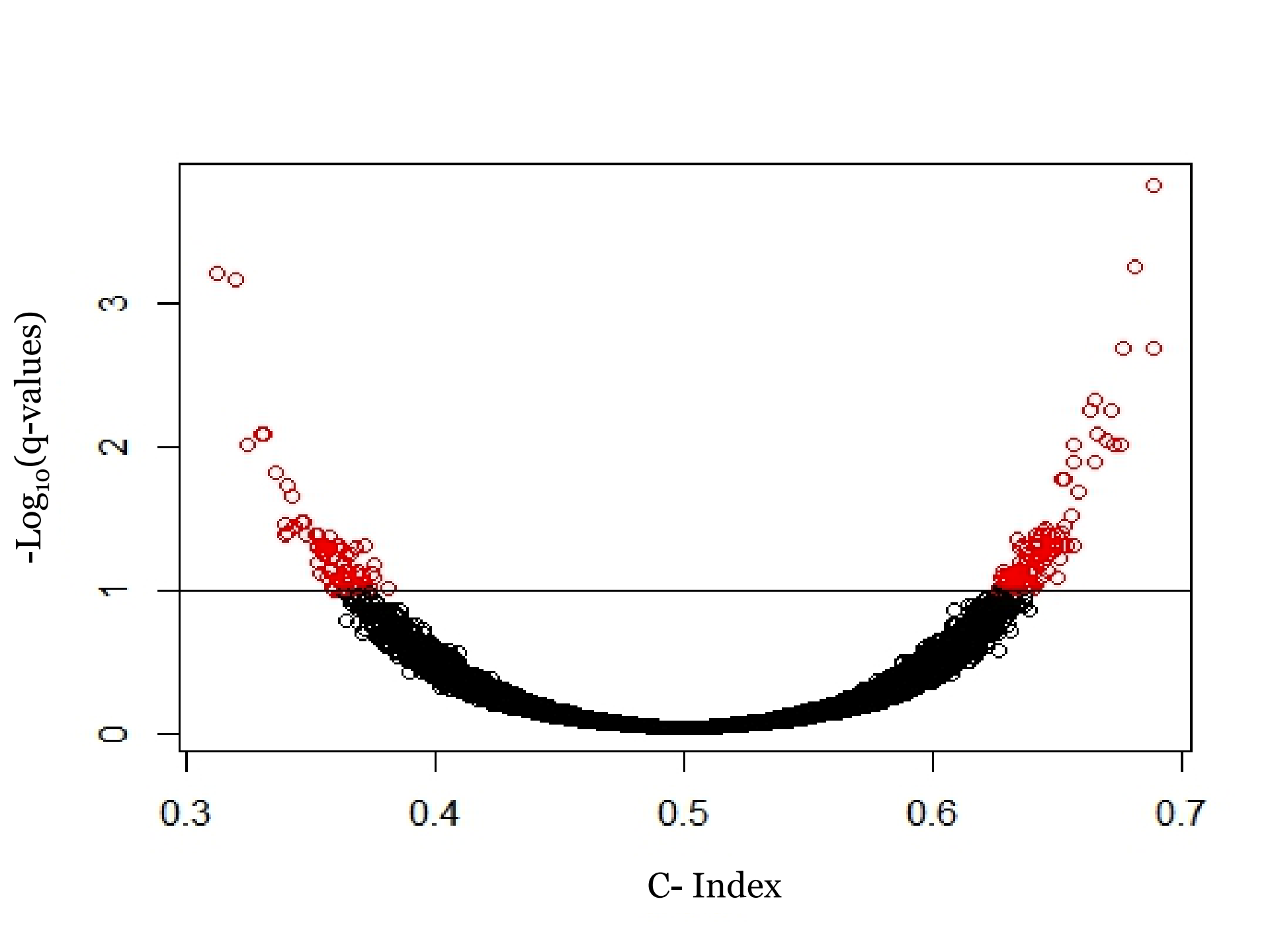
```
install.packages("Hmisc")
library(Hmisc)
genes = t(exprs(etrain))
rcorrcens1 <- function(e) {
  t = e$futime
  d = e$fustatus

  p = nrow(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

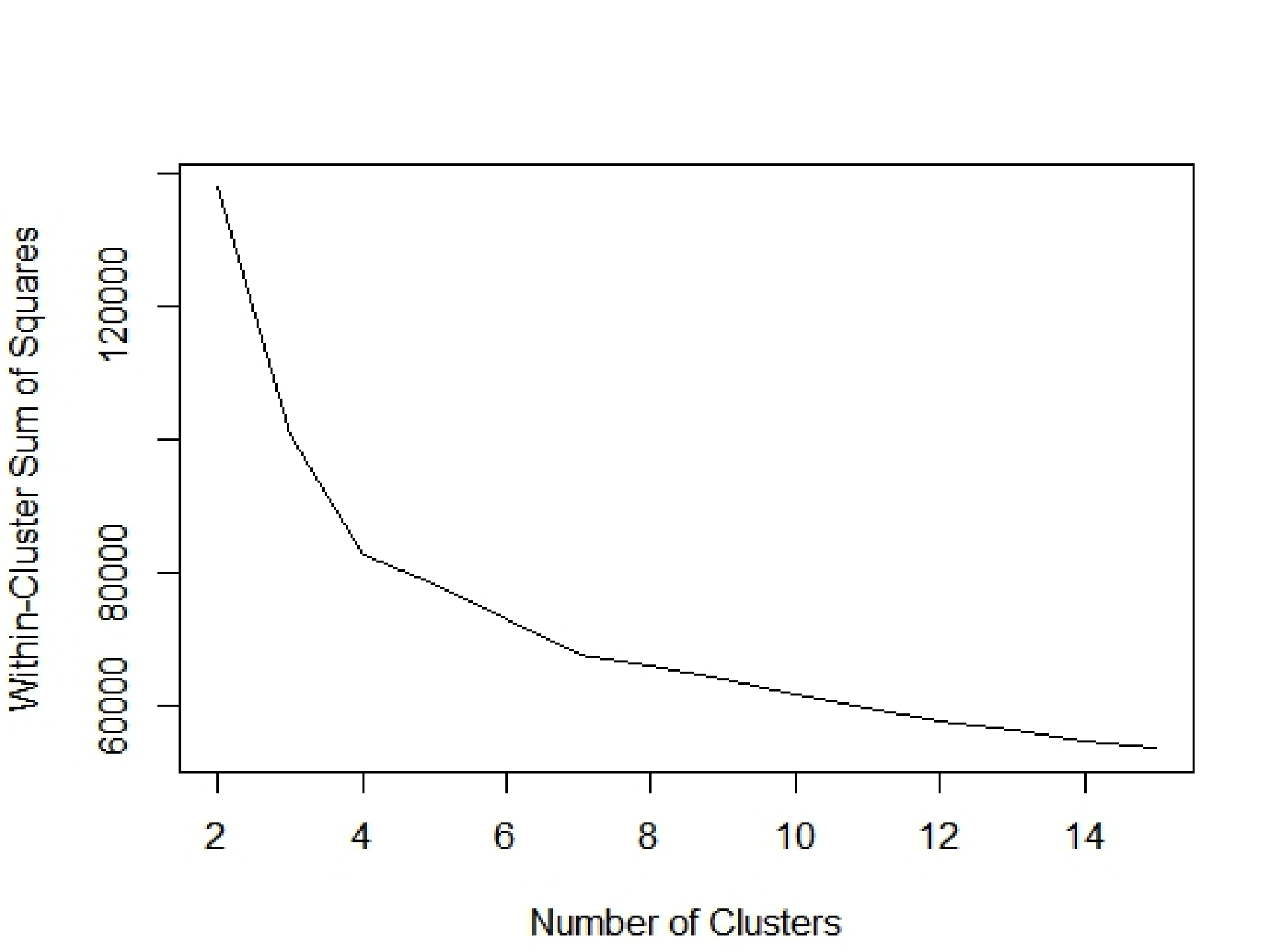
- False discovery rate : among those selected, the average number of genes thought to be significant that proved NOT to be significant
- FDR =10%- manageable number and benchmark FDR
- Lenz study used p values instead of using a false discovery rate to identify significant genes in the screening process


-Log10(q-values)
0
1
2
3
0.3
0.4
0.5
0.6
0.7
C- Index

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
- ^ μ_k- mean of cluster k

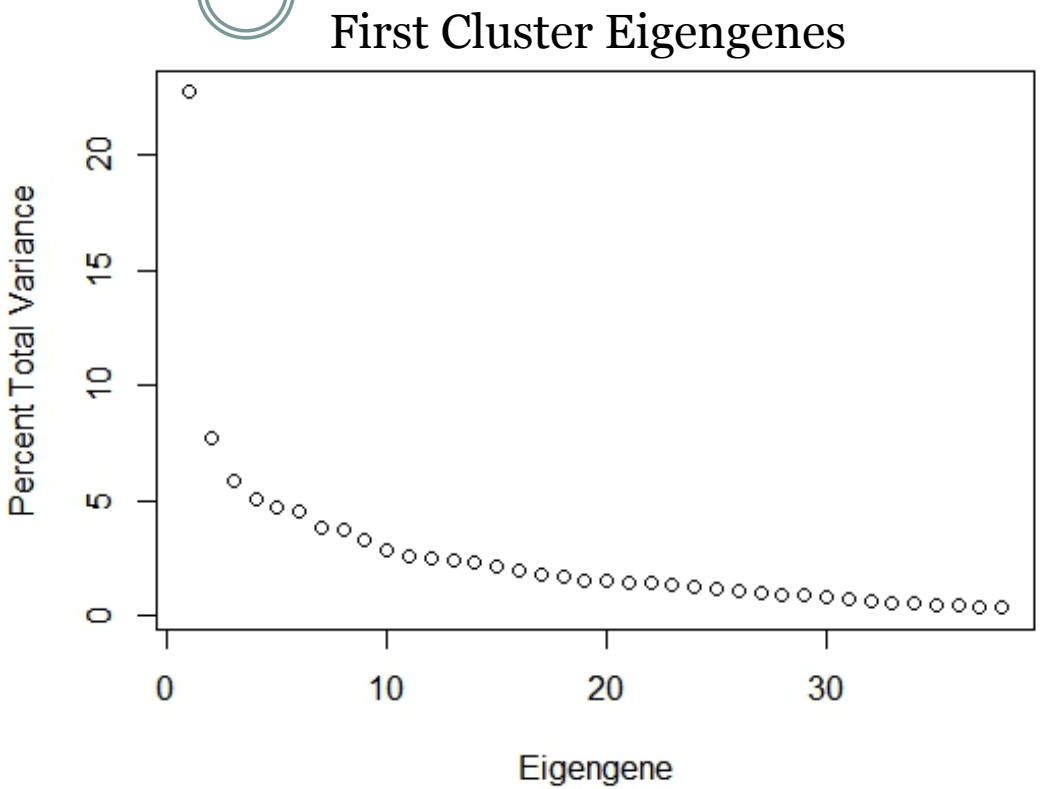
$$WSS = \sum_{k=1}^{K} \sum_{C(i)=k} ||x_i - \hat{\mu}_k||.$$


Within-Cluster Sum of Squares
60000
80000
120000
2
4
6
8
10
12
14
Number of Clusters

Cluster 1: 8 genes
"1552531_a_at" "1553499_s_at" "203434_s_at" "203435_s_at" "206310_at" "219874_at" "231887_s_at" "244467_at"
Cluster 2: 34 genes
"1555275_a_at" "1560397_s_at" "201866_s_at" "202022_at" "202172_at" "202740_at" "203285_s_at" "203524_s_at" "203633_at" "203723_at" "204012_s_at" "204866_at" "206003_at" "206181_at" "208456_s_at" "209621_s_at" "209825_s_at" "210461_s_at" "212133_at" "213534_s_at" "218324_s_at" "221036_s_at" "221912_s_at" "222482_at" "223159_s_at" "225207_at" "226930_at" "227220_at" "227684_at" "227904_at" "230509_at" "235213_at" "235692_at" "235743_at"
Cluster 3: 35 genes
"1554306_at" "1559867_at" "1568600_at" "1570156_s_at" "202751_at" "203516_at" "203634_s_at" "204530_s_at" "204584_at" "206653_at" "206698_at" "206756_at" "207949_s_at" "207954_at" "209938_at" "213116_at" "215011_at" "218296_x_at" "219101_x_at" "219232_s_at" "219241_x_at" "219420_s_at" "221845_s_at" "224357_s_at" "227055_at" "228000_at" "228977_at" "229849_at" "230640_at" "230888_at" "239427_at" "239973_at" "240616_at" "241599_at" "242240_at"
Cluster 4: 26 genes
"1553979_at" "200644_at" "200788_s_at" "201160_s_at" "201865_x_at" "203140_at" "204249_s_at" "205668_at" "209306_s_at" "209337_at" "209397_at" "209924_at" "211275_s_at" "211671_s_at" "212129_at" "212589_at" "212646_at" "213168_at" "213708_s_at" "216321_s_at" "218331_s_at" "225331_at" "226496_at" "228167_at" "228812_at" "32128_at"
Cluster 5: 27 genes
"1555209_at" "1555729_a_at" "1563621_at" "205450_at" "205960_at" "209840_s_at" "210192_at" "210330_at" "210688_s_at" "214071_at" "214597_at" "215056_at" "215784_at" "215828_at" "217455_s_at" "219491_at" "224417_at" "231049_at" "232664_at" "233310_at" "233458_at" "234284_at" "236231_at" "240921_at" "241453_at" "242934_at" "244367_at"
Cluster 6: 17 genes
"204428_s_at" "211870_s_at" "213544_at" "216617_s_at" "220983_s_at" "229276_at" "229361_at" "231367_s_at""231391_at" "232534_at" "234871_at" "237241_at" "238232_at" "243392_at" "243733_at" "243762_at" "243905_at"
Cluster 7: 31 genes
"1555728_a_at" "201161_s_at" "201512_s_at" "201554_x_at" "202020_s_at" "202171_at" "203645_s_at" "205255_x_at" "209100_at" "212685_s_at" "213106_at" "213189_at" "213327_s_at" "215049_x_at" "216945_x_at" "218134_s_at" "218862_at" "219061_s_at" "219607_s_at" "221675_s_at" "222592_s_at" "222593_s_at" "223158_s_at" "223414_s_at" "224523_s_at" "225537_at" "226001_at" "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

$$\lambda(t,x) = \lambda_0(t) \exp\{\beta_1 X_1 + \ldots + \beta_p X_p\}$$

- λ- represents death rate (over time)
- λ_0 (t) -represents base line of death rate
- β-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

Eigengene	β	SE (β)	p-value
Cluster 2	0.116	0.040	0.0041
Cluster 4	-0.209	0.061	0.0006

Goodness of Fit:
R^2=0.24 (modest)
C-Index=0.79 (good)

Fitted Model Profile

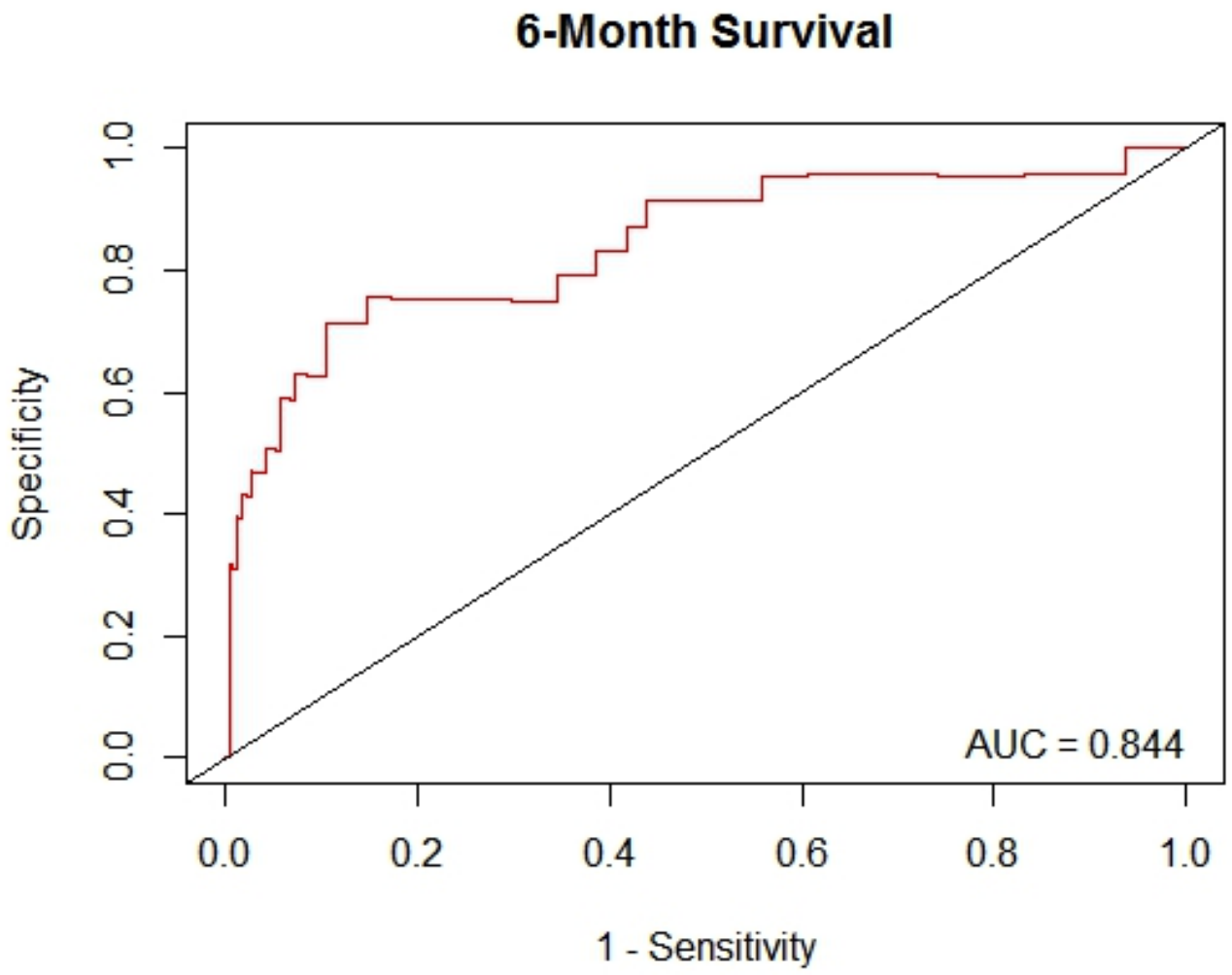
$$\lambda(t,x) = \lambda_0(t) \exp\{0.116Y_2 - 0.209Y_4\}$$

Gene Profile= $0.116Y_2 - 0.209Y_4$

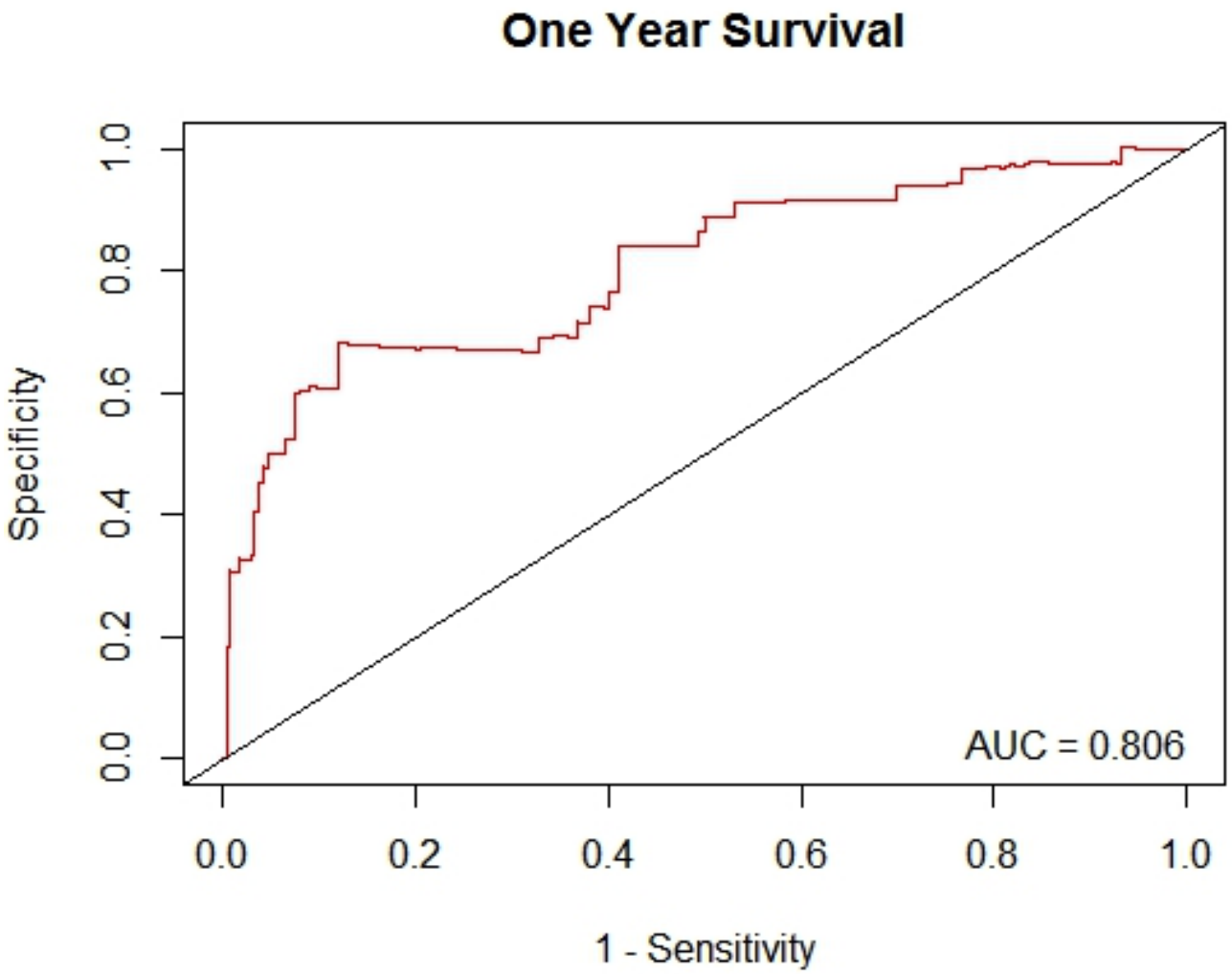
ROC Analysis

- 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

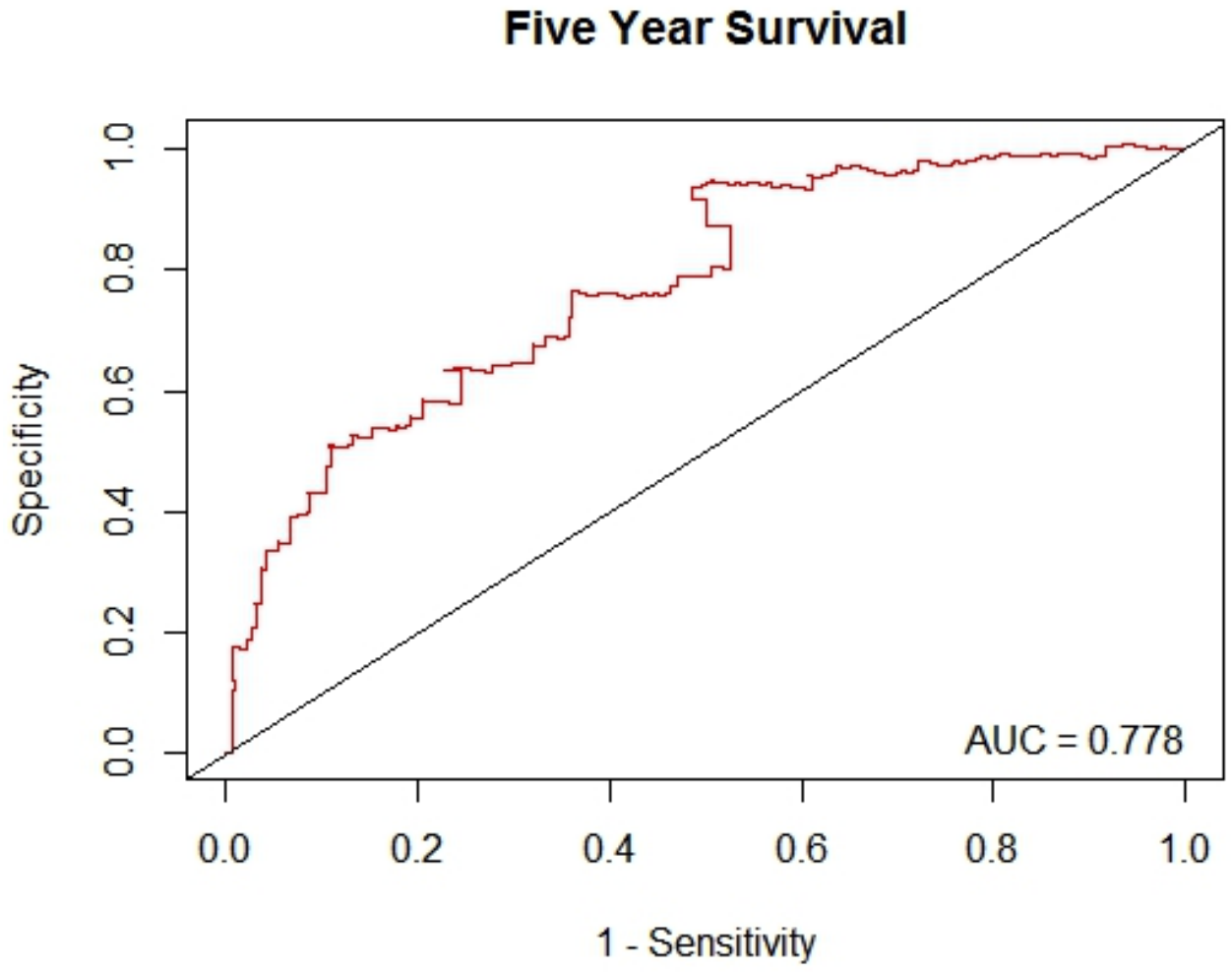
ROC Curves


6-Month Survival
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.844
0.0
0.2
0.4
0.6
0.8
1.0
1 - Sensitivity

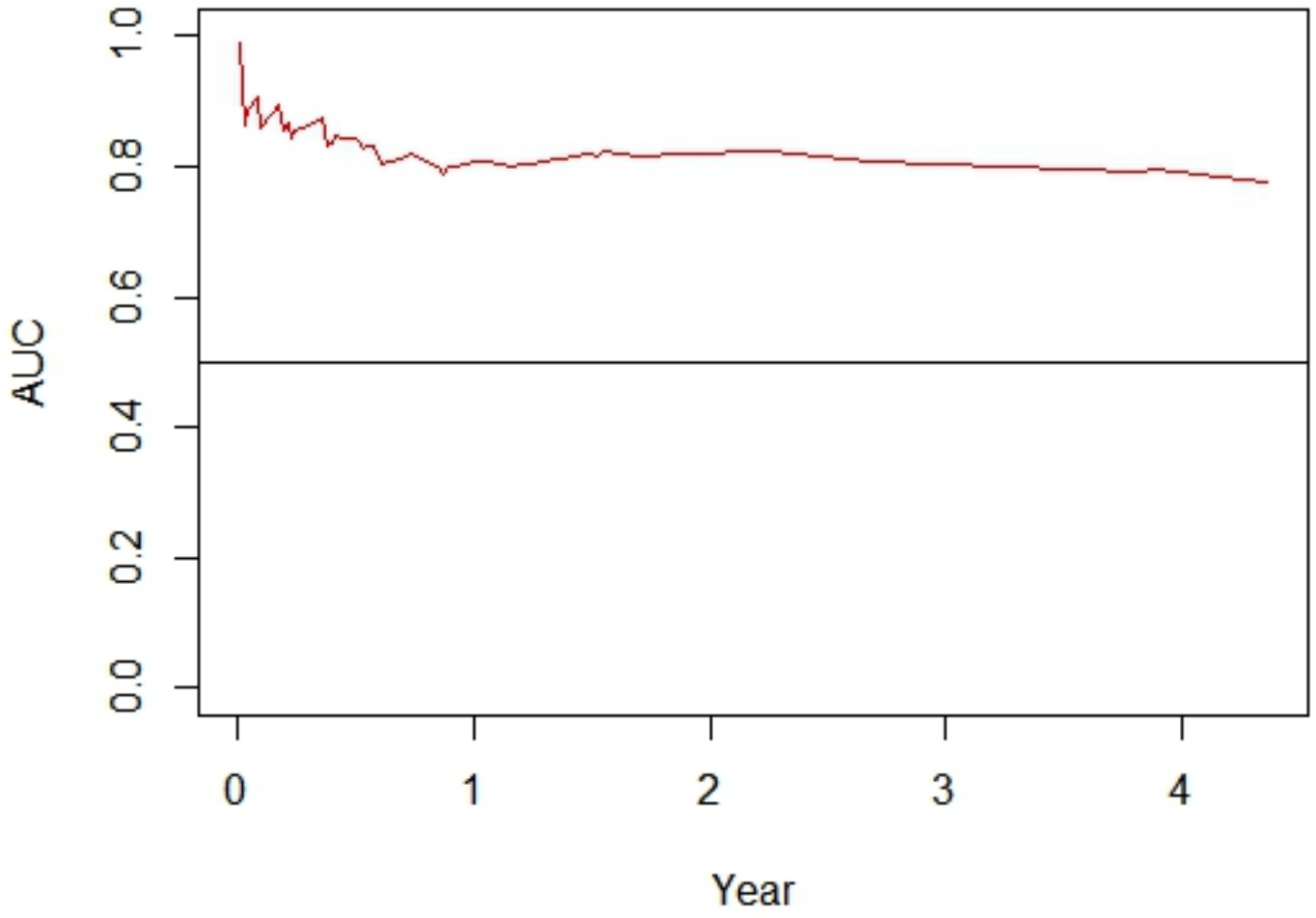
ROC Curves


One Year Survival
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.806
0.0
0.2
0.4
0.6
0.8
1.0
1 - Sensitivity

ROC Curves


Five Year Survival
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.778
0.0
0.2
0.4
0.6
0.8
1.0
1 - Sensitivity

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
   - http://seer.cancer.gov/statfacts/html/lymph.html