

Statistical Methods for Biomarker Discovery

An Application to Diagnosis and Prognosis of ARDS

Tatsuki Koyama

Division of Cancer Biostatistics
Department of Biostatistics, Vanderbilt University School of Medicine

Cancer Biostatistics Workshop
April 17th, 2009

Acknowledgement

Vanderbilt Clinical Proteomics Program

- Lorraine Ware
- Richard Fremont
- Addison May

Biostatistics

- Dean Billheimer
- William Wu

Acute Respiratory Distress Syndrome

- Respiratory failure
- Caused by sepsis, pneumonia, pulmonary aspiration, trauma
- May cause hypoxemia and multiple organ failure
- 200,000 / year in US
- 30% to 50% mortality
- Diagnosis is based on clinical definition
 - Bilateral infiltrates
 - $\text{PaO}_2 / \text{FiO}_2 < 300$ Acute Lung Injury (ALI)
 - $\text{PaO}_2 / \text{FiO}_2 < 200$ ARDS
 - No evidence of heart failure as cause

Bernard et al. *Am J Respir Crit Care Med* 1994; 149: 818-824

Biological markers in ALI/ARDS

- plasma / serum, pulmonary edema fluid, urine
- Markers of
 - inflammation (IL6, IL8)
 - oxidant stress
 - lung epithelial / endothelial injury (RAGE, SPD, CC16, VWF)
 - collagen deposition and altered coagulation and fibrinolysis (PCPIII)
- Rationale for studying a panel of biomarkers
 - No single marker has been sufficient to accurately differentiate patients with ALI/ARDS from those at risk.
 - ARDS is a complex syndrome characterized by injury to multiple cell types, inflammation and dysregulated coagulation and fibrinolysis

Study design : diagnostic panel of biomarkers

- Trauma prospective cohort study with 1,020 critically injured trauma patients
- ALI/ARDS incidence 20 to 30%
- Nested case control study:
 - 85 patients with clear chest X ray
 - 105 patients with ALI/ARDS
 - selection criteria include availability of blood sample / CXR within 72 hours of study enrollment
- Goal is to develop a model that distinguishes ALI/ARDS patients from control patients, and more importantly, choose useful biomarkers from a panel of 21.
- VALID study (Validation of Acute Lung Injury Markers for Diagnosis): prospective ICU study with 2550 patients.

Study design : prognosis in ALI/ARDS

- 528 patients from the ARDS clinical trials network multi-center trial (the ALVEOLI study)
- 8 biomarkers (selected based on prior demonstration of their association with adverse outcomes in ALVEOLI study)
- Goal is to determine the prognostic value of a panel of plasma markers for death in patients with ARDS

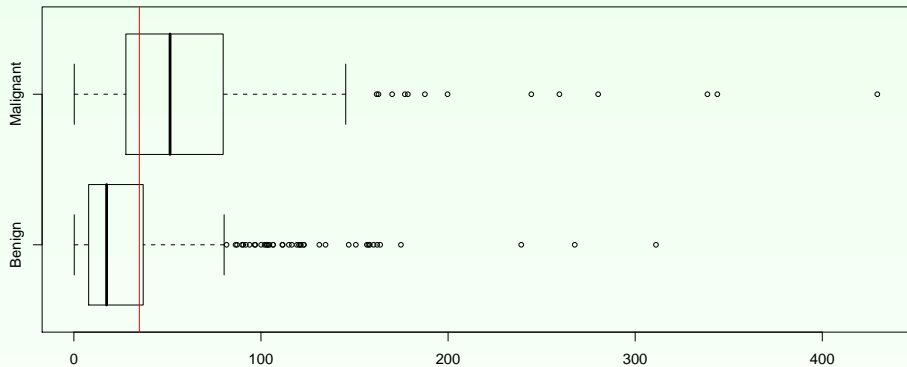
Sensitivity and Specificity

| | | TRUTH | |
|----------------------|---|-----------------------|-----------------------|
| | | + | - |
| TEST (Prediction) | + | a True Positive | b False Positive |
| | - | c False Negative | d True Negative |

$$\text{Sensitivity} = \frac{a}{a+c} = \text{P}[\text{Test } + \mid \text{Truth } +]$$

$$\text{Specificity} = \frac{d}{b+d} = \text{P}[\text{Test } - \mid \text{Truth } -]$$

Example: PCA3 and prostate cancer



Example: PCA3 and prostate cancer

Malignant group has elevated PCA3 levels. The cutoff of 35 was suggested.

| | Malignant | Benign |
|-----------|-----------|--------|
| PCA3 > 35 | 108 | 160 |
| PCA3 ≤ 35 | 45 | 423 |

$$\text{Sensitivity} = P[\text{Test} + \mid \text{Truth} +] = \frac{108}{108 + 45} = 0.71$$

$$\text{Specificity} = P[\text{Test} - \mid \text{Truth} -] = \frac{423}{160 + 423} = 0.73$$

Receiver Operating Characteristic

In the previous example, if the cutoff point is not 35, sensitivity and specificity would be different.

| | Malignant | Benign |
|-----------|-----------|--------|
| PCA3 > 10 | 136 | 405 |
| PCA3 ≤ 10 | 17 | 178 |

$$\text{Sensitivity} = 0.89$$

$$\text{Specificity} = 0.31$$

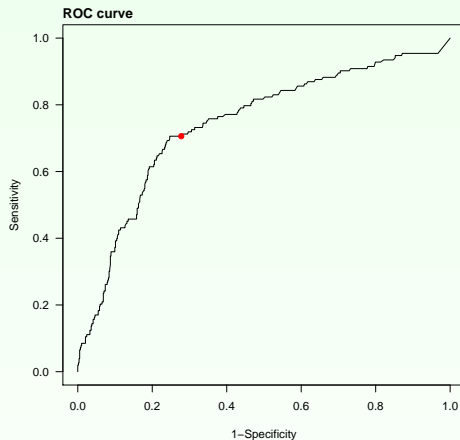
| | Malignant | Benign |
|------------|-----------|--------|
| PCA3 > 150 | 13 | 11 |
| PCA3 ≤ 150 | 140 | 572 |

$$\text{Sensitivity} = 0.08$$

$$\text{Specificity} = 0.98$$

PCA3 and prostate cancer: ROC curve

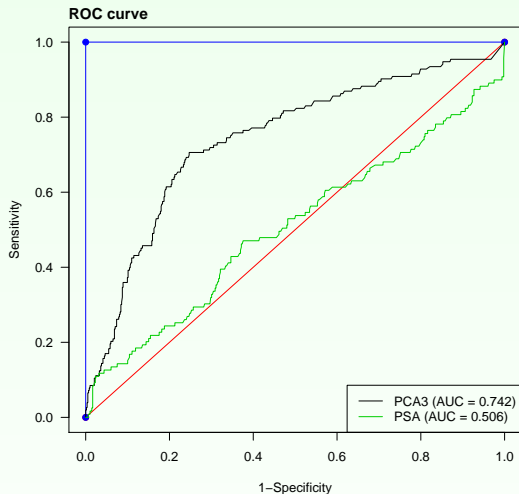
ROC curve is a plot of sensitivity and 1-specificity.



AUC: area under curve

- The plot of sensitivity and 1-specificity ($P[\text{false negative}]$) generally does not help you with choosing the optimal cutoff, but it can be used to assess the usefulness of the predictor.

ROC curve



- The blue line corresponds to a perfect prediction. ($AUC = 1$)
- The red line corresponds to the worst case (coin flip). ($AUC = 0.5$)
- In this example, AUC (aka C -statistic) of PCA3 is 0.742.

Interpreting *AUC*

- The average value of sensitivity for all possible values of specificity.
- The average value of specificity for all possible values of sensitivity.
- The probability that the randomly selected “condition +” patient has a test result indicating greater suspicion than a randomly selected “condition -” patient.
- There is a connection between *AUC* and a Wilcoxon 2-sample rank-sum test.

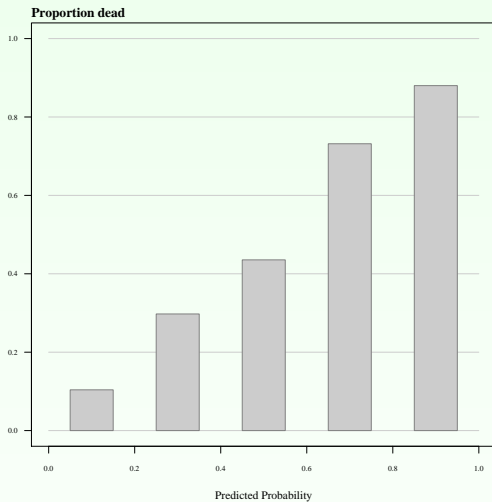
Discrimination and calibration

So far, we have looked at the model's ability to discriminate cases from controls.

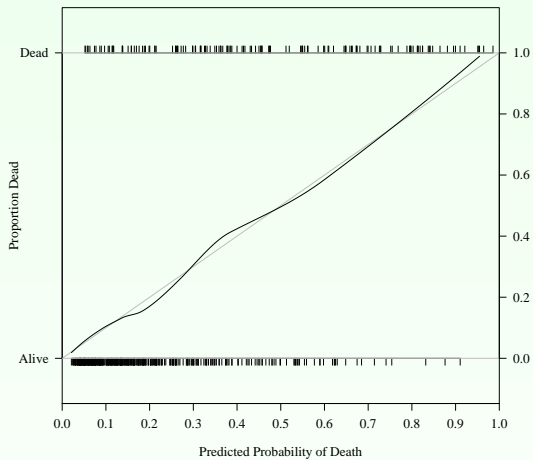
Usefulness of a model should also be judged on calibration.

- A model that assigns a predicted probability of 0.51 to all cases and 0.49 to all controls enjoys perfect discrimination, but it is not calibrated well.
- Predicted and actual probabilities are about the same in a well calibrated model. e.g., about 30% of people with 30% probability of death die.

Calibration



Calibration



Predicting death from ARDS

We would like to develop a model with a few predictors that predicts death from ARDS fairly well.

$N = 528$, (alive 384, dead 144)

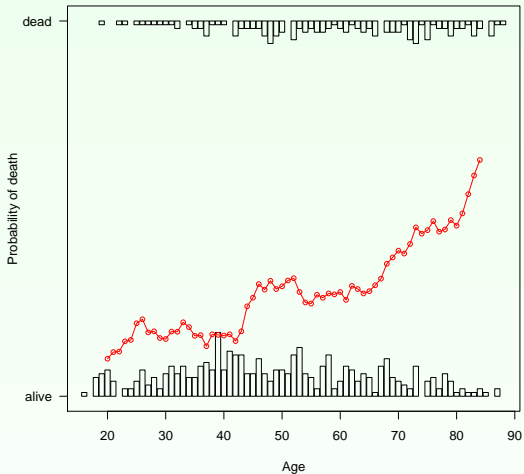
Consider the following models:

- Full model with 13 predictors
- Clinical variables only
- Biomarkers only
- “Best” model

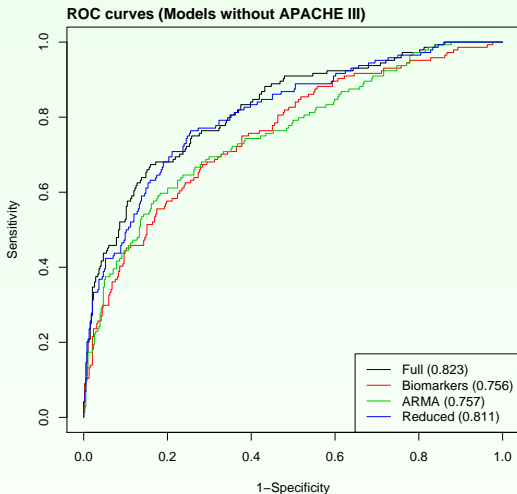
Some additional considerations:

- Interactions?
- Nonlinear effects?

Nonlinear effects?



ROC Curves from various models



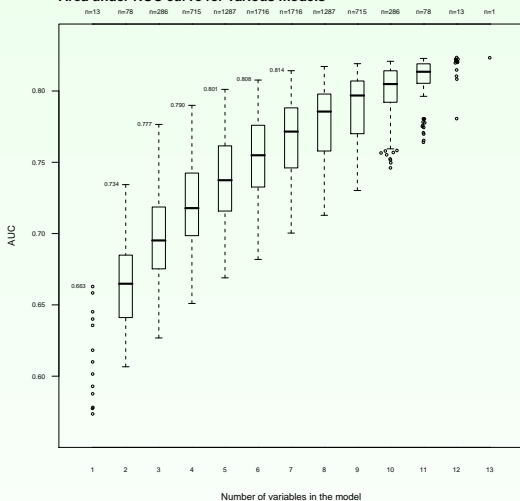
- Full model has 13 predictors.
- Biomarkers-only model has 8 predictors.
- Clinical-variables-only model has 5 predictors.
- Reduced model has 7 predictors (4 biomarkers).

How to reduce from the full model?

- Not all decisions were statistical!
(*"We want to use 5 biomarkers at most..."*)
- Univariable contribution suggested by backward elimination.
(Bootstrap?)
- Computed AUC for all possible combinations of predictors

$AUC \times 8,191$

Area under ROC curve for various models



- 1 Age
- 2 Age + IL8
- 3 Age + IL8 + SPD
- 4 Age + IL8 + SPD + Organ failures
- 5 Age + IL8 + SPD + o.f. + PAI1
- 6 Age + IL8 + SPD + o.f. + PAI1 + Alveolar-arterial O₂ difference
- 7 Age + IL8 + SPD + o.f. + PAI1 + AA + TNFR

So which model?

We would like to have a model that is as good as the full model with fewer predictors.

- Is the reduced model good enough?
- How can we compare the models?

Reclassification Table

- Cook NR, “Use and misuse of the receiver operating characteristic curve in risk prediction”. *Circulation*, 2007; 115: 928-935.

Reclassification table from Cook (2007)

Close

TABLE 3. Comparison of Observed and Predicted Risks Among Women in the Women's Health Study*

| Model Without HDL 10-Year Risk (%) | Model With HDL 10-Year Risk (%) | | | | % Reclassified |
|------------------------------------|---------------------------------|-----------|------------|------|----------------|
| | 0 to <5% | 5 to <10% | 10 to <20% | 20%+ | |
| 0% to <5% | | | | | |
| Total, n | 22655 | 696 | 6 | 0 | ... |
| %† | 97.0 | 3.0 | 0.0 | 0.0 | 3.0 |
| Observed 10-year risk (%)‡ | 1.5 | 5.9 | 0.0 | ... | ... |
| 5% to <10% | | | | | |
| Total, n | 593 | 1712 | 291 | 0 | ... |
| % | 22.8 | 66.0 | 11.2 | 0.0 | 34.0 |
| Observed 10-year risk (%) | 3.7 | 7.6 | 14.7 | ... | ... |
| 10% to <20% | | | | | |
| Total, n | 3 | 214 | 512 | 76 | ... |
| % | 0.4 | 26.6 | 63.6 | 9.4 | 36.4 |
| Observed 10-year risk (%) | 0.0 | 7.5 | 10.7 | 23.3 | ... |
| 20%+ | | | | | |
| Total, n | 0 | 0 | 41 | 102 | ... |
| % | 0.0 | 0.0 | 28.7 | 71.3 | 28.7 |
| Observed 10-year risk (%) | ... | ... | 15.8 | 32.5 | ... |

*This comparison uses models that include Framingham risk factors with and without HDL. All estimated and observed risks represent 10-year risk of cardiovascular disease.
 †Percent classified in each risk stratum by the model with HDL.
 ‡Observed proportion of participants developing cardiovascular disease in each category.

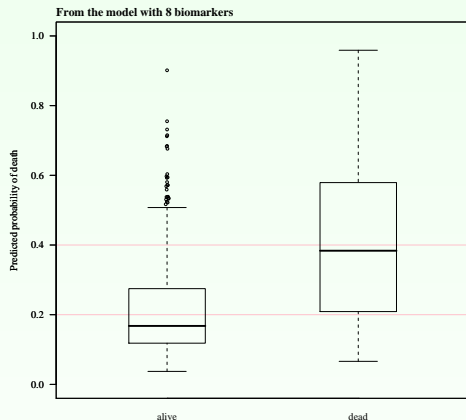
TABLE 3. Comparison of Observed and Predicted Risks Among Women in the Women's Health Study

Some improvements on reclassification table

- Put the predicted probabilities from Model 1 into a few categories like “low risk”, “medium risk”, “high risk”.
- Put the predicted probabilities from Model 2 into the same categories.
- Look at the cases that Model 2 reclassified into better categories (i.e., if the sample is from “dead” group, categorizing into “medium risk” is better than “low risk”; if the sample is from “alive” group, “medium risk” is better than “high risk”).
- Is Model 2 an improvement from Model 1?

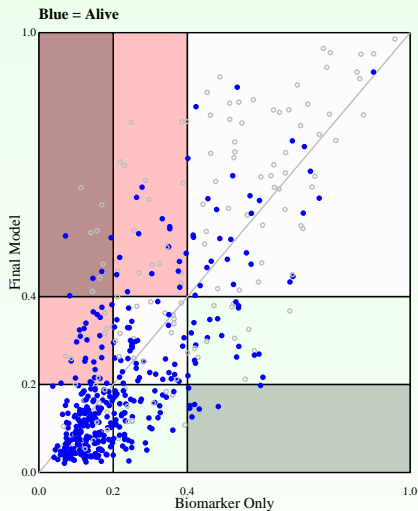
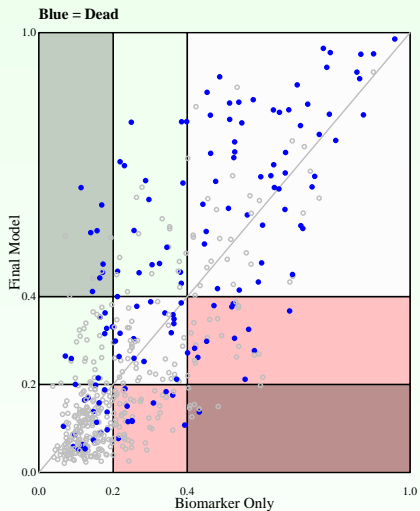
Biomarker-only model as an example

- With a logistic regression model, we compute the predicted probability of death for each patient.



- Cutoff = 0.2
Sensitivity = 0.76
Specificity = 0.58
- Cutoff = 0.4
Sensitivity = 0.47
Specificity = 0.86

Biomarker-only Model and Final Model



Problems with Reclassification Table

- Limited means of evaluating improvement in performance of the models
- Requires pre-set categories

Beyond Reclassification Tables

“Statisticians should seek new ways, beyond the ROC curve, to evaluate clinically useful new markers for their ability to improve upon current models ...” [Greenland and O’Malley 2005]

- People with and without events should be considered separately.
- Reclassification improvement
= Proportion of upward movement in event group
+ Proportion of downward movement in non-event group
- We should compute the reclassification improvement without preset categories.
- Just imagine every single person is in his/her own category. \Rightarrow Compare predicted probabilities of event.

Integrated Discrimination Improvement [Pencina *et al.* 2007]

- *IDI* is a measure of improvement of prediction that does not require pre-set categories.
 - Compute the differences (new - old) in predicted probabilities of the event (e.g., death) in the event group and average them. If the new model is better, this will be positive.
 - Compute the difference (new - old) in predicted probabilities of the event in the non-event group and average them. If the new model is better, this will be negative.
 - *IDI* is the differences of these two measures.

More on *IDI*

- *IDI* does not require cutoffs.
- *IDI* takes magnitudes (predicted probabilities of an event) into account.
- Asymptotic distribution of *IDI* is simple, and testing $H_0 : IDI = 0$ (no improvement) is simple when sample size is large.
- *IDI* seems much more sensitive than *AUC*.
 - Example in Pencina *et al.*
Conventional model for coronary heart disease risk includes sex, diabetes, smoking, age, systolic blood pressure.
New model includes HDL cholesterol.
 $N = 3264$
AUC increases from 0.762 to 0.774 (p -value = 0.092).
IDI estimate is 0.009 (p -value = 0.008)

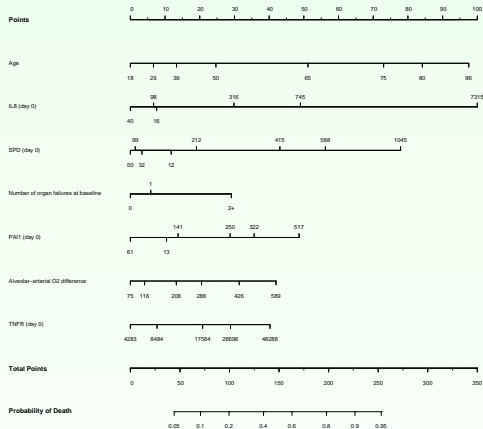
AUC and *IDI*

- Both *AUC* and *IDI* are a form of average sensitivity.
 - *AUC* weights more heavily sensitivity corresponding to small cutoffs (large sensitivity).
 - *IDI* uses a uniform weight.
 - Assuming no change in specificity and a uniform increase in sensitivity by 0.01 at every cutoff point, both *AUC* and *IDI* will go up by 0.01.
- Is increase of 0.01 in average sensitivity meaningful?

Choosing the model

- Choosing the final model often requires decisions not based on statistical significance. (Difference between two models may be statistically significant with *IDI* but not with *AUC*.)
- The model should have good predictive ability (good sensitivity and specificity → large *AUC*.)
- The model should be simple.

Nomogram from the final model



- Fremont, Koyama, Calfee, Wu, Dossett, Bossert, Mitchell, Wickersham, Bernard, Matthay, May, Ware. “Acute lung injury in patients with traumatic injuries: utility of a panel of biomarkers for diagnosis and pathogenesis” *under revision*
- Ware, Koyama, Billheimer, Wu, Bernard, Thompson, Brower, Standiford, Martin, Mattheay, NHLBI ARDS Clinical Trials Network. “Prognostic and pathogenetic value of combining clinical and biochemical indices in patients with acute lung injury” *under review*
- Blum, Koyama, M’Koma, Iturregui, Martinez-Ferrer, Uwamariya, Smith, Clark, Bhowmick. Chemokine markers predict biochemical recurrence of prostate cancer following prostatectomy. *Clinical Cancer Research*, 2008, **14**(23): 7790-7797.

Statistical Methods for Biomarker Discovery

An Application to Diagnosis and Prognosis of ARDS

Tatsuki Koyama

Division of Cancer Biostatistics
Department of Biostatistics, Vanderbilt University School of Medicine

Cancer Biostatistics Workshop
April 17th, 2009