Part [1.1] – Measures of Classification Accuracy for the Prediction of Survival Times

- Patrick J. Heagerty PhD
- Department of Biostatistics
- University of Washington

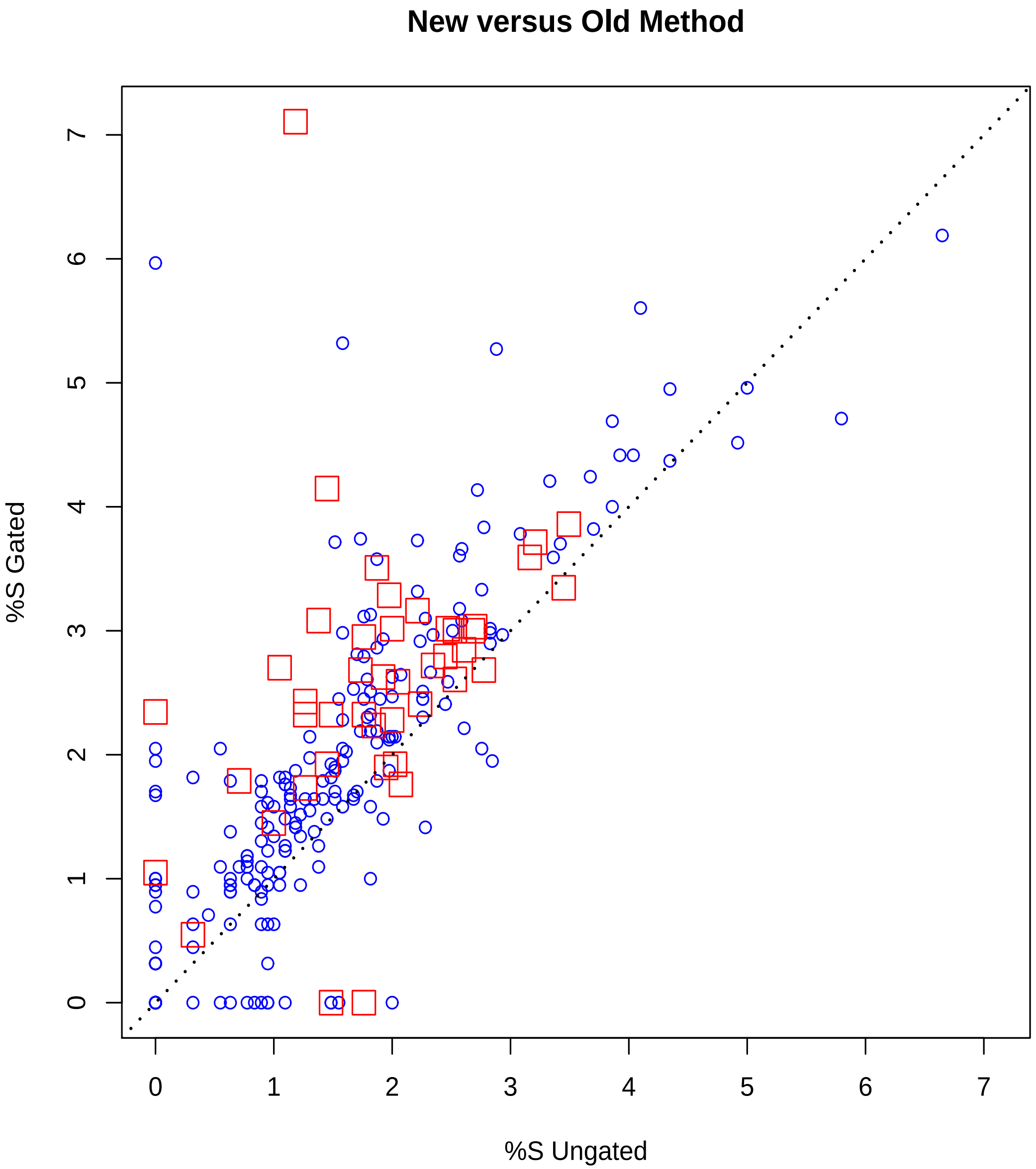
Session Outline

- Examples
  - Breast Cancer Cytology Data
  - Mayo PBC Data
  - Cystic Fibrosis Foundation Registry Data
- Previous approaches
- ROC overview
- TP, FP for survival outcomes
- $ROC_t^{\mathbb{C}/\mathbb{D}}(p)$
- $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$, $AUC(t)$, and concordance, C^τ
- $\overline{ROC}(p)$ interpretation and dynamic criteria, $c^p(t)$
- Further work

Example: Comparing cytometry measures

Breast Cancer among Younger Women

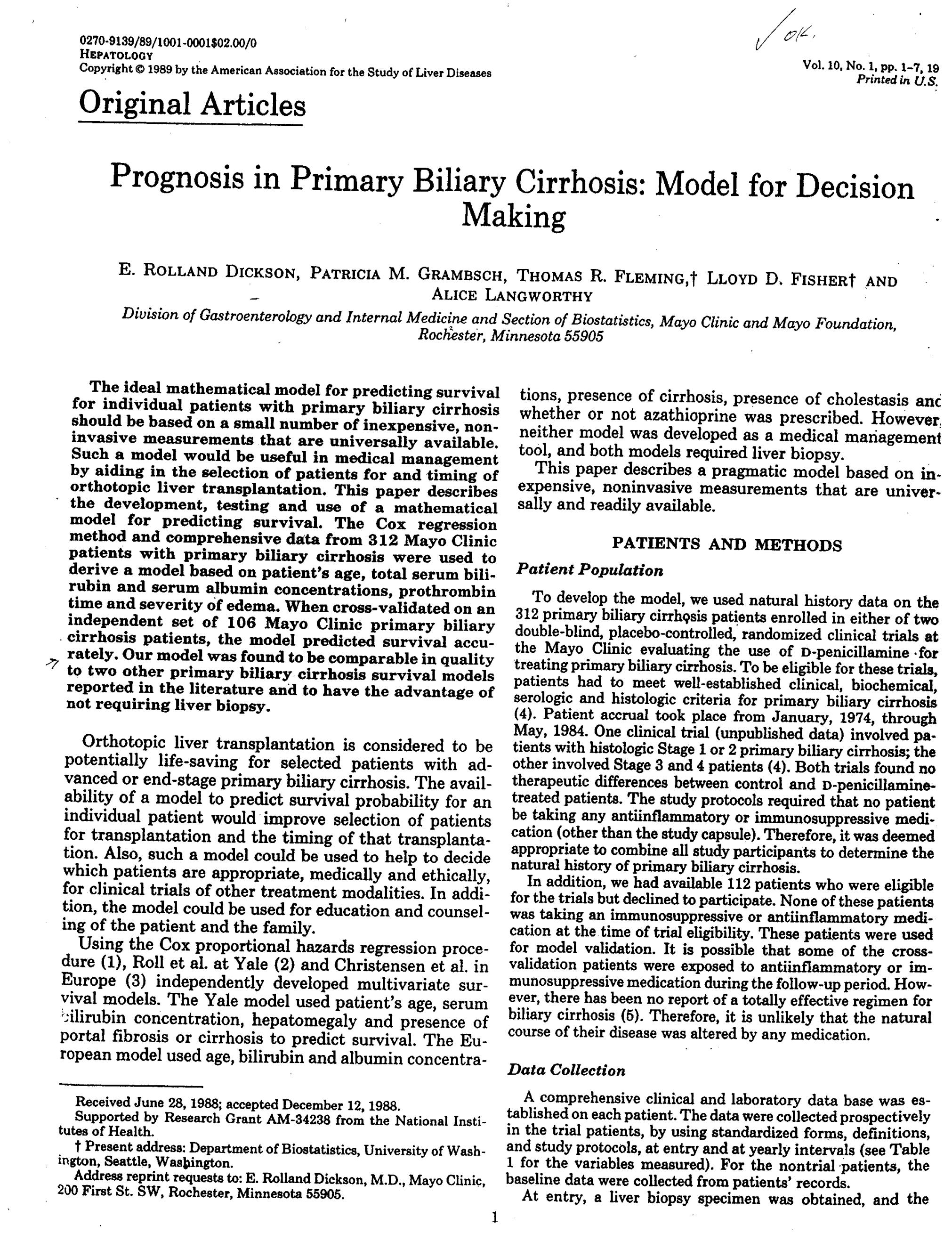
- $N = 253$ women BC diagnosed aged 20 to 44.
- Endpoint: time-until-death (any cause)
- Cytometry measurements:
  - ▷ "old" (S-phase ungated)
  - ▷ "new" (S-phase gated)
- Goal: compare measures as predictors of mortality
- Heagerty, Lumley & Pepe (2000)


New versus Old Method
%S Gated
0
1
2
3
4
5
6
7
%S Ungated
0
1
2
3
4
5
6
7

Predictive Survival Models

Mayo PBC Data

- $N = 312$ subjects, 125 deaths, 1974-1986
- Baseline measurements: bilirubin, prothrombin time, albumin...
- Goal: predict mortality; "medical management"



√ ok.

0270-9139/89/1001-0001$02.00/0
HEPATOLOGY
Copyright © 1989 by the American Association for the Study of Liver Diseases

Vol. 10, No. 1, pp. 1–7, 19
Printed in U.S.

Original Articles

Prognosis in Primary Biliary Cirrhosis: Model for Decision Making

E. ROLLAND DICKSON, PATRICIA M. GRAMBSCH, THOMAS R. FLEMING,† LLOYD D. FISHER† AND ALICE LANGWORTHY

Division of Gastroenterology and Internal Medicine and Section of Biostatistics, Mayo Clinic and Mayo Foundation, Rochester, Minnesota 55905

The ideal mathematical model for predicting survival for individual patients with primary biliary cirrhosis should be based on a small number of inexpensive, noninvasive measurements that are universally available. Such a model would be useful in medical management by aiding in the selection of patients for and timing of orthotopic liver transplantation. This paper describes the development, testing and use of a mathematical model for predicting survival. The Cox regression method and comprehensive data from 312 Mayo Clinic patients with primary biliary cirrhosis were used to derive a model based on patient's age, total serum bilirubin and serum albumin concentrations, prothrombin time and severity of edema. When cross-validated on an independent set of 106 Mayo Clinic primary biliary cirrhosis patients, the model predicted survival accurately. Our model was found to be comparable in quality to two other primary biliary cirrhosis survival models reported in the literature and to have the advantage of not requiring liver biopsy.

Orthotopic liver transplantation is considered to be potentially life-saving for selected patients with advanced or end-stage primary biliary cirrhosis. The availability of a model to predict survival probability for an individual patient would improve selection of patients for transplantation and the timing of that transplantation. Also, such a model could be used to help to decide which patients are appropriate, medically and ethically, for clinical trials of other treatment modalities. In addition, the model could be used for education and counseling of the patient and the family.

Using the Cox proportional hazards regression procedure (1), Roll et al. at Yale (2) and Christensen et al. in Europe (3) independently developed multivariate survival models. The Yale model used patient's age, serum bilirubin concentration, hepatomegaly and presence of portal fibrosis or cirrhosis to predict survival. The European model used age, bilirubin and albumin concentrations, presence of cirrhosis, presence of cholestasis and whether or not azathioprine was prescribed. However, neither model was developed as a medical management tool, and both models required liver biopsy.

This paper describes a pragmatic model based on inexpensive, noninvasive measurements that are universally and readily available.

PATIENTS AND METHODS

Patient Population

To develop the model, we used natural history data on the 312 primary biliary cirrhosis patients enrolled in either of two double-blind, placebo-controlled, randomized clinical trials at the Mayo Clinic evaluating the use of D-penicillamine for treating primary biliary cirrhosis. To be eligible for these trials, patients had to meet well-established clinical, biochemical, serologic and histologic criteria for primary biliary cirrhosis (4). Patient accrual took place from January, 1974, through May, 1984. One clinical trial (unpublished data) involved patients with histologic Stage 1 or 2 primary biliary cirrhosis; the other involved Stage 3 and 4 patients (4). Both trials found no therapeutic differences between control and D-penicillamine-treated patients. The study protocols required that no patient be taking any antiinflammatory or immunosuppressive medication (other than the study capsule). Therefore, it was deemed appropriate to combine all study participants to determine the natural history of primary biliary cirrhosis.

In addition, we had available 112 patients who were eligible for the trials but declined to participate. None of these patients was taking an immunosuppressive or antiinflammatory medication at the time of trial eligibility. These patients were used for model validation. It is possible that some of the cross-validation patients were exposed to antiinflammatory or immunosuppressive medication during the follow-up period. However, there has been no report of a totally effective regimen for biliary cirrhosis (5). Therefore, it is unlikely that the natural course of their disease was altered by any medication.

Data Collection

A comprehensive clinical and laboratory data base was established on each patient. The data were collected prospectively in the trial patients, by using standardized forms, definitions, and study protocols, at entry and at yearly intervals (see Table 1 for the variables measured). For the nontrial patients, the baseline data were collected from patients' records.

At entry, a liver biopsy specimen was obtained, and the

Received June 28, 1988; accepted December 12, 1988.

Supported by Research Grant AM-34238 from the National Institutes of Health.

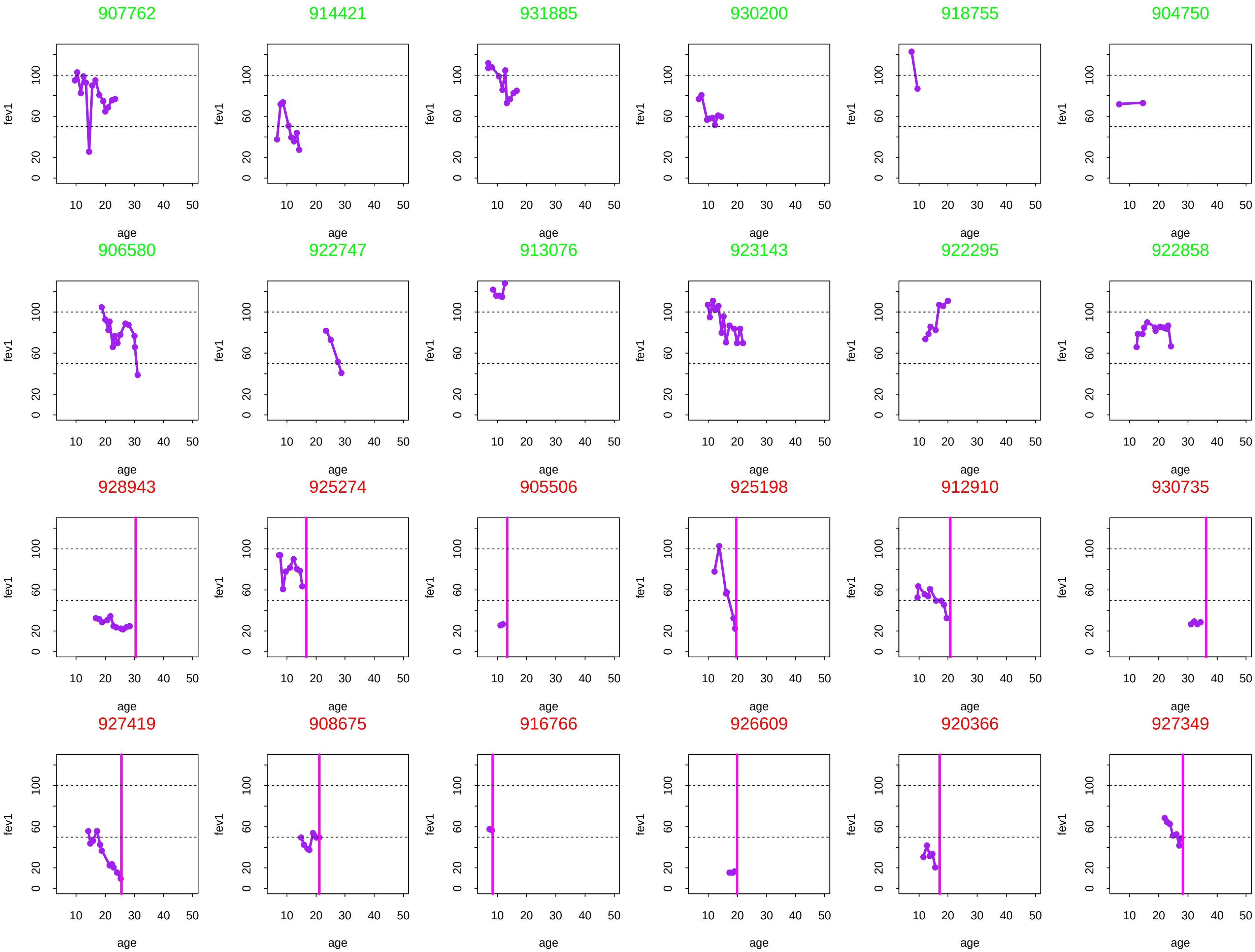
† Present address: Department of Biostatistics, University of Washington, Seattle, Washington.

Address reprint requests to: E. Rolland Dickson, M.D., Mayo Clinic, 200 First St. SW, Rochester, Minnesota 55905.

Predictive Survival Models

Cystic Fibrosis Data

- $N = 23,530$ subjects, $4,772$ deaths, 1986-2000
- $n = 160,005$ longitudinal observations
- Longitudinal measurements: FEV1, weight, height
- Goal: predict mortality; transplantation selection


907762
914421
931885
930200
918755
904750
906580
922747
913076
923143
922295
922858
928943
925274
905506
925198
912910
930735
927419
908675
916766
926609
920366
927349
fev1
0
20
60
100
age
10
20
30
40
50

Accuracy: Some proposals

R^2 Generalizations

- Korn and Simon (1990)
- Schemper and Henderson (2000)
- O'Quigley and Xu (2001)

TP, FP, ROC Generalizations

- Etzioni et al. (1999); Slate and Turnbull (1999)
- Heagerty, Lumley, and Pepe (2000)
- Heagerty and Zheng (2005)

Some Comments

- Schemper and Henderson (2000), p. 249:

  > “Consequently, there have been a number of attempts to develop measures akin to R^2 for Cox proportional hazards models [[$references$]], though as yet, none have been generally accepted.”

- These versions of R^2 are not about variance in T. They focus on average variances of the counting process:

$$N(t) = 1(T \leq t)$$

R^2: Schemper and Henderson (2000)

Idea: $N(t) = 1(T \leq t)$ with $E[N(t)] = 1 - S(t)$

	Without Covariates	With Covariates
variance	$S(t)[1 - S(t)]$	$S(t \mid X)[1 - S(t \mid X)]$
average (X)		$E_X \{S(t \mid X)[1 - S(t \mid X)]\}$
average (T)	$\int_t S(t)[1 - S(t)] f(t) dt$	$\int_t E_X \{S(t \mid X)[1 - S(t \mid X)]\} f(t) dt$
	$\downarrow$	$\downarrow$
	D_0	D_X

Proposal: $R^2 = (D_0 - D_X)/D_0$

Some Comments

- Natural to think of survival through counting process $N(t)$.
- Common to use ROC curves for logistic regression / binary classification.

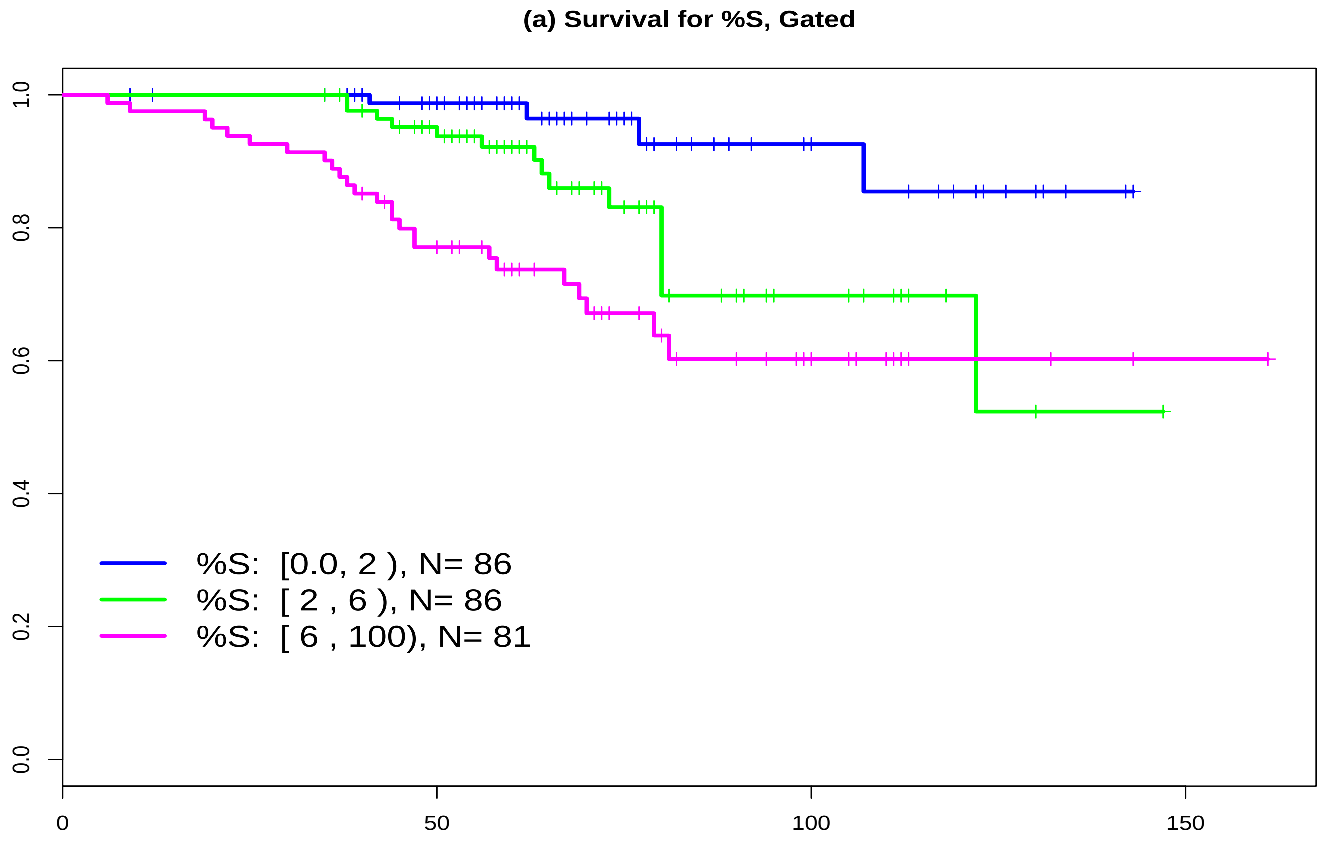
$\boxed{\text{Q:}}$ Extend classification error rate concepts to survival data?

Components of Accuracy

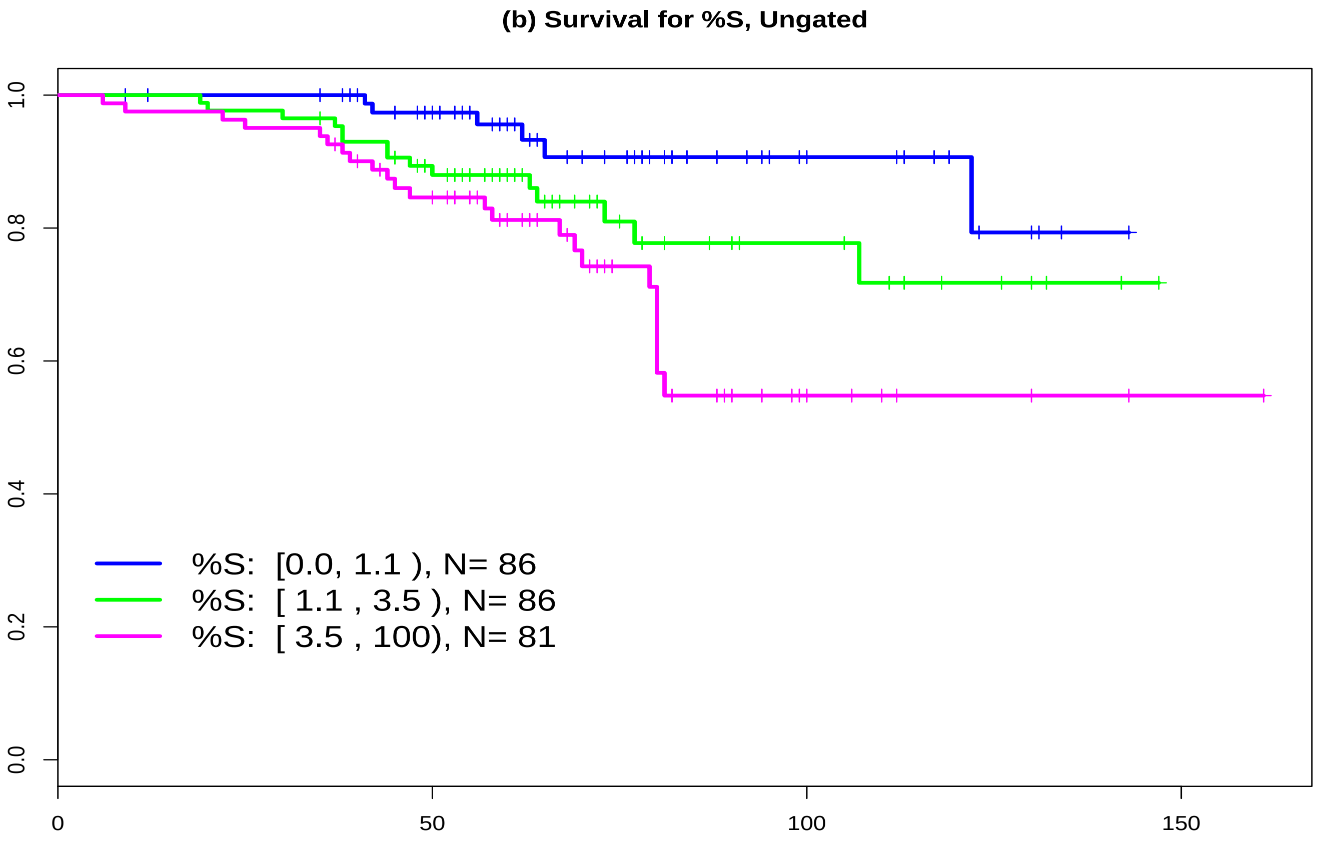
- **Calibration**
  - ▷ Bias – does observed match predicted?
  - ▷ Evaluated graphically and formally.
- **Discrimination**
  - ▷ Does prediction separate subjects with different risks?
  - ▷ Evaluated qualitatively based on K-M plots.

Harrell, Lee and Mark (1996)

BC Survival: New Measurement (gated)



BC Survival: Old Measurement (ungated)



Binary Classification

Sensitivity “**True Positive**”

BINARY TEST : $P(T+ \mid D = 1)$

CONTINUOUS MARKER : $P(M > c \mid D = 1)$

Specificity “**True Negative**”

BINARY TEST : $P(T- \mid D = 0)$

CONTINUOUS MARKER : $P(M \leq c \mid D = 0)$

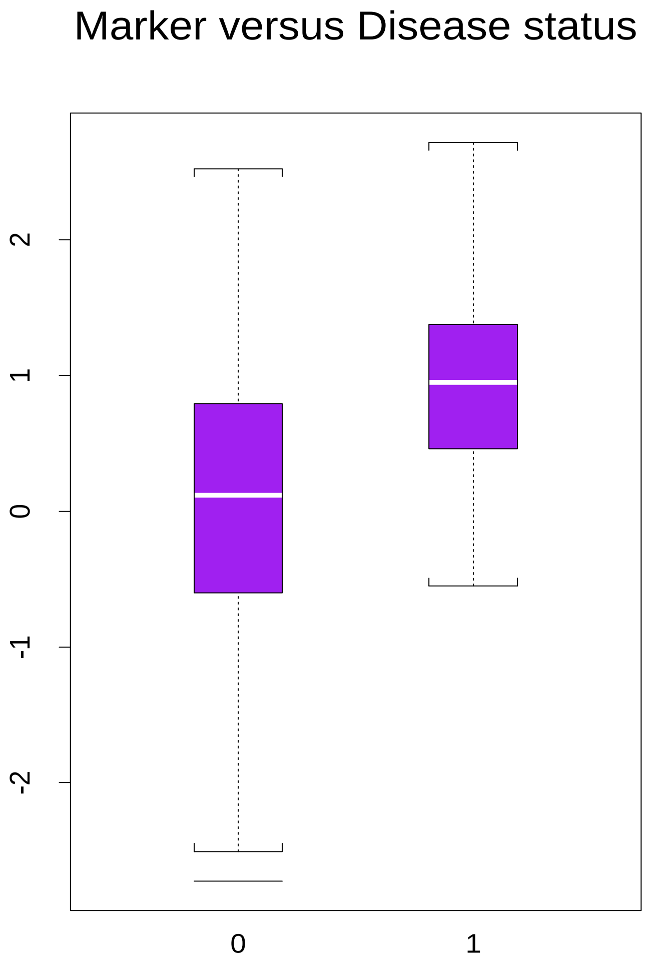
ROC Curve

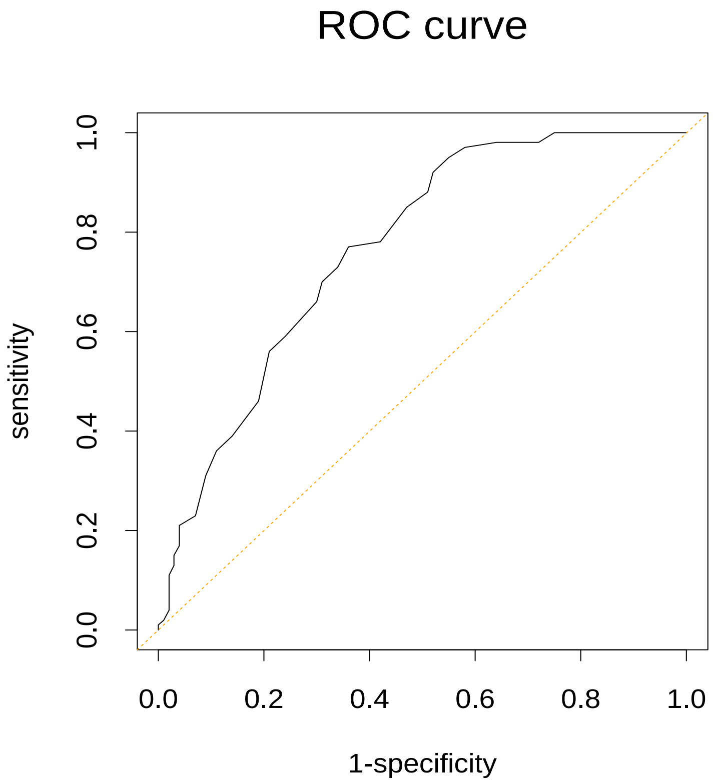
An ROC curve plots the **True Positive Rate**, TP(**c**), versus the False Positive Rate, FP(c) for all possible cutpoints, **c**:

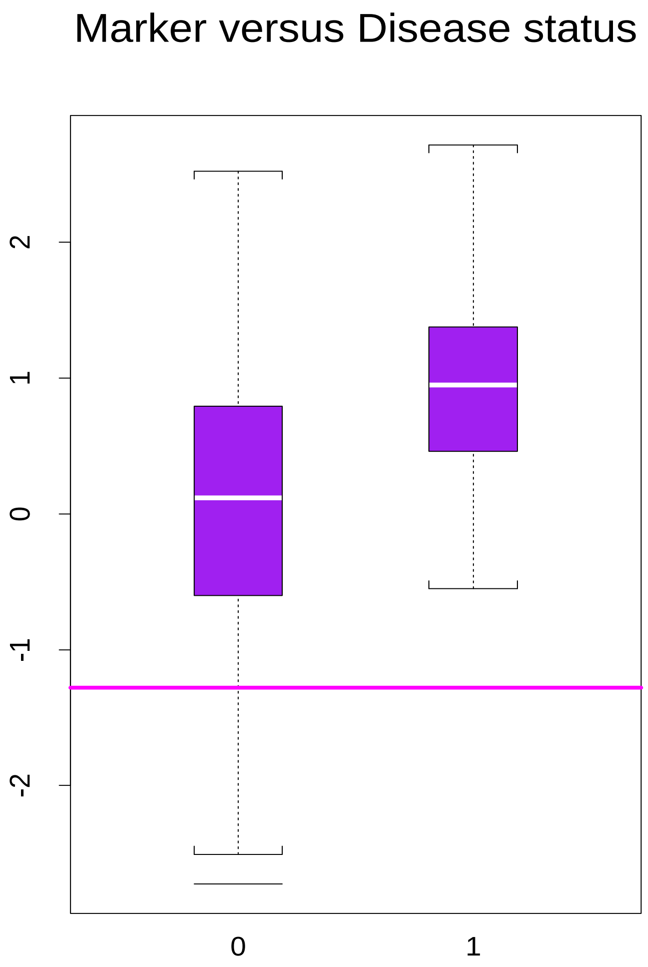
$$\text{FP}(c) \quad = \quad P(M > c \mid D = 0)$$

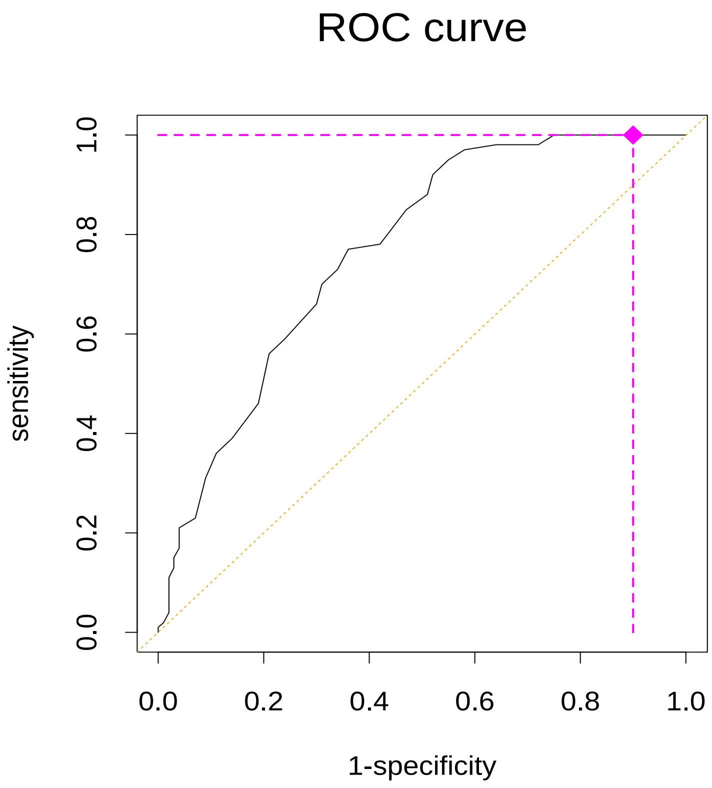
$$\text{TP}(c) \quad = \quad P(M > c \mid D = 1)$$

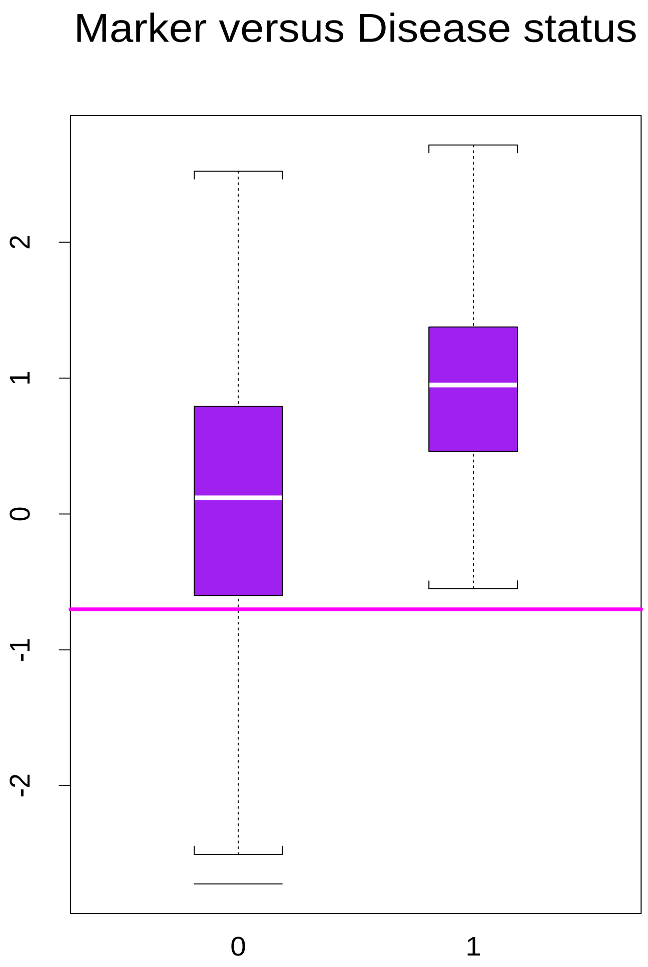
$$\text{ROC Curve} \quad : \quad [\, FP(c),\, TP(c)\,] \quad \forall c$$

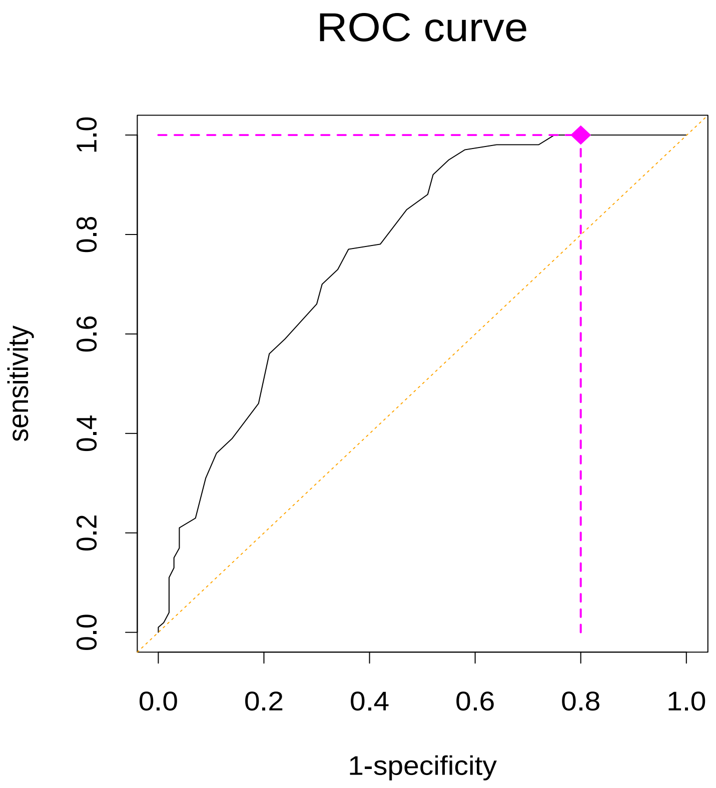

Marker versus Disease status
2
1
0
-1
-2
0
1

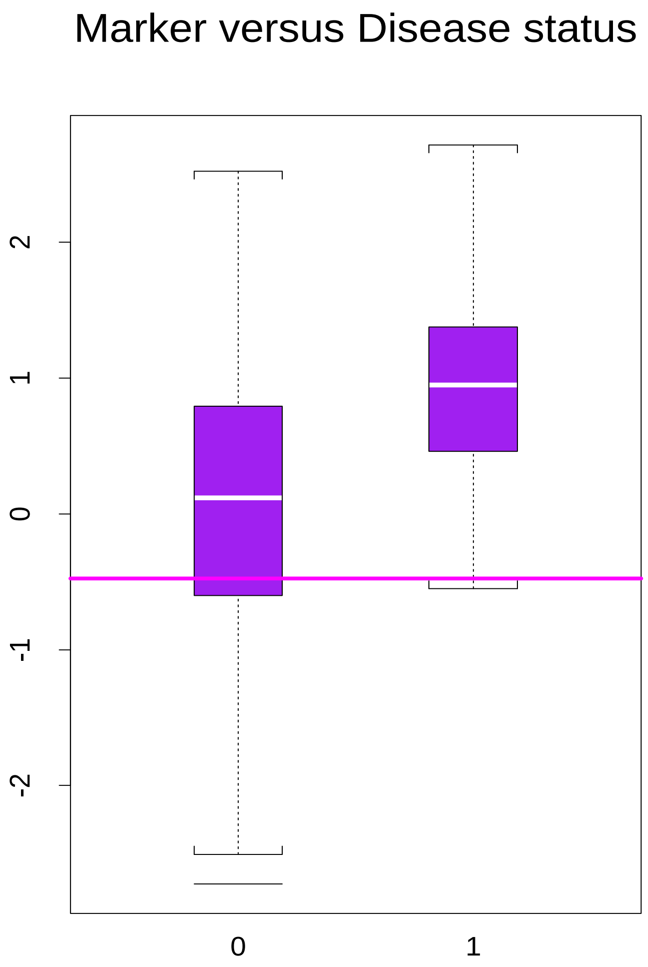

ROC curve
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

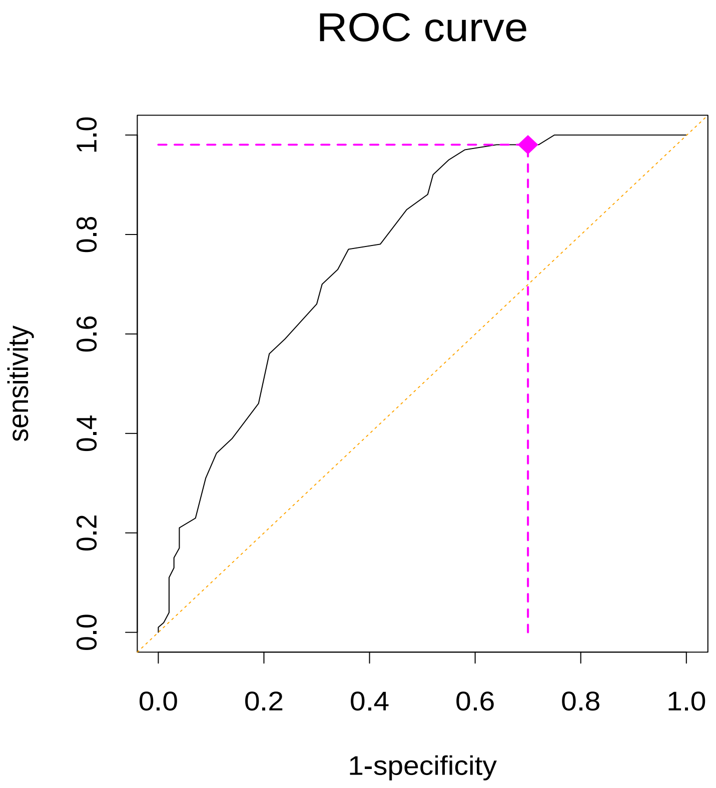

Marker versus Disease status
2
1
0
-1
-2
0
1

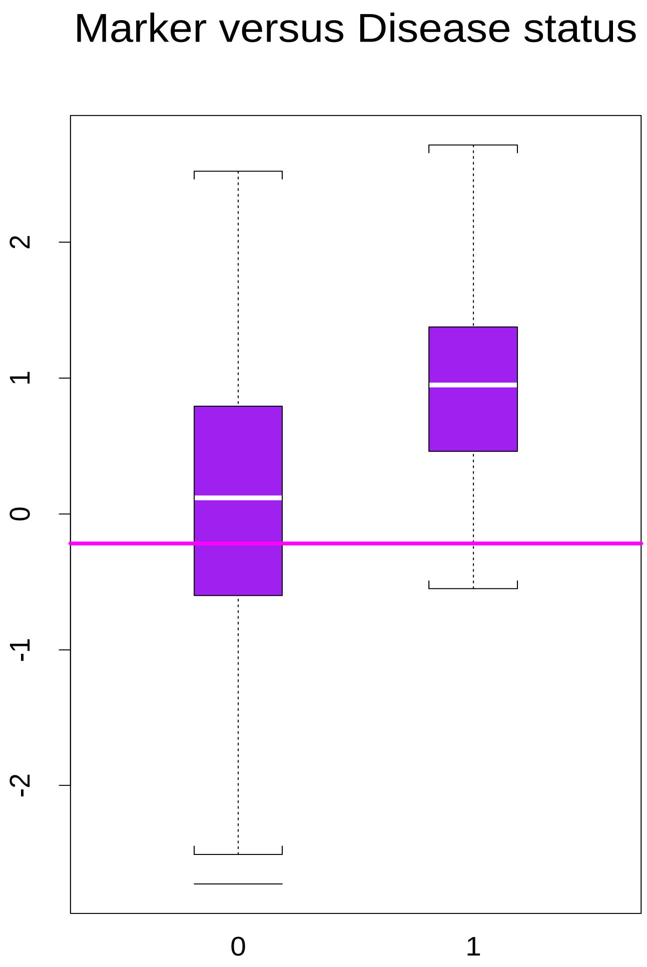

ROC curve
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

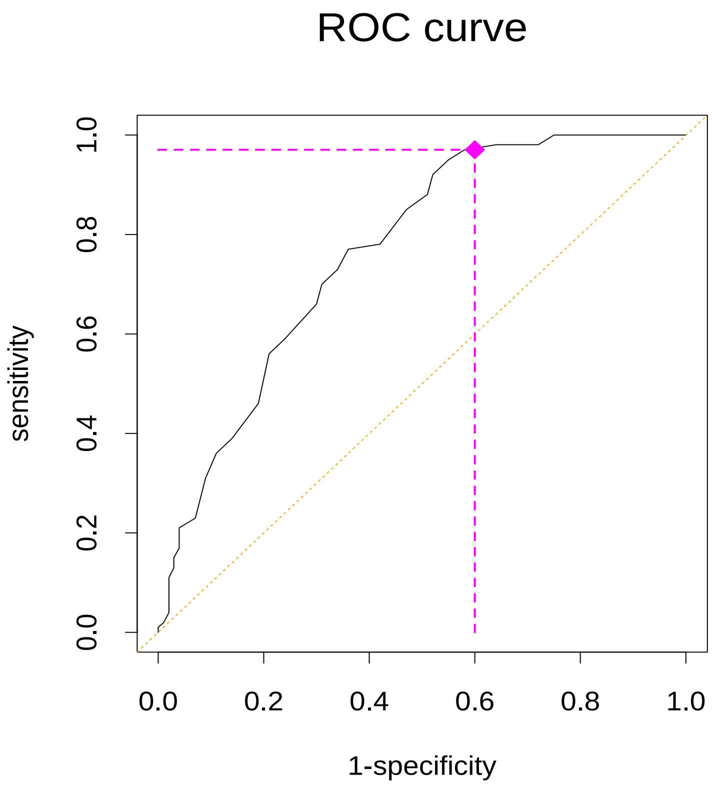

Marker versus Disease status
2
1
0
-1
-2
0
1

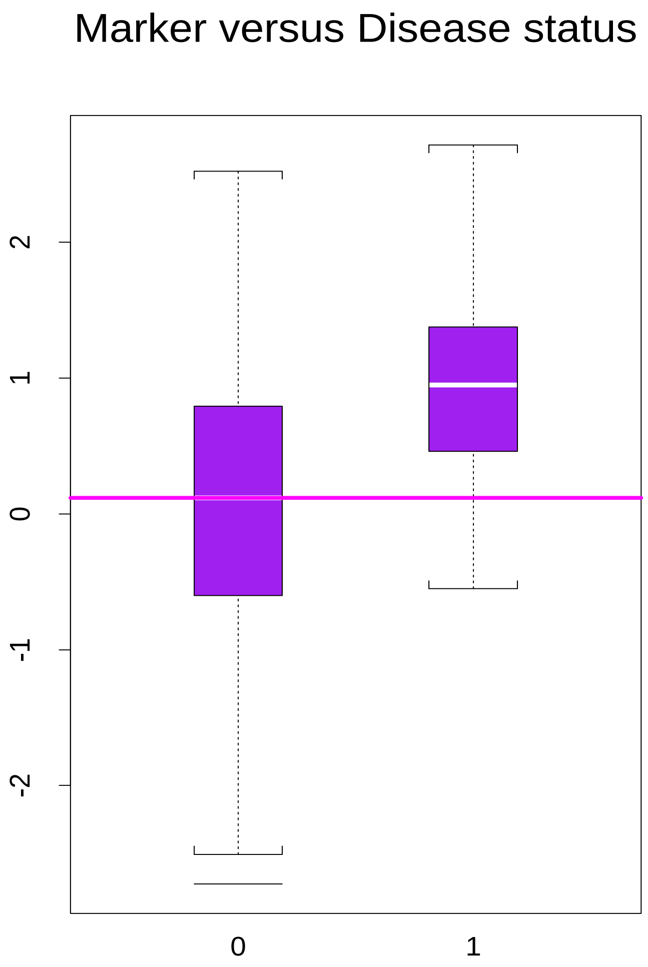

ROC curve
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity


Marker versus Disease status
2
1
0
-1
-2
0
1

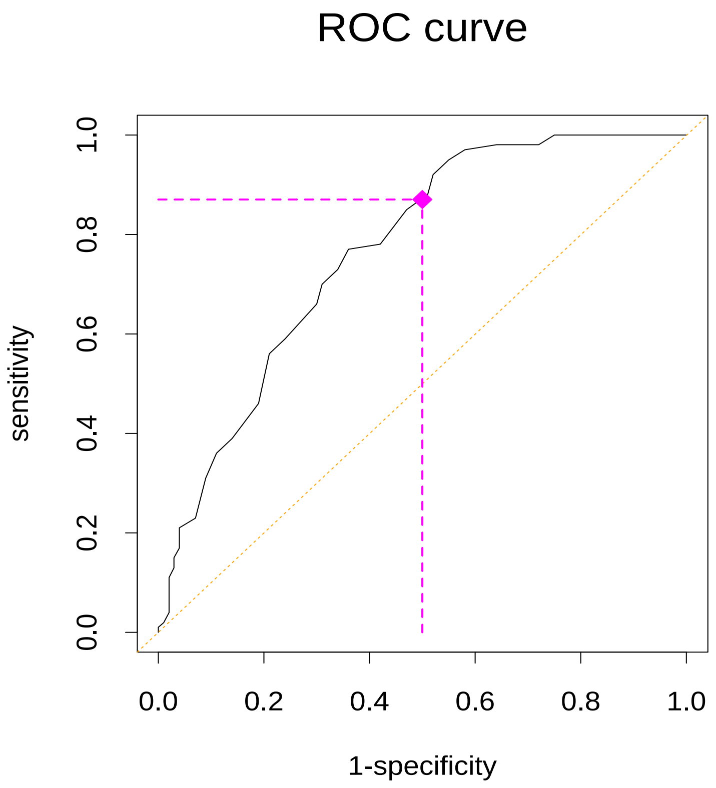

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity


Marker versus Disease status
2
1
0
-1
-2
0
1

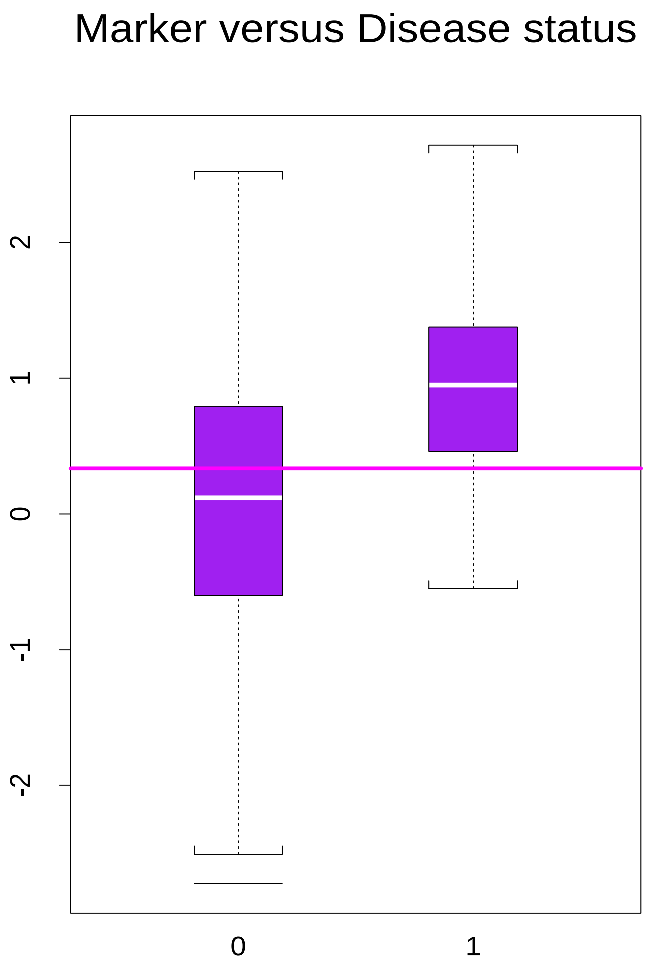

ROC curve
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity



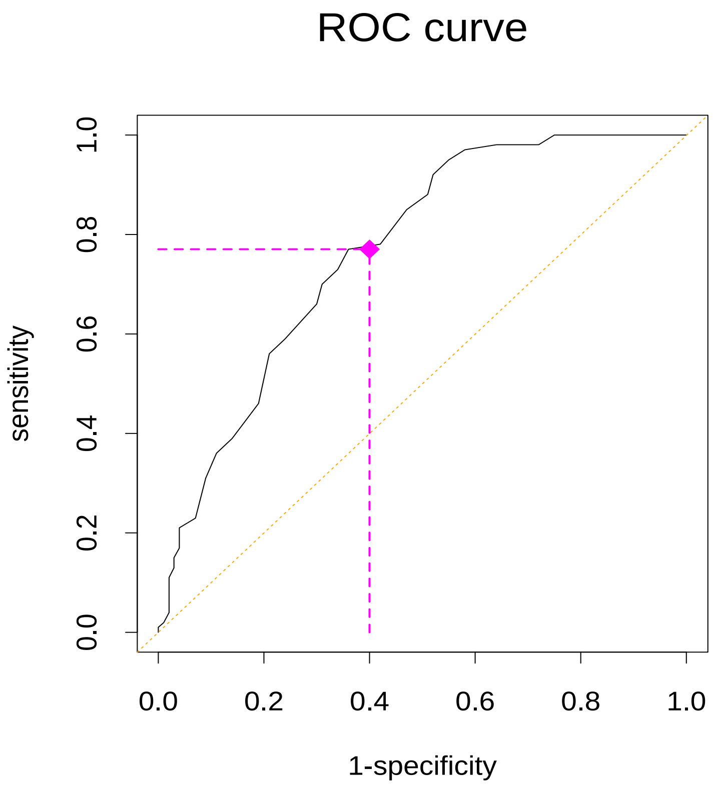
Marker versus Disease status



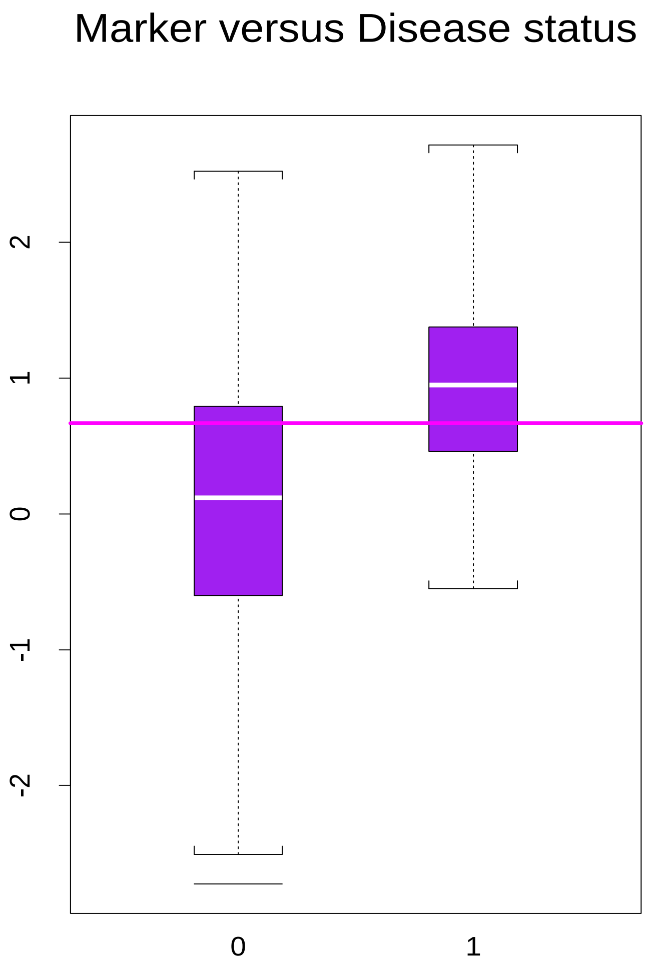
ROC curve

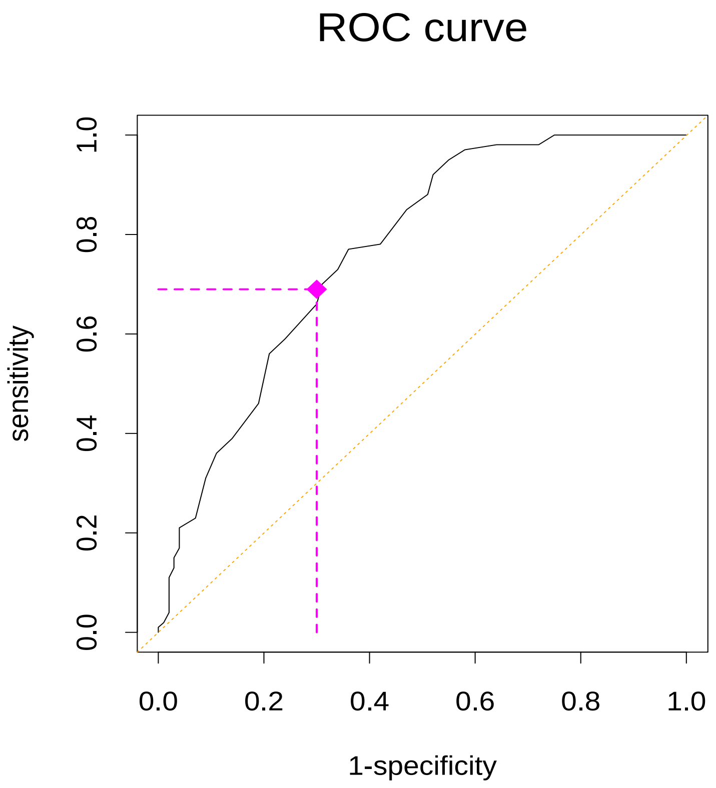


Marker versus Disease status

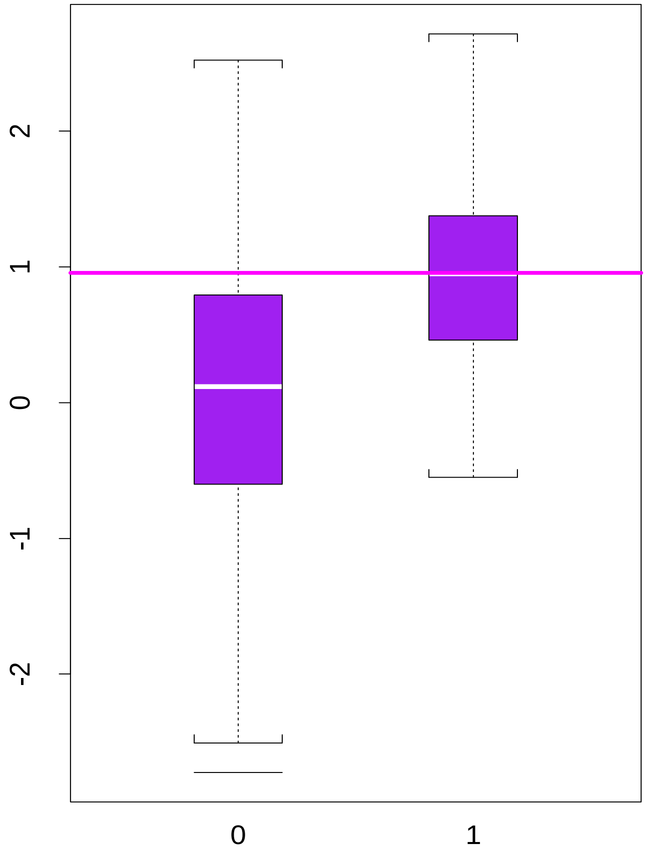


ROC curve

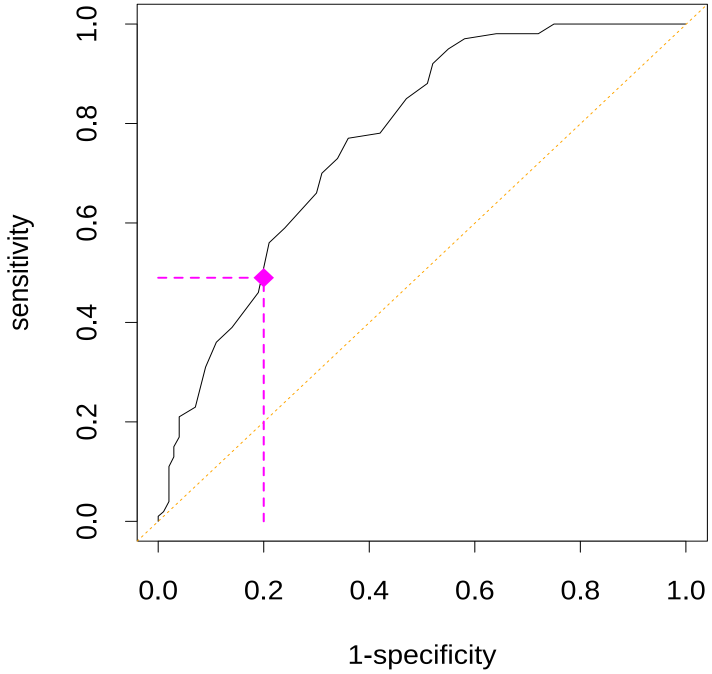

Marker versus Disease status
2
1
0
-1
-2
0
1

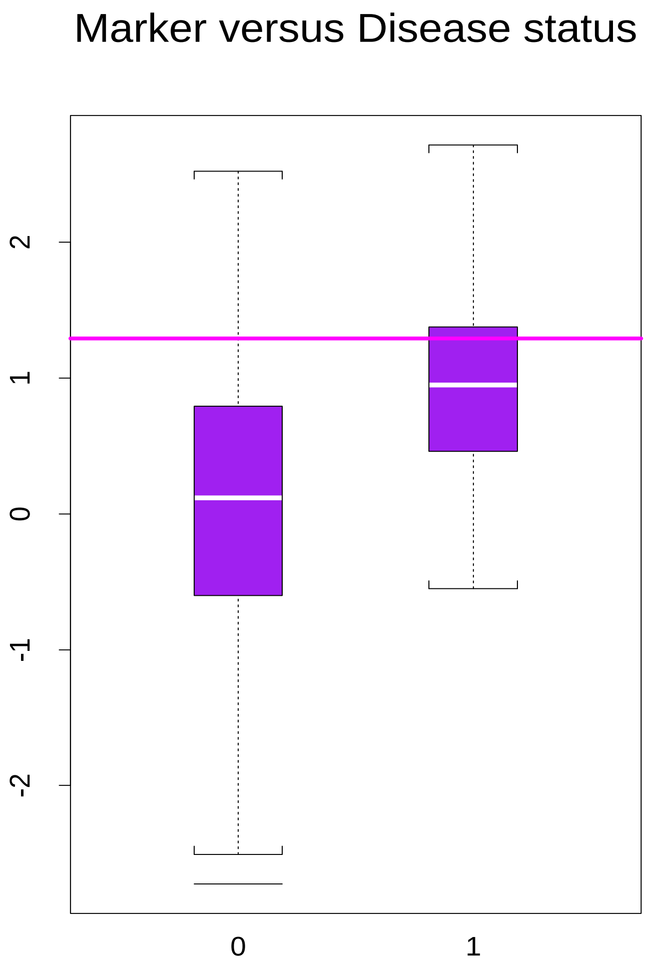

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

Marker versus Disease status

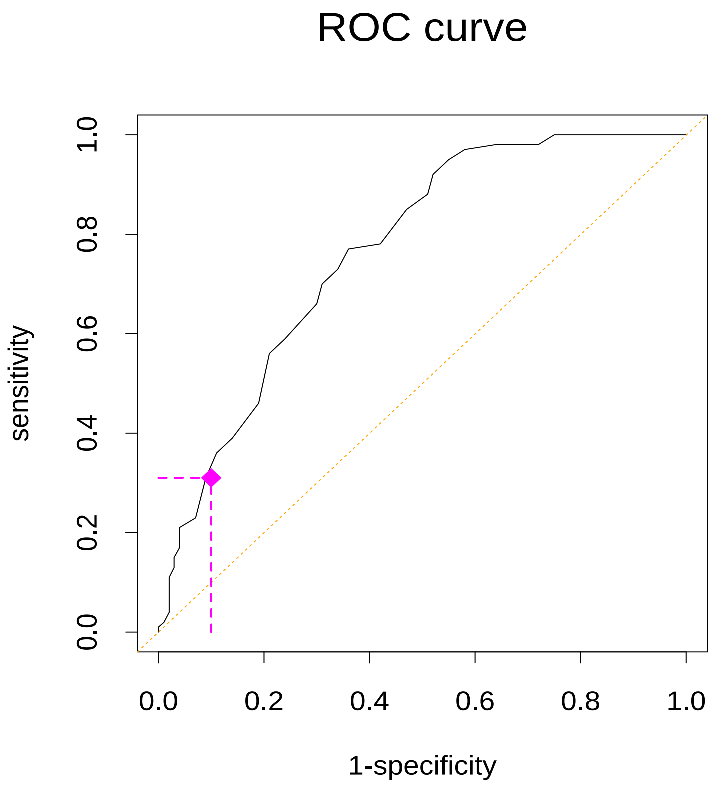


ROC curve





Marker versus Disease status



ROC curve

ROC Curves

1. Compare different markers over full spectrum of error combinations.
2. Compare sensitivity when controlling specificity (eg. TP when FP=10%).
3. AUC interpretation:
   "For a randomly chosen case and control, the area under the ROC curve is the probability that the marker for the case is greater than the marker for the control."
4. AUC is a marker-outcome concordance summary.

Does a (repeated) measurement predict onset?

- $\boxed{\textbf{Q}:}$ Can a **measurement** accurately predict which **CASES** will experience an event (soon)?
  - ▷ e.g. FEV1
  - ▷ e.g. death time

- $\boxed{\textbf{Q}:}$ Can a **measurement** be used to accurately guide longitudinal treatment decisions?
  - ▷ e.g. lung transplantation

Classification Errors

- **True Positive Rate**

  $$P(\text{ high measurement } | \textbf{ CASE })$$

- **False Positive Rate**

  $$P(\text{ high measurement } | \textbf{ CONTROL })$$

Issues Related to Time

- **Q**: When is the **measurement** taken?
  - At baseline: $Y(0)$
  - At a follow-up time t: $Y(t)$
- **Q**: What time is used to determine when someone is a **CASE** or a **CONTROL**?
  - **Case** $=$ event (disease, death) before time t.
  - **Case** $=$ event (disease, death) at time t.
  - **Control** $=$ event-free through time t.
  - **Control** $=$ event-free through a large follow-up time $t^{\star}$.

Our approaches

	measurement	case	control
Heagerty, Lumley & Pepe (2000)	baseline	$T \leq t$	$T > t$
Zheng & Heagerty (2004)	longitudinal	$T = t$	$T > t^{\star}$
Heagerty & Zheng (2005)	baseline	$T = t$	$T > t$
Zheng & Heagerty (2007)	longitudinal	$T \leq t$	$T > t$
Saha & Heagerty (2011)	longitudinal	$T = t$, $\delta = j$	$T > t$

Sensitivity and Specificity for Survival

Let T denote the survival time, and let $N(t)$ denote the counting process for the uncensored outcome:

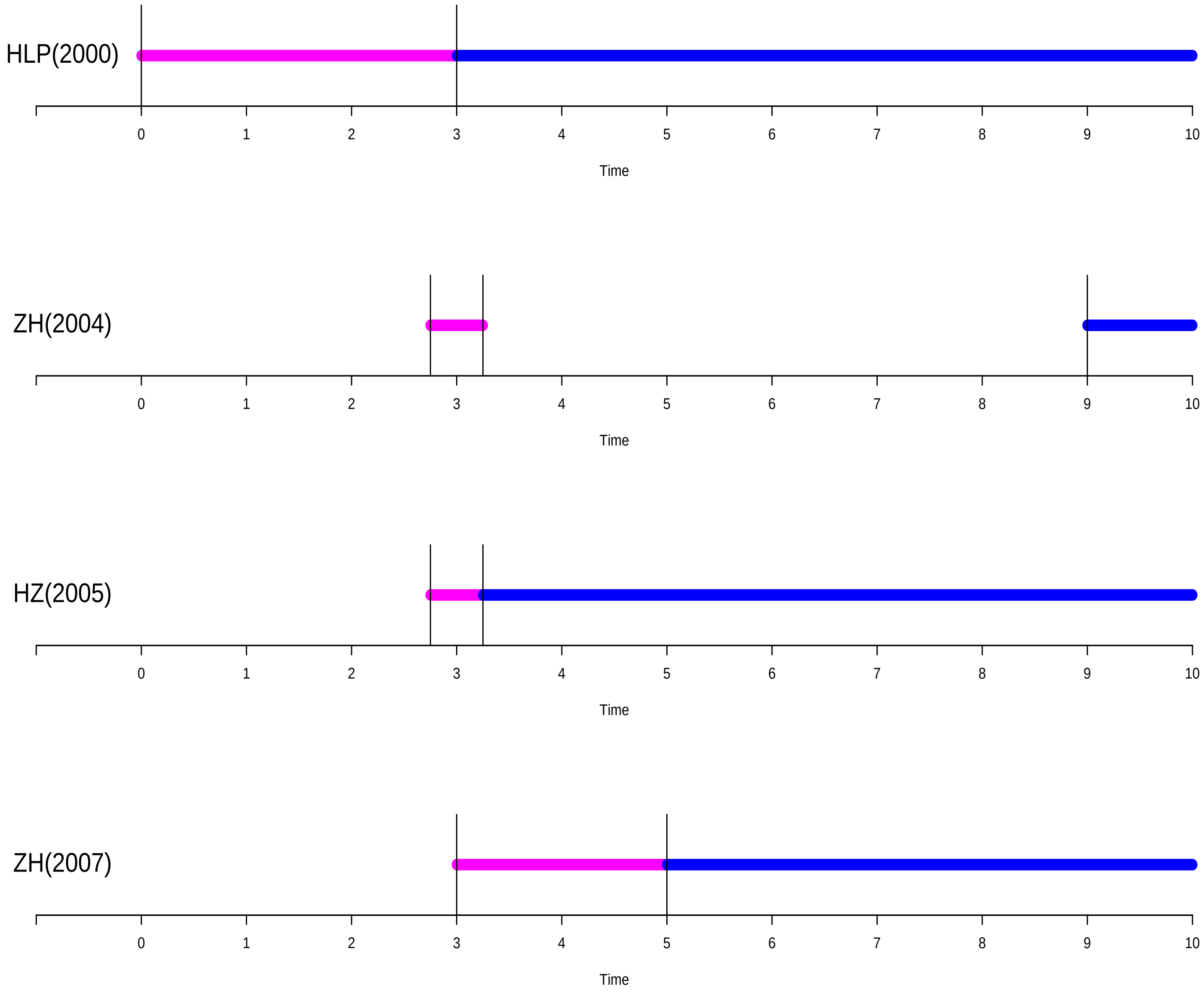
$$N(t) = 1(T \leq t)$$

Possible definitions:

$$\text{CASE}(t) \quad : \quad \begin{cases} \text{Cumulative} & N(t) = 1 \\ \text{Incident} & dN(t) = 1 \end{cases}$$

$$\text{CONTROL}(t) \quad : \quad \begin{cases} \text{Static} & N(t^\star) = 0 \\ \text{Dynamic} & N(t) = 0 \end{cases}$$

○ Where $t^\star$ is a fixed “large” time, $t^\star >> t$.


HLP(2000)
0
1
2
3
4
5
6
7
8
9
10
Time
ZH(2004)
0
1
2
3
4
5
6
7
8
9
10
Time
HZ(2005)
0
1
2
3
4
5
6
7
8
9
10
Time
ZH(2007)
0
1
2
3
4
5
6
7
8
9
10
Time

[1] Sensitivity and Specificity for Survival

Define: Heagerty, Lumley & Pepe (2000)

$$\text{sensitivity}^{\mathbb{C}}(c,t) \quad : \quad P(M > c \mid T \leq t)$$

$$P(M > c \mid N(t) = 1)$$

$$\text{specificity}^{\mathbb{D}}(c,t) \quad : \quad P(M \leq c \mid T > t)$$

$$P(M \leq c \mid N(t) = 0)$$

$$TP_t^{\mathbb{C}}(c) = P(M > c \mid N(t) = 1)$$

$$FP_t^{\mathbb{D}}(c) = P(M > c \mid N(t) = 0)$$

[1] Time-dependent ROC Curve

An Cumulative/Dynamic ROC curve shows the ability of a marker to separate the cumulative cases through time t (e.g. $T \leq t$) from the controls at time t (e.g. $T > t$).

Define curve $[\ p,\ ROC_t^{\mathbb{C}/\mathbb{D}}(p)\ ]$:

$$\begin{aligned}
\text{ROC}_t^{\mathbb{C}/\mathbb{D}}(p) &= TP^{\mathbb{C}}\left\{[FP^{\mathbb{D}}]^{-1}(p)\right\} \\
&= TP^{\mathbb{C}}(c^p) \\
\text{where} \quad & c^p \ : \ p = FP^{\mathbb{D}}(c^p)
\end{aligned}$$

Define AUC as:

$$AUC(t) = \int ROC_t^{\mathbb{C}/\mathbb{D}}(p)\, dp$$

Estimation: NNE for $S(m, t)$

With censored times a valid ROC solution can be provided by using an estimator of the bivariate distribution function for $[M, T]$.

$\boxed{\text{Define:}}$

$$\begin{aligned} S(c,t) &= P[M > c, T > t] \\ S(c,t) &= \int_c^{\infty} S(t \mid M = u) dF_M(u) \end{aligned}$$

where $F_M(u)$ is the distribution function for M.

$\boxed{\text{Akritas (1994):}}$

$$\widehat{S}_{\lambda_n}(c,t) = \frac{1}{n}\sum_i \widehat{S}_{\lambda_n}(t \mid M = M_i)\mathbf{1}(M_i > c)$$

where $\widehat{S}_{\lambda_n}(t \mid M = M_i)$ is a suitable estimator of the conditional survival function characterized by a smoothing parameter λ_n.

Estimation: NNE for $S(m, t)$

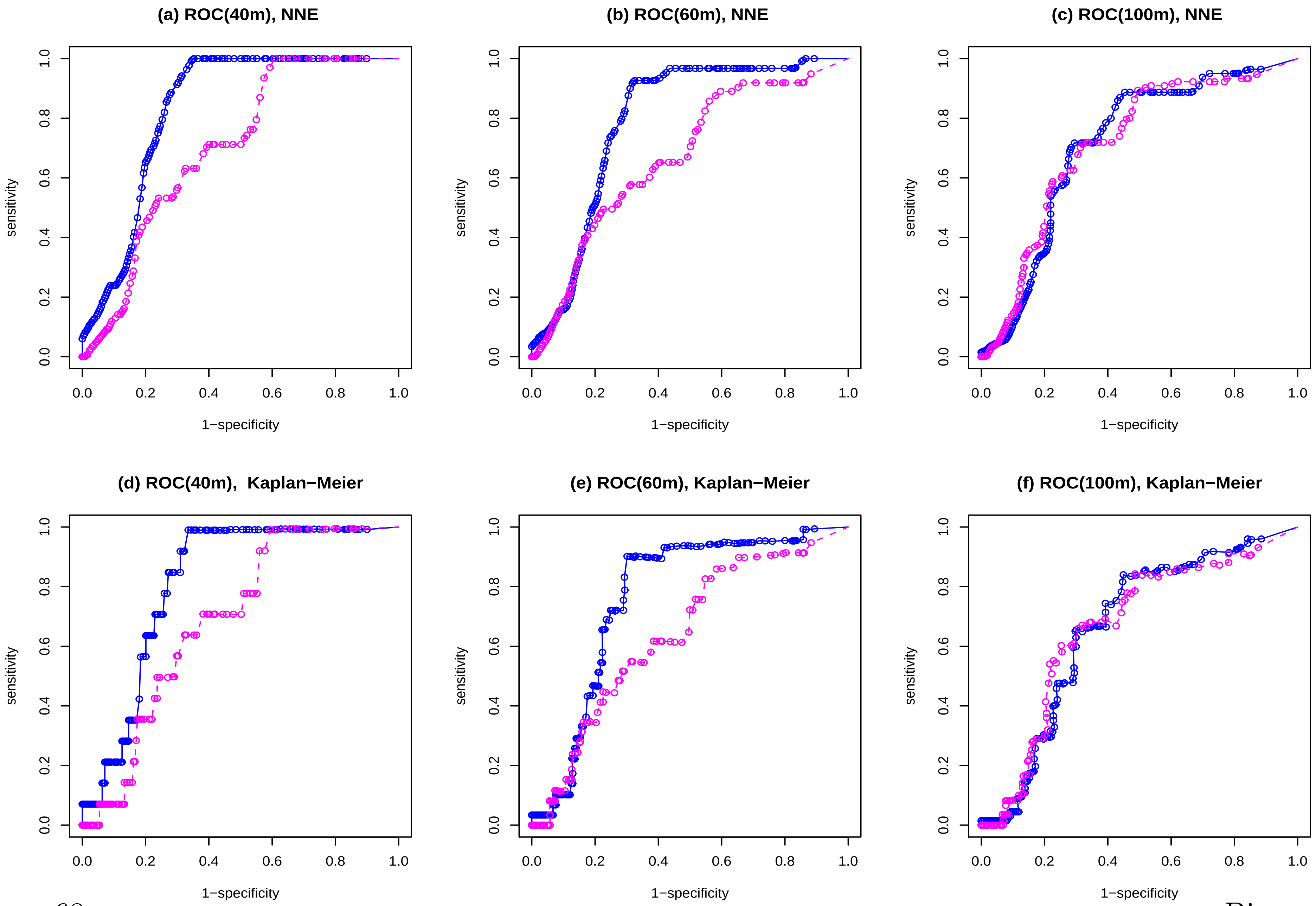
Define:

$$\begin{aligned}
\widehat{P}_{\lambda_n}[M > c \mid N(t) = 1] &= \frac{\{[1 - \widehat{F}_M(c)] - \widehat{S}_{\lambda_n}(c,t)\}}{\{1 - \widehat{S}_{\lambda_n}(t)\}} \\
\widehat{P}_{\lambda_n}[M \leq c \mid N(t) = 0] &= 1 - \frac{\widehat{S}_{\lambda_n}(c,t)}{\widehat{S}_{\lambda_n}(t)}
\end{aligned}$$

where $\widehat{S}_{\lambda_n}(t) = \widehat{S}_{\lambda_n}(-\infty, t)$.

- For NNE $\lambda_n = O(n^{-1/3})$ sufficient for weak consistency.

- Results from Akritas (1994) and van de Vaart and Wellner (1996) imply that the bootstrap can be used for inference.



(a) ROC(40m), NNE
(b) ROC(60m), NNE
(c) ROC(100m), NNE
(d) ROC(40m), Kaplan−Meier
(e) ROC(60m), Kaplan−Meier
(f) ROC(100m), Kaplan−Meier
sensitivity
1−specificity
0.0
0.2
0.4
0.6
0.8
1.0

Accuracy Comparisons

Sensitivity

- Using gated (new) measurement:
  - $\widehat{P}[M_1 \leq 5.4 \mid N(60m) = 0] = 0.71$
  - $\widehat{P}[M_1 > 5.4 \mid N(60m) = 1] = 0.82$

- Using ungated (old) measurement:
  - $\widehat{P}[M_2 \leq \boxed{3.5} \mid N(60m) = 0] = \boxed{0.71}$
  - $\widehat{P}[M_2 > 3.5 \mid N(60m) = 1] = 0.54$

- Therefore, controlling M_1 and M_2 to have equal specificity, the new measure, M_1, has a greater sensitivity.

Accuracy Comparisons

AUC Calculations

- AUC for gated (new): 0.80,
  Bootstrap 95% CI: (0.72,0.89)

- AUC for ungated (old): 0.68,
  Bootstrap 95% CI: (0.56,0.77)

- 95% CI for difference in areas (new-old): (0.03, 0.26)

Summary

- Define time-dependent ROC curves based on prospective data and **cumulative** occurence of events.
- Local survival estimation handles censored event times.
- Tool to evaluate a marker or a survival regression model score.
- R package: `survivalROC` (see web for doc/validation)
- ⋆ Alternative definitions?
- ⋆ More general longitudinal marker scenario?

• Heagerty, Lumley and Pepe (2000) "Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker" *Biometrics*

[2] Sensitivity and Specificity for Survival

Define: Heagerty & Zheng (2005)

$$\begin{aligned} \text{sensitivity}^{\mathbb{I}}(c,t) &: \quad P(M > c \mid T = t) \\ & \quad\;\; P(M > c \mid dN(t) = 1) \end{aligned}$$

$$\begin{aligned} \text{specificity}^{\mathbb{D}}(c,t) &: \quad P(M \leq c \mid T > t) \\ & \quad\;\; P(M \leq c \mid N(t) = 0) \end{aligned}$$

$$TP_t^{\mathbb{I}}(c) = P(M > c \mid dN(t) = 1)$$

$$FP_t^{\mathbb{D}}(c) = P(M > c \mid N(t) = 0)$$

[2] Time-dependent ROC Curve

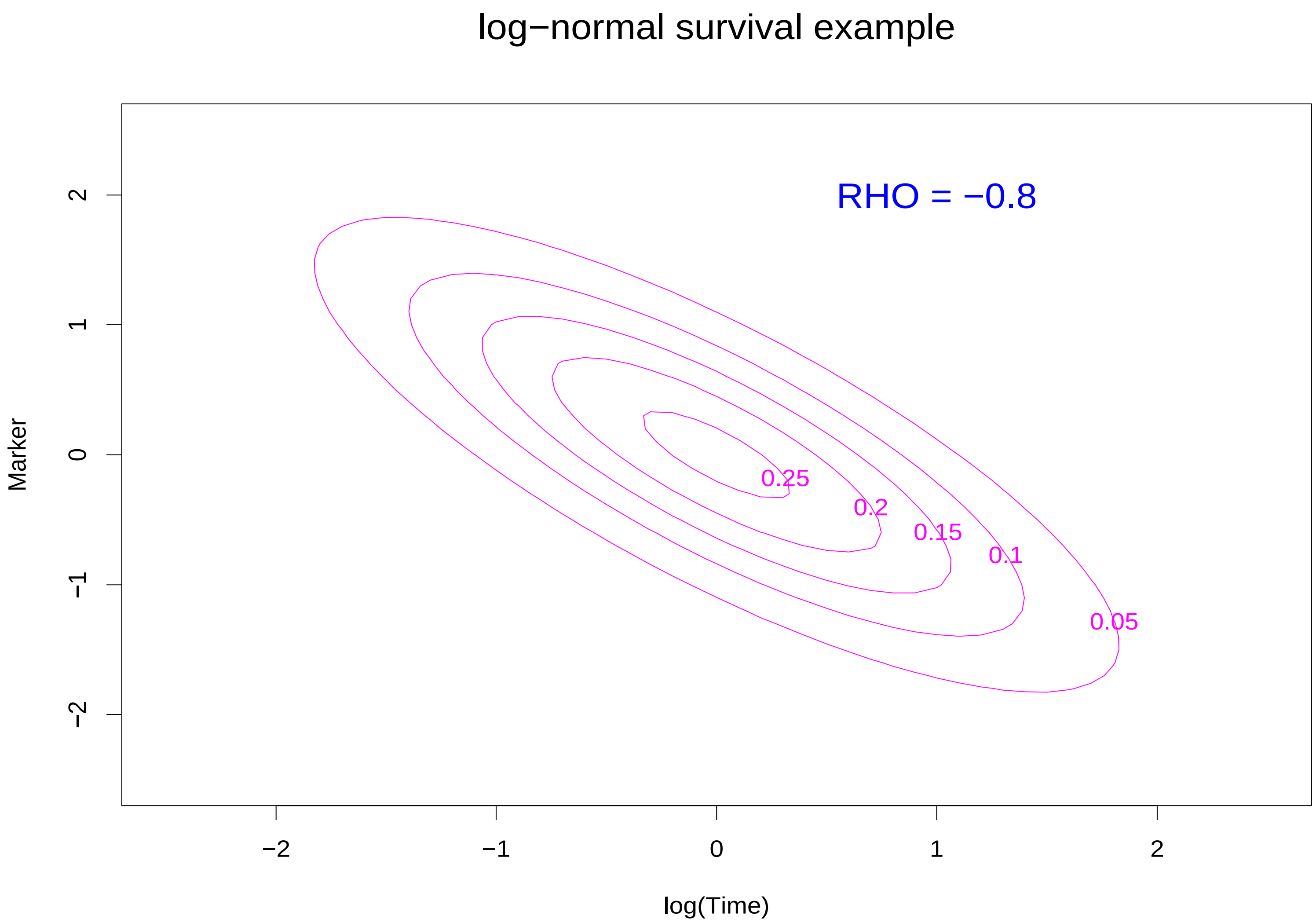
An Incident/Dynamic ROC curve shows the ability of a marker to separate the cases ($T = t$) from the controls ($T > t$) within a (potential) risk-set ($T \geq t$).

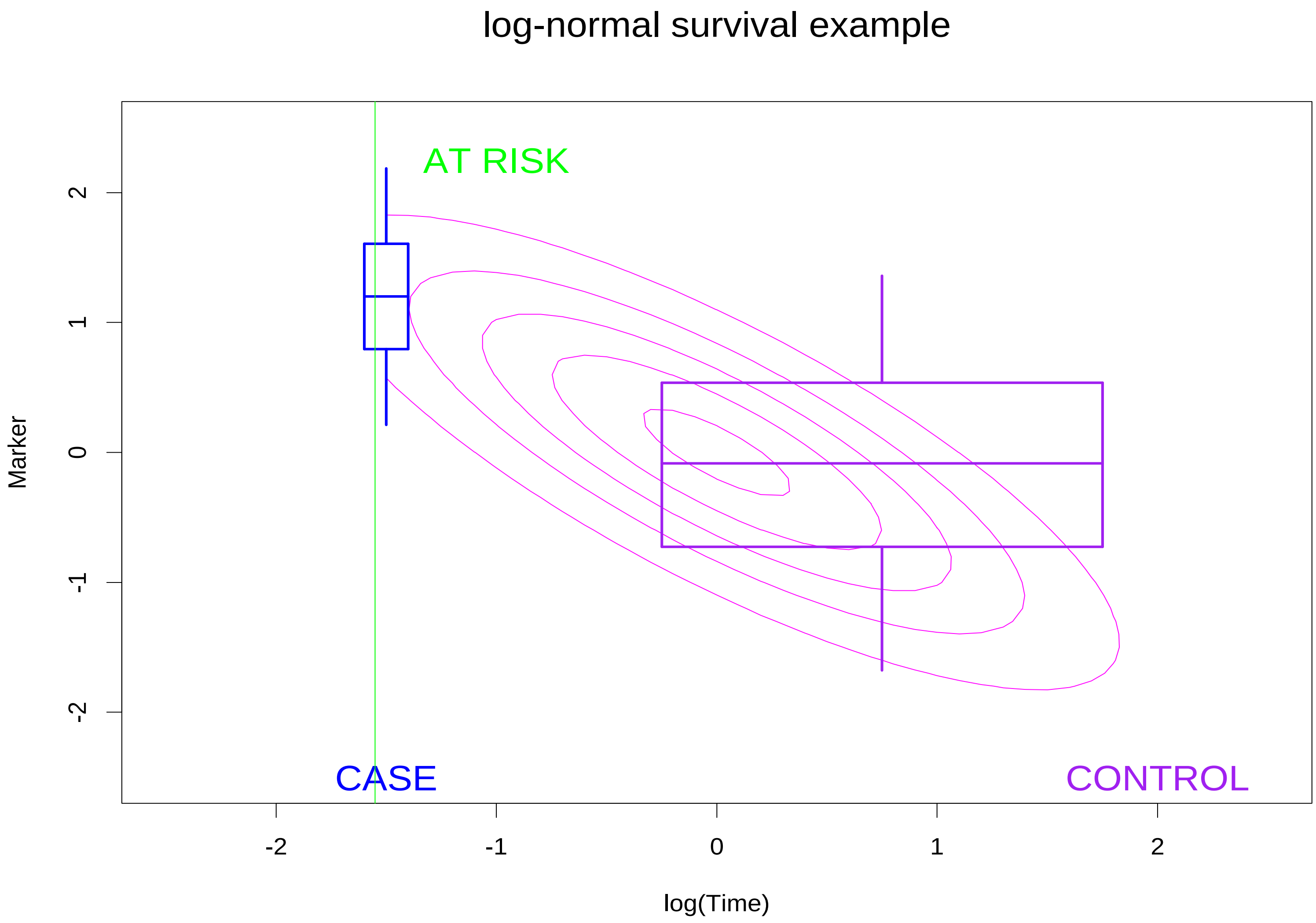
Define curve $[\ p,\ ROC_t^{\mathbb{I}/\mathbb{D}}(p)\ ]$:

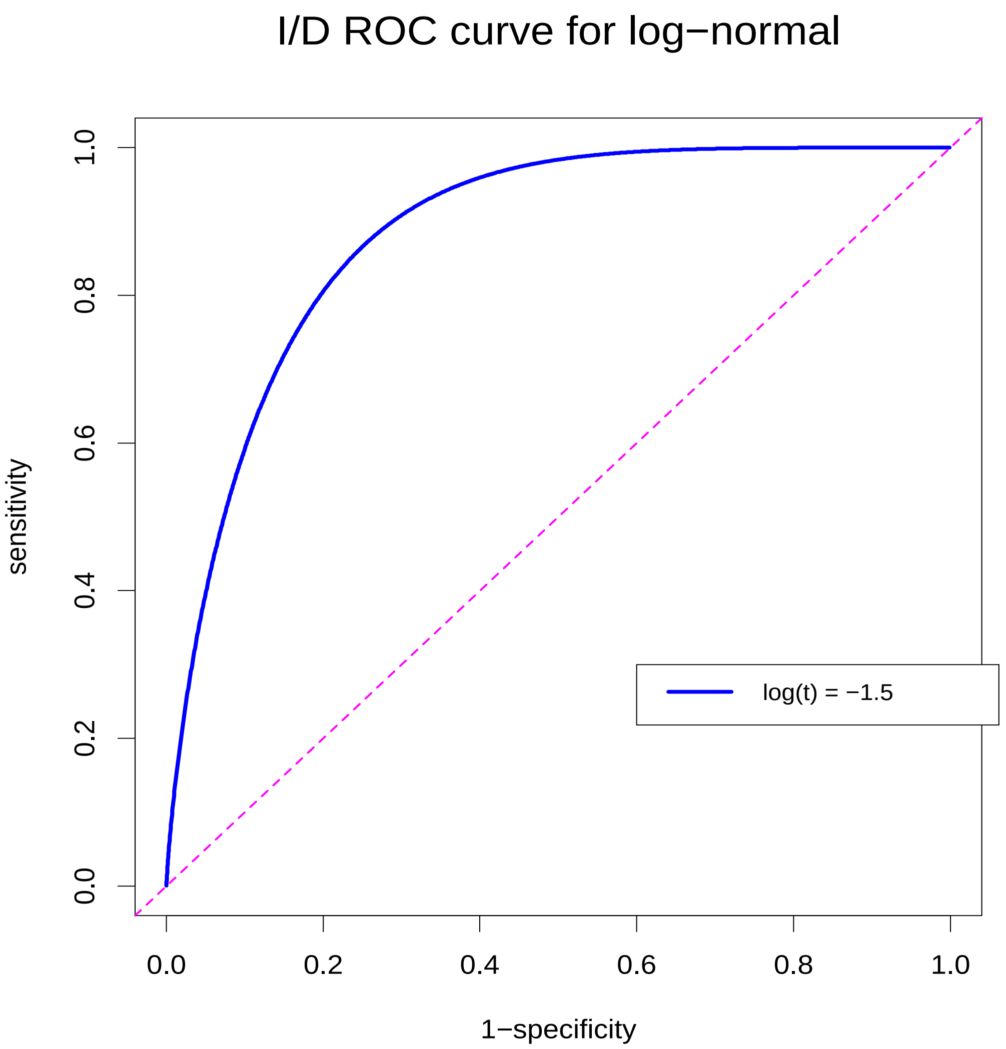
$$
\begin{aligned}
\text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p) &= TP^{\mathbb{I}}\left\{[FP^{\mathbb{D}}]^{-1}(p)\right\} \\
&= TP^{\mathbb{I}}(c^p) \\
\text{where} \quad & \quad c^p \ : \ p = FP^{\mathbb{D}}(c^p)
\end{aligned}
$$

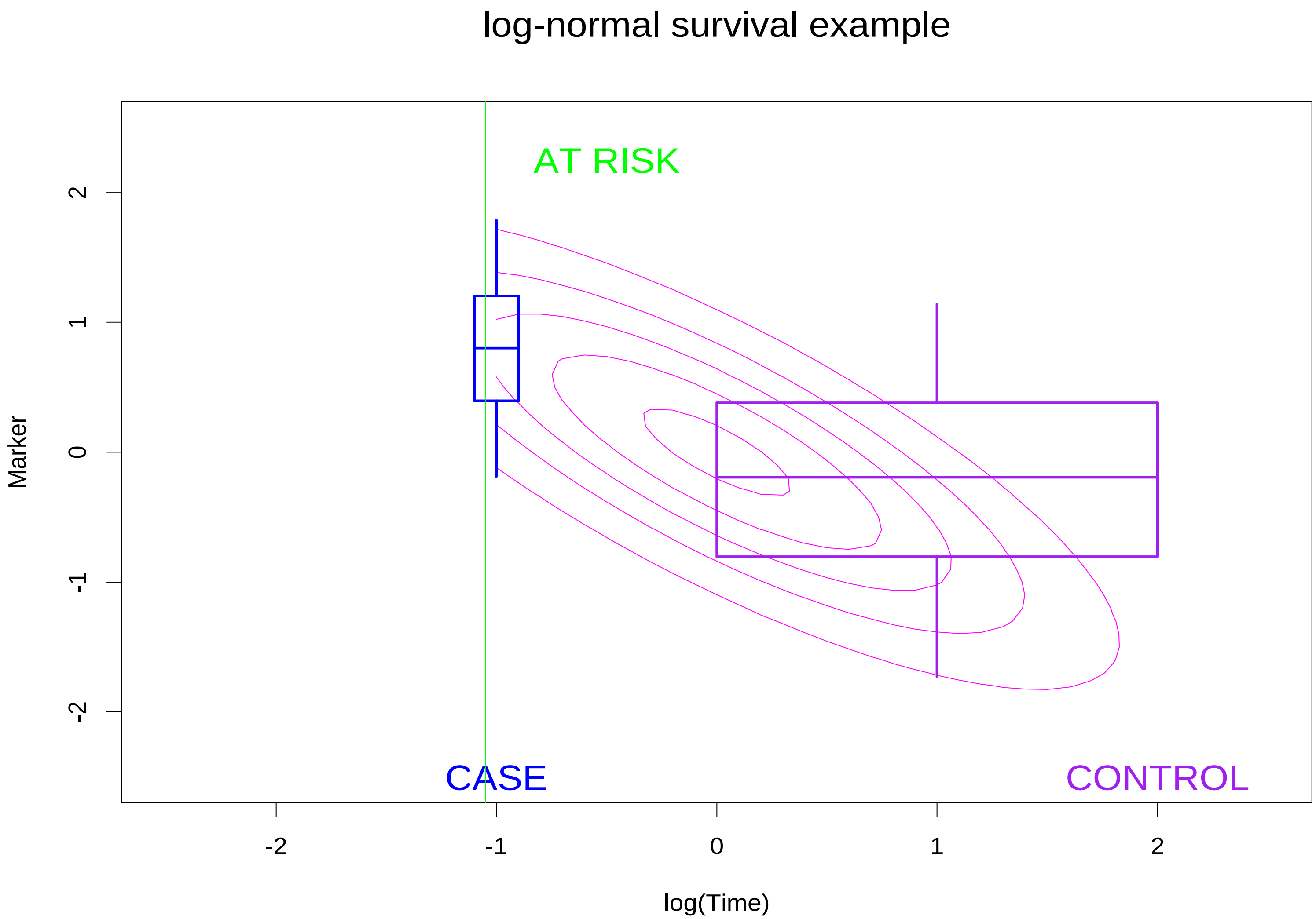
Define AUC as function of time:

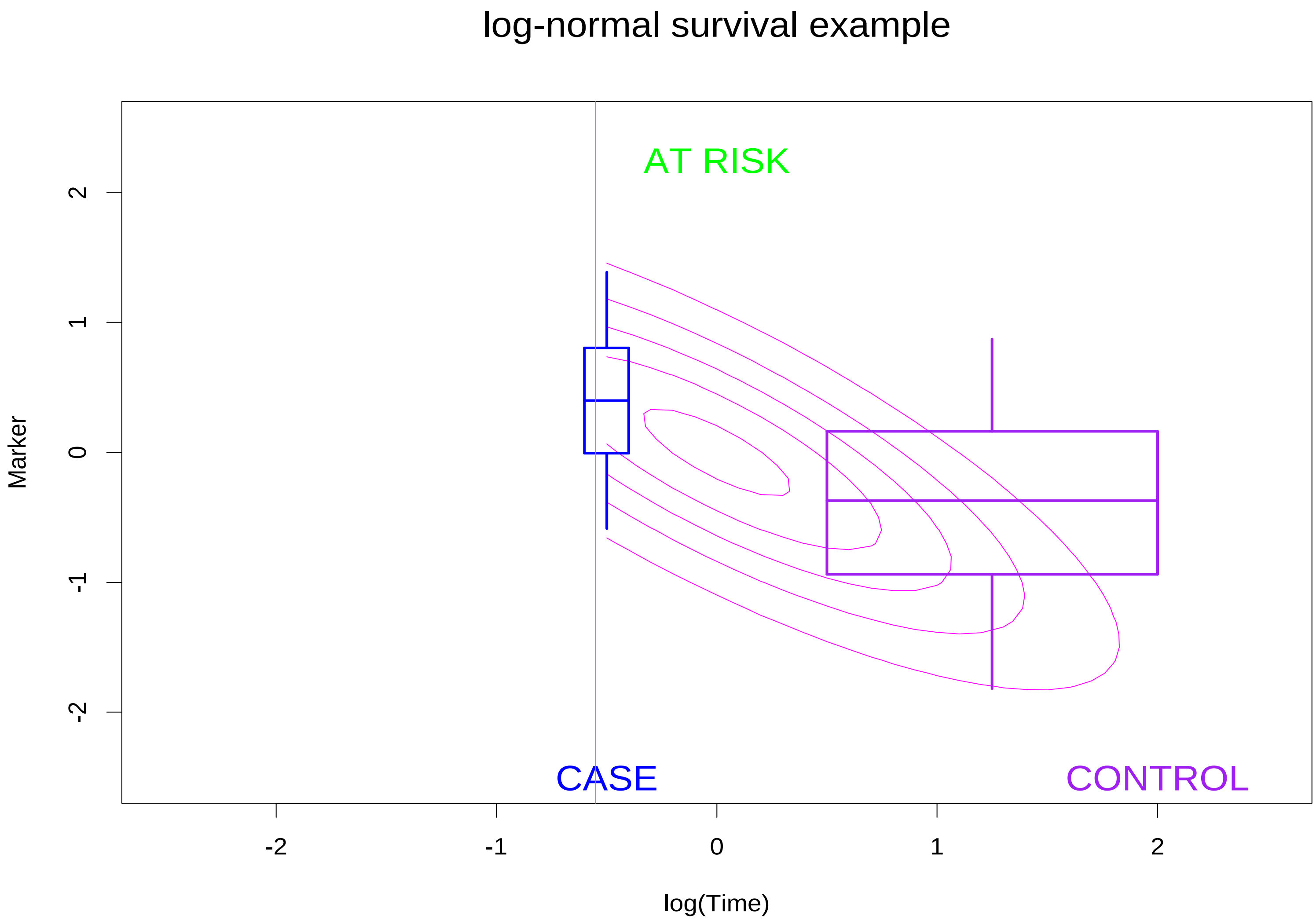
$$
AUC(t) = \int ROC_t^{\mathbb{I}/\mathbb{D}}(p)\, dp
$$

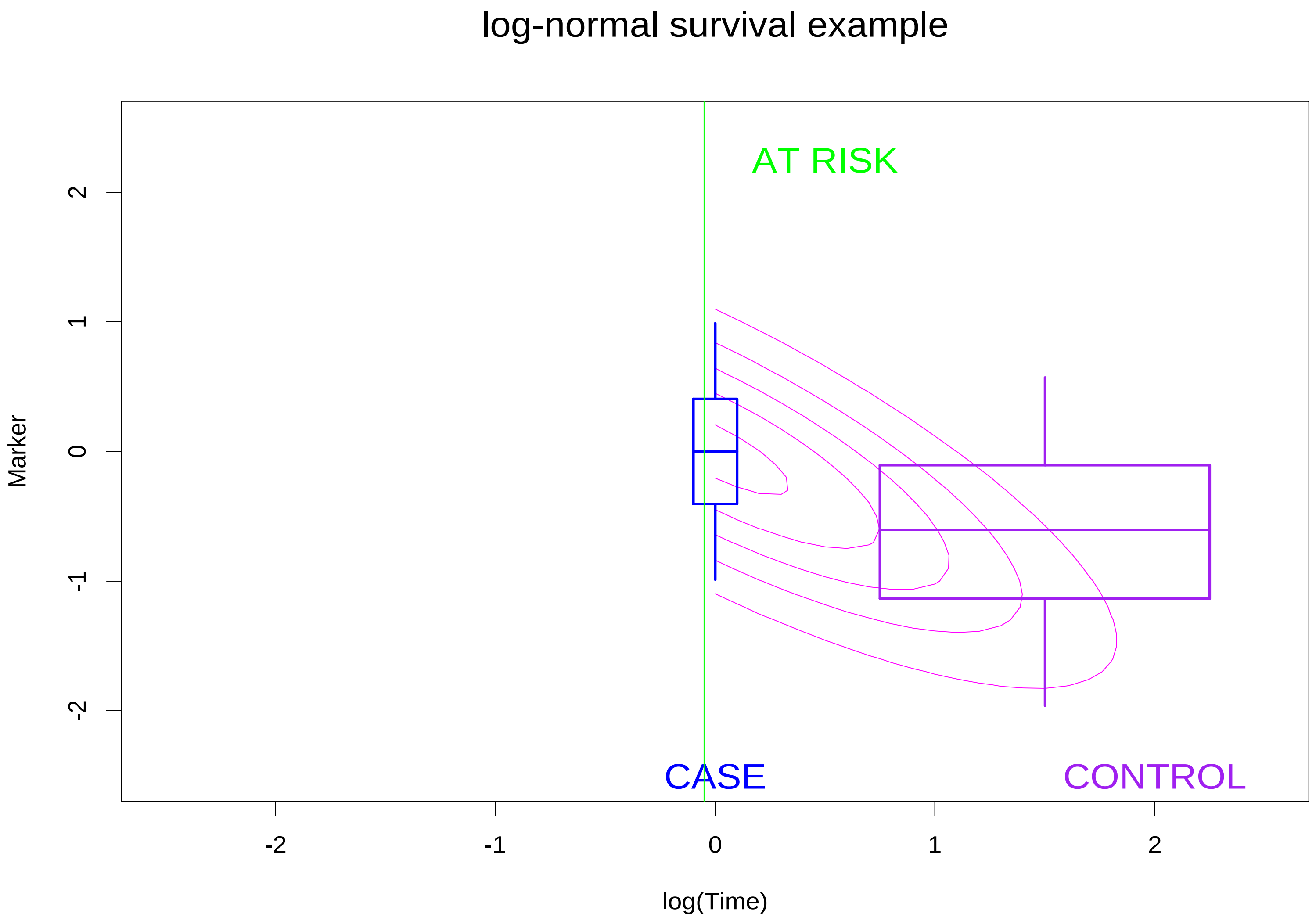

log−normal survival example
RHO = −0.8
0.25
0.2
0.15
0.1
0.05
Marker
log(Time)
−2
−1
0
1
2
−2
−1
0
1
2

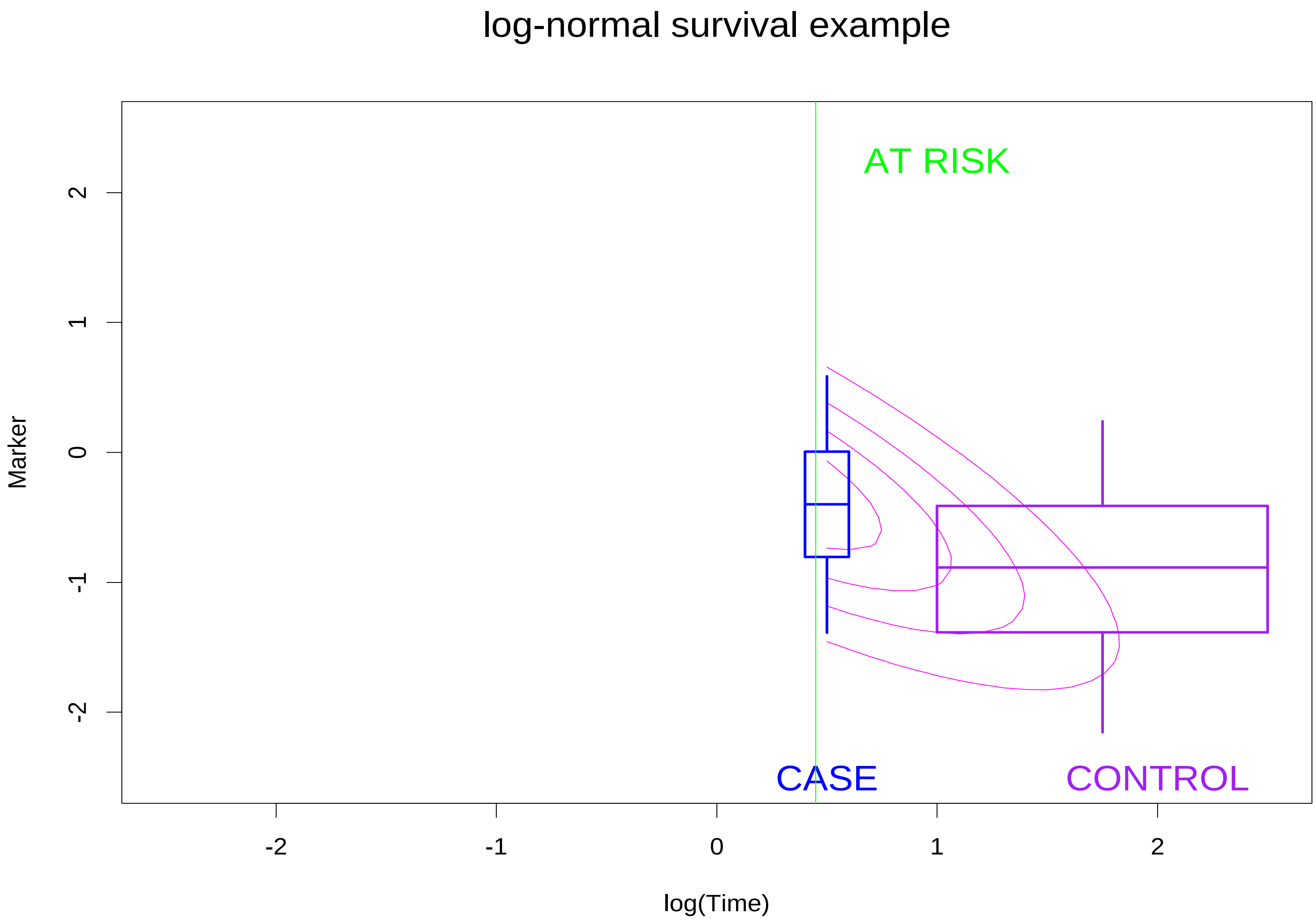

log-normal survival example
AT RISK
CASE
CONTROL
Marker
2
1
0
-1
-2
log(Time)
-2
-1
0
1
2

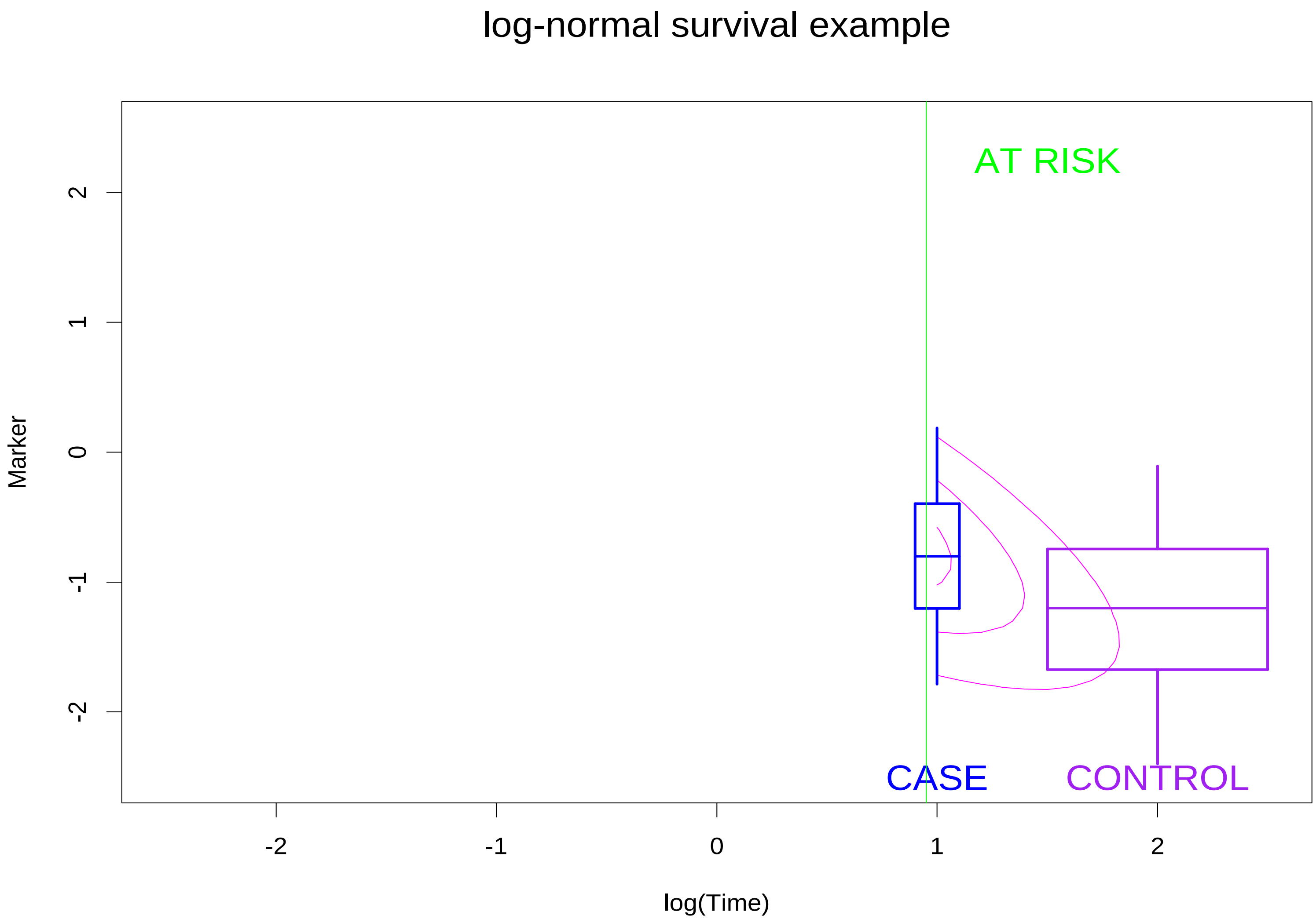

I/D ROC curve for log−normal
sensitivity
1−specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
log(t) = −1.5

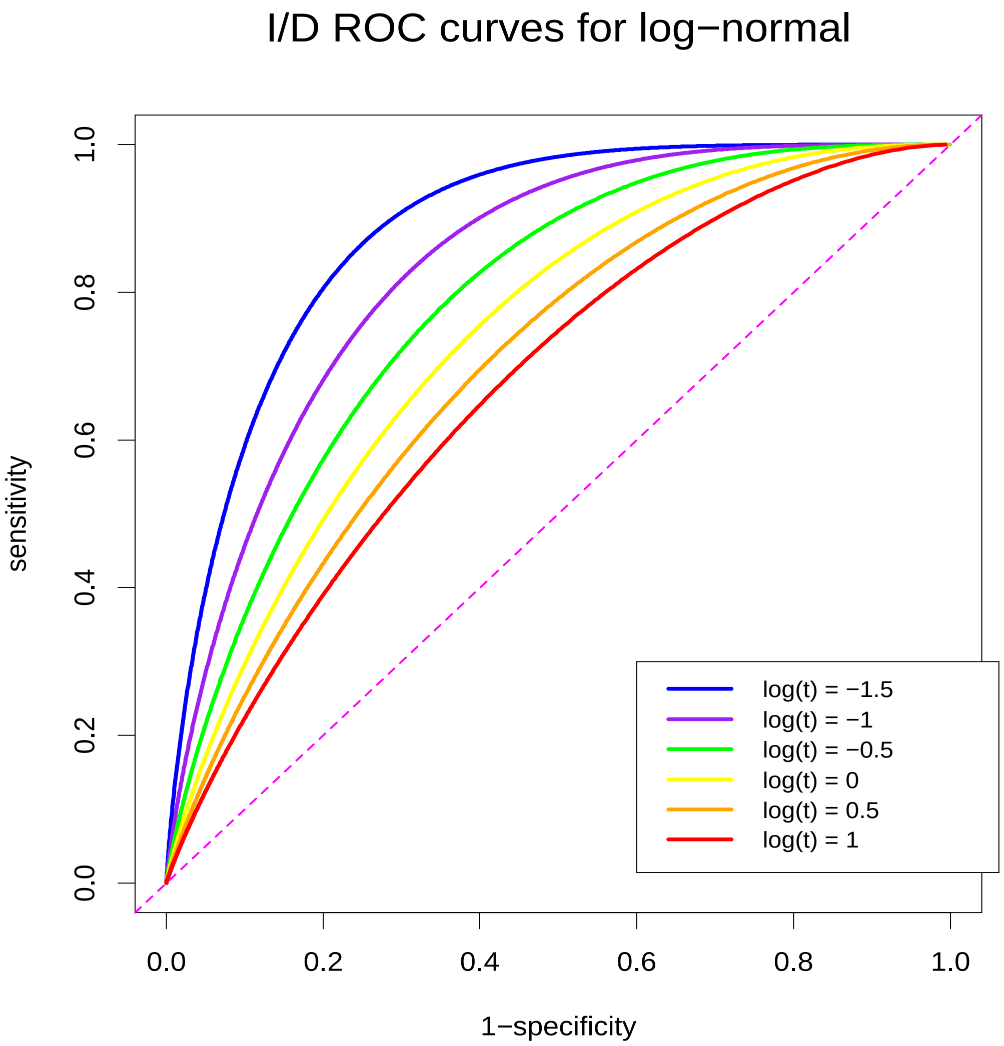

log-normal survival example
AT RISK
CASE
CONTROL
Marker
2
1
0
-1
-2
-2
-1
0
1
2
log(Time)


log-normal survival example
AT RISK
CASE
CONTROL
Marker
log(Time)
-2
-1
0
1
2
-2
-1
0
1
2


log-normal survival example
AT RISK
CASE
CONTROL
Marker
2
1
0
-1
-2
-2
-1
0
1
2
log(Time)


log-normal survival example
AT RISK
CASE
CONTROL
Marker
log(Time)
-2
-1
0
1
2
2
1
0
-1
-2


log-normal survival example
AT RISK
CASE
CONTROL
Marker
2
1
0
-1
-2
-2
-1
0
1
2
log(Time)


I/D ROC curves for log−normal
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1−specificity
0.0
0.2
0.4
0.6
0.8
1.0
log(t) = −1.5
log(t) = −1
log(t) = −0.5
log(t) = 0
log(t) = 0.5
log(t) = 1

AUC(t) and Concordance

Q: The I/D ROC curve and AUC(t) provide time-specific summaries of accuracy, but is there a single global summary?

Concordance:

$$C = P(M_j > M_k \mid T_j < T_k)$$

$$C = \int_t AUC(t) \cdot w(t) \, dt$$

with $w(t) = 2 \cdot f(t) \cdot S(t)$

AUC(t) and Concordance

- Time can be restricted to $(0, \tau)$ to obtain:

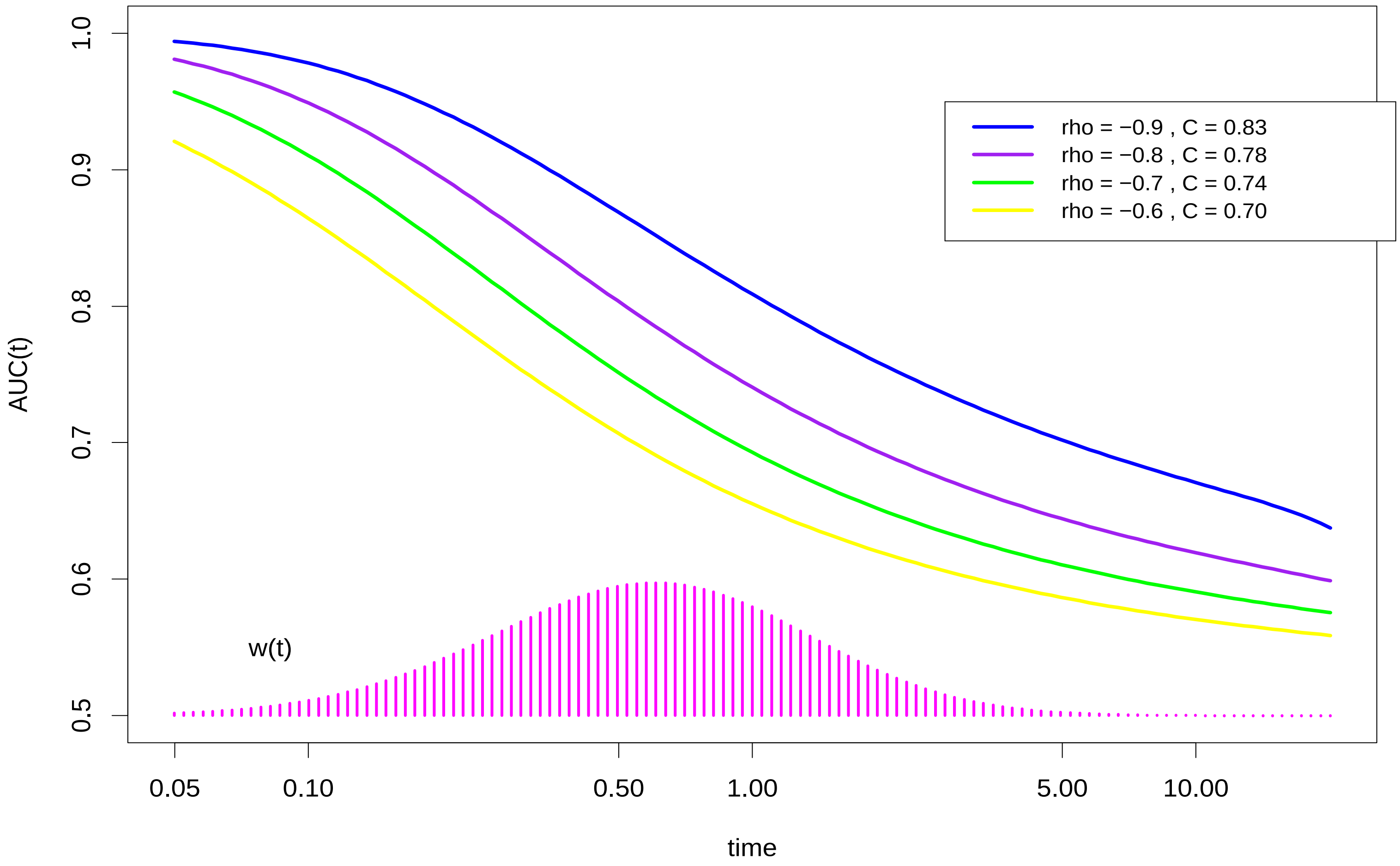
$$\begin{aligned} C^\tau &= P(M_j > M_k \mid T_j < T_k, T_j < \tau) \\ &= \int_0^\tau AUC(t) \cdot w^\tau(t)\, dt \\ &\quad \text{with } w^\tau(t) = w(t)/[1 - S^2(\tau)] \end{aligned}$$

- C is directly related to Kendall's tau, K:

$$C = K/2 + 1/2$$

  - ▷ Korn and Simon (1990)
  - ▷ Harrell, Lee and Mark (1996)

AUC(t) curves for log−normal



Estimation: Issues

- $FP_t^{\mathbb{D}}(c) = P(M > c \mid T > t)$ can be estimated non-parametrically for times when $\sum_i R_i(t)$ moderate-to-large, where $R_i(t) = 1(T_i^* > t)$, “at-risk” indicator.

- However, estimation of $TP_t^{\mathbb{I}}(c) = P(M > c \mid T = t)$ requires some sort of smoothing since the observed subset with $T_i = t$ may only contain one observation.

- Essentially we are interested in regression quantiles for the marker as a function of time, $T = t$, but T may be censored (coarsened covariate).

- The hazard can be used as a “bridge” to estimate $TP_t^{\mathbb{I}}(p)$.

Estimation: Proportional Hazards

Assume:

- (M_i, T_i^*, Δ_i) *iid*

$$T_i^* = \min(T_i, C_i),\ \Delta_i = 1(T_i < C_i).$$

- Independent censoring, C_i.
- No assumption for marker distribution.

Hazard Model:

$$\lambda(t \mid M_i) = \lambda_0(t) \exp(\gamma \cdot M_i)$$

Estimation: Proportional Hazards

To estimate $FP_t^{\mathbb{D}}(c)$ we use the empirical distribution for the “control set” $(T^* > t)$:

$$\widehat{FP}_t^{\mathbb{D}}(c) \quad = \quad \sum_i \frac{1}{n_t} \cdot R_i(t+) \cdot 1(M_i > c)$$

where $\qquad R_i(t+) = 1(T_i^* > t), \;\; n_t = \sum_i R_i(t+)$

Estimation: Proportional Hazards

To estimate $TP_t^{\mathbb{I}}(c)$ we use the "exponential tilt" of the empirical distribution for the risk set $(T^* \geq t)$:

$$\widehat{TP}_t^{\mathbb{I}}(c) \quad = \quad \sum_i \pi_i(t,\gamma) \cdot 1(M_i > c)$$

where
$$\pi_i(t,\gamma) = R_i(t) \cdot \exp(\gamma \cdot M_i)/W_t$$
$$W_t = \sum_i R_i(t) \cdot \exp(\gamma \cdot M_i), \quad R_i(t) = 1(T_i^* \geq t)$$

- Xu and O'Quigley (2000) show that the weights, $\pi_i(t,\gamma)$, applied to the risk set provide consistent estimation of $P(M_i > c \mid T_i = t)$.

- Partial likelihood connections: $E(M \mid T = t)$.

Estimation: Hazard as Bridge

A general definition for the hazard is

$$\lambda(t \mid M_i) = \frac{P(T_i = t \mid M_i)}{P(T_i \geq t \mid M_i)}$$

Then using a little algebra yields

$$P(M_i = m \mid T_i = t) \quad \propto \quad \lambda(t \mid M_i = m) \cdot P(M_i = m \mid T_i \geq t)$$

Estimation: Hazard as Bridge

A general definition for the hazard is

$$\lambda(t \mid M_i) = \frac{P(T_i = t \mid M_i)}{P(T_i \geq t \mid M_i)}$$

Then using a little algebra yields

$$\begin{array}{ccccc} P(M_i = m \mid T_i = t) & \propto & \underbrace{\lambda(t \mid M_i = m)} \cdot \underbrace{P(M_i = m \mid T_i \geq t)} \\ \text{Estimate} & \Longleftarrow & \text{Smooth model} \; + \; \text{Empirical} \end{array}$$

Estimation: non-PH

Estimation assuming PH can be relaxed using a varying coefficient model:

$$\lambda(t \mid M_i) = \lambda_0(t) \exp[\gamma(t) \cdot M_i]$$

- Estimation of $\gamma(t)$
  - ▷ Hastie and Tibshirani (1993)
  - ▷ Cai and Sun (2003) [local linear MPLE]
  - ▷ simple smoothing of scaled Schoenfeld residuals
- Only assumes smooth hazard ratios, and linearity in M.
- Linearity in M can be relaxed using functions $f(M)$.

Estimation: non-PH

Only the estimate of $TP_t^{\mathbb{I}}(c)$ is modified:

$$\widehat{TP}_t^{\mathbb{I}}(c) \quad = \quad \sum_i \pi_i[t, \gamma(t)] \cdot 1(M_i > c)$$

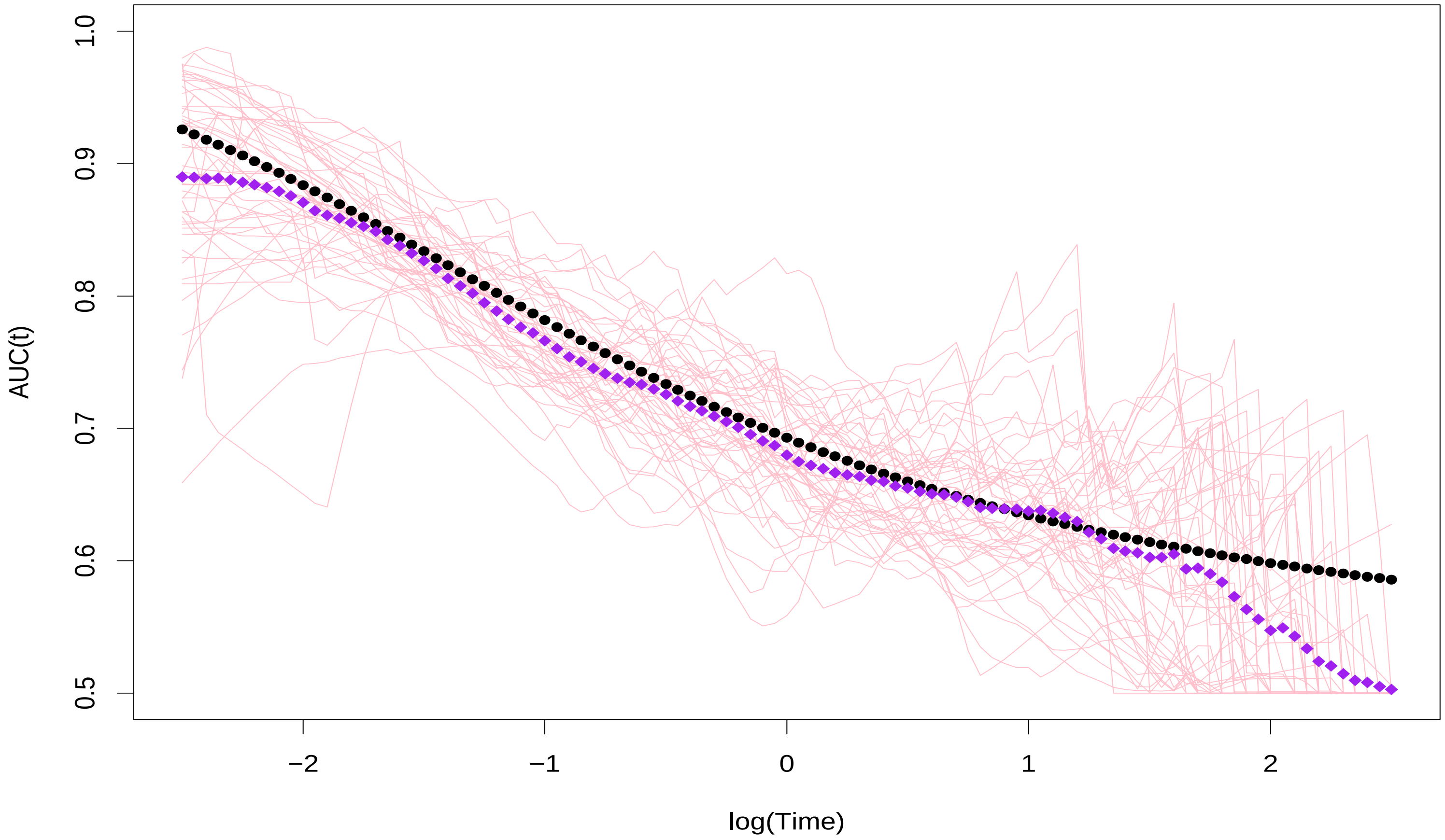
where $$\pi_i[t, \gamma(t)] = R_i(t) \cdot \exp[\gamma(t) \cdot M_i] / W_t$$

$$W_t = \sum_i R_i(t) \cdot \exp[\gamma(t) \cdot M_i]$$

Some Simulations

- Data $(M_i, \log T_i)$ were generated as bivariate normal with a correlation of $\rho = -0.7$.
- The sample size for each simulated data set was $N = 200$.
- The $AUC(t)$ curve and the integrated curve, C^τ, was estimated using:
  - ▷ maximum likelihood assuming a bivariate normal model
  - ▷ Cox model which assumes proportional hazards
  - ▷ local maximum partial likelihood for the varying-coefficient model $\lambda(t) = \lambda_0(t) \exp[\gamma(t) \cdot M_i]$
  - ▷ local linear smooth of the scaled Schoenfeld residuals to estimate the varying-coefficient model.

AUC(t) Estimated Using Local-linear MPL



40% Censoring

		MLE		Cox model		local MPLE		residual smooth	
log time	$AUC(t)$	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
-2.0	0.884	0.884	(0.019)	0.749	(0.031)	0.859	(0.054)	0.875	(0.048)
-1.5	0.833	0.834	(0.021)	0.742	(0.029)	0.818	(0.035)	0.827	(0.037)
-1.0	0.782	0.782	(0.021)	0.732	(0.026)	0.770	(0.035)	0.772	(0.035)
-0.5	0.734	0.734	(0.020)	0.722	(0.024)	0.724	(0.038)	0.722	(0.039)
0.0	0.693	0.693	(0.019)	0.712	(0.024)	0.689	(0.042)	0.687	(0.041)
0.5	0.660	0.660	(0.018)	0.702	(0.026)	0.654	(0.045)	0.655	(0.043)
1.0	0.634	0.635	(0.016)	0.689	(0.035)	0.633	(0.057)	0.637	(0.048)
1.5	0.614	0.614	(0.015)	0.653	(0.055)	0.617	(0.075)	0.614	(0.051)
2.0	0.598	0.599	(0.013)	0.560	(0.073)	0.555	(0.075)	0.546	(0.058)
C^{τ}	0.741	0.741	(0.017)	0.727	(0.022)	0.740	(0.021)	0.742	(0.021)

Illustration: PBC Data

- Marker, M_i, derived as linear predictor from Cox model:

Covariate	estimate	s.e.	Z
log(bilirubin)	0.877	0.099	8.87
log(prothrombin time)	3.015	1.033	2.94
edema	0.785	0.300	2.62
albumin	-0.944	0.237	-3.99
age	0.092	0.024	3.88

- Accuracy evaluated using $\lambda_0(t) \exp[\gamma(t) \cdot M_i]$
- local-linear MPLE for $\gamma(t)$
- (5) predictor model compared to (4) predictor excluding bilirubin

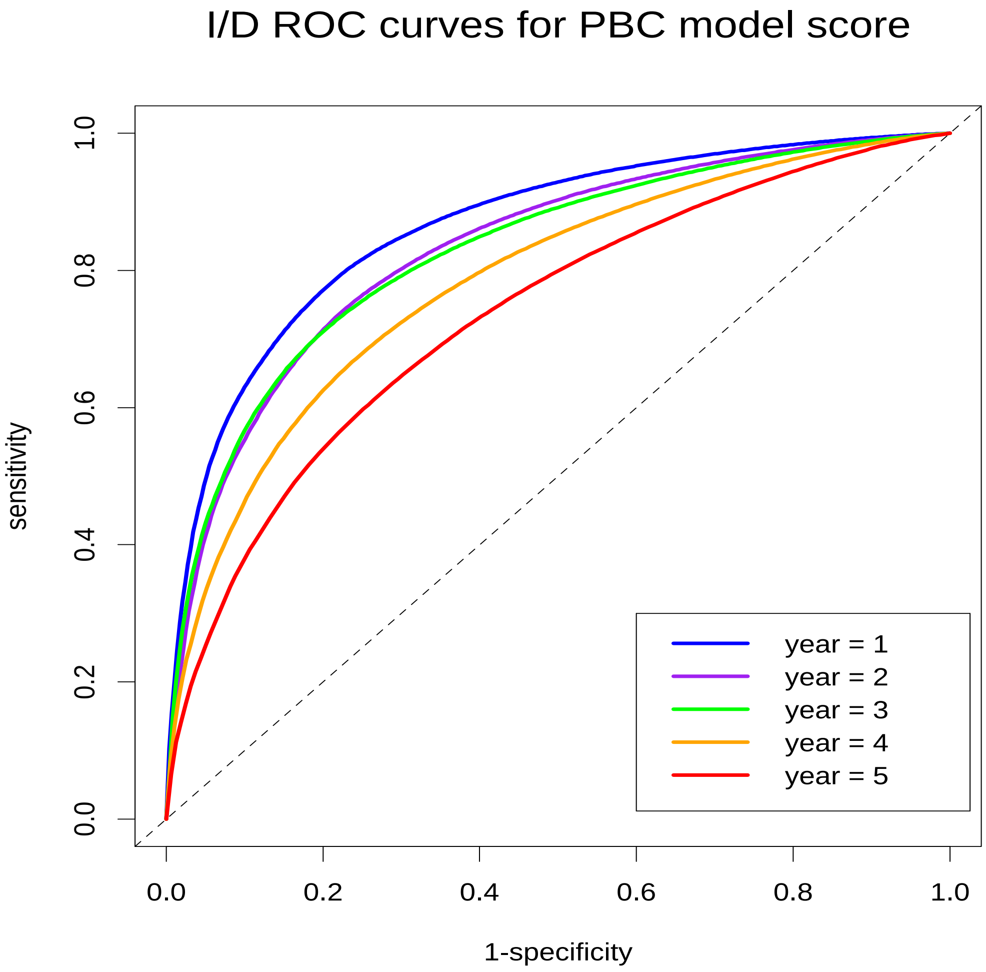
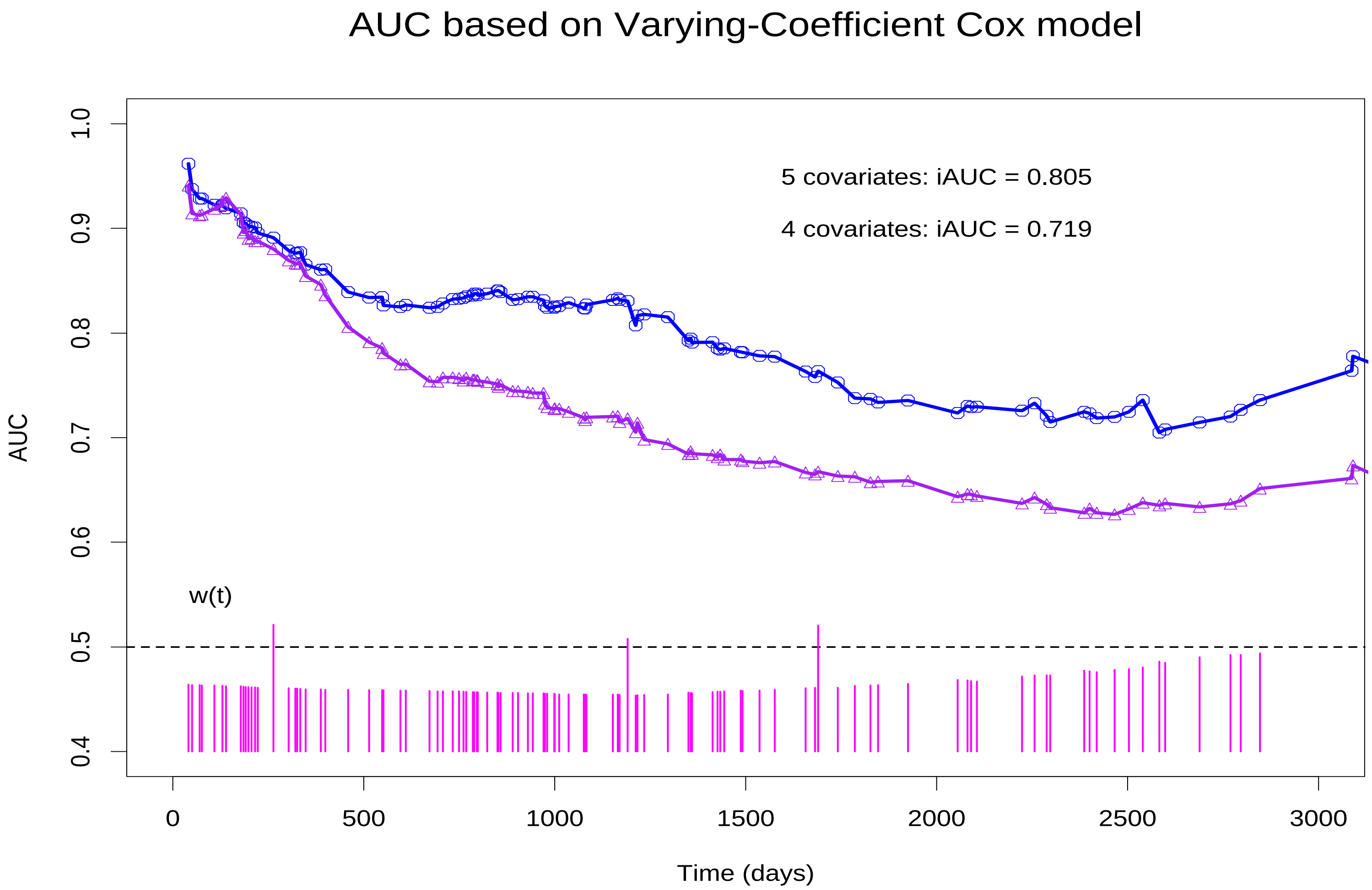

I/D ROC curves for PBC model score
sensitivity
1-specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
year = 1
year = 2
year = 3
year = 4
year = 5

Illustration: PBC Data



Summary

- Extension of ROC concepts to risk sets.
- Estimation based on hazard model.
- Varying coefficient – simple methods look promising.
- All summaries can be obtained from routine Cox model output.
- Criterion for marker and/or model comparison.
- Separation of marker generation and marker evaluation.
- Note: R^2 for PBC data estimated as 0.32 (max possible 0.98)

- Heagerty and Zheng (2005) "Survival Model Predictive Accuracy and ROC Curves" *Biometrics*

Extension to Longitudinal Markers

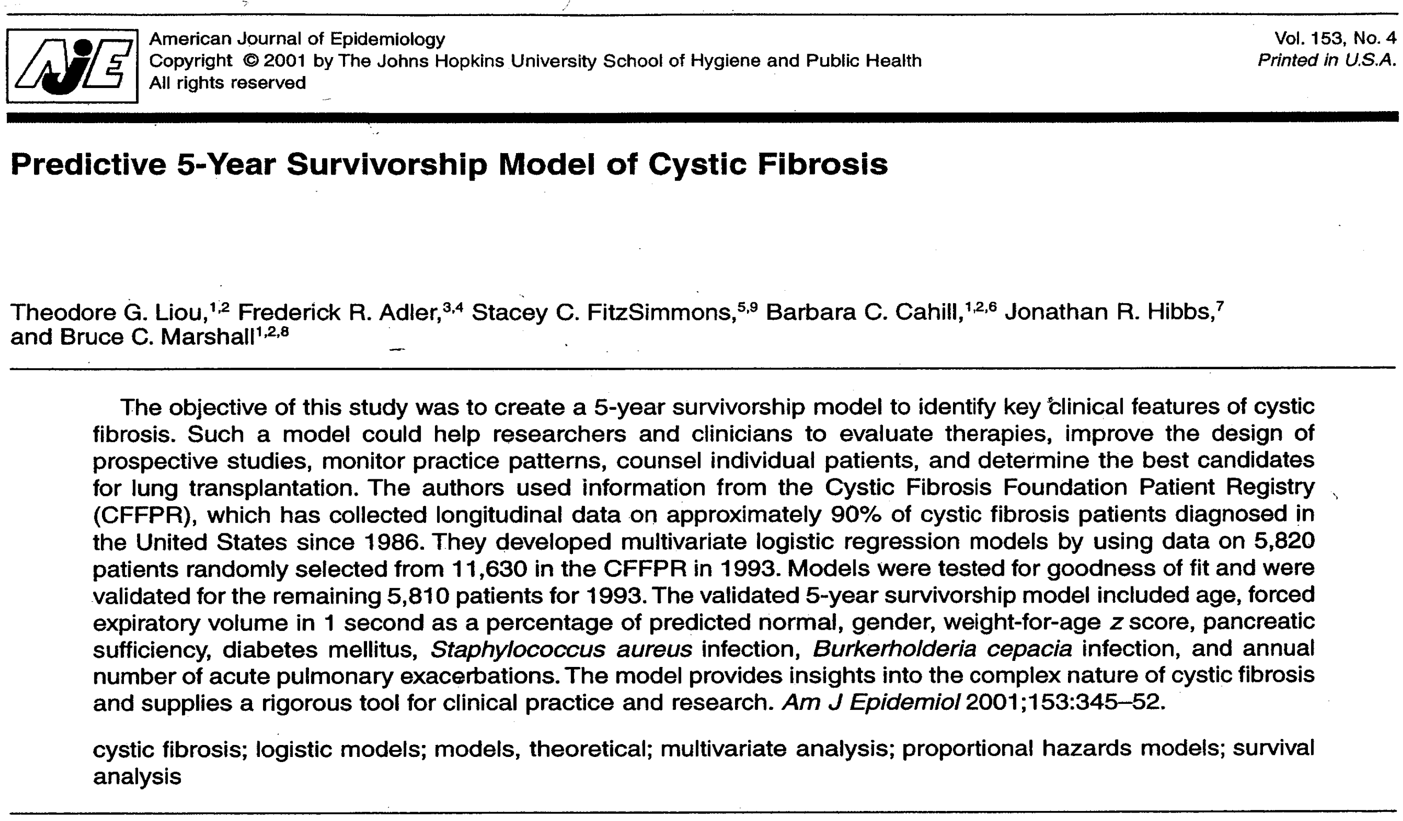
- TP, FP based on longitudinal marker

$$\begin{aligned} TP_t^{\mathbb{I}}(c) &= P[M(t) > c \mid T = t] \\ FP_t^{\mathbb{D}}(c) &= P[M(t) > c \mid T > t] \end{aligned}$$

- No simple "concordance", but $\int AUC(t) \cdot w(t)dt$
- Dynamic criterion, $c^p(t)$, controls specificity

$$c^p(t) \quad : \quad p = P[M(t) > c^p(t) \mid T > t]$$

- Summary ROC curve shows total percent test positive when controlling FP using dynamic criterion.



American Journal of Epidemiology
Copyright © 2001 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol. 153, No. 4
Printed in U.S.A.

Predictive 5-Year Survivorship Model of Cystic Fibrosis

Theodore G. Liou,[1,2] Frederick R. Adler,[3,4] Stacey C. FitzSimmons,[5,9] Barbara C. Cahill,[1,2,6] Jonathan R. Hibbs,[7]
and Bruce C. Marshall[1,2,8]

The objective of this study was to create a 5-year survivorship model to identify key clinical features of cystic fibrosis. Such a model could help researchers and clinicians to evaluate therapies, improve the design of prospective studies, monitor practice patterns, counsel individual patients, and determine the best candidates for lung transplantation. The authors used information from the Cystic Fibrosis Foundation Patient Registry (CFFPR), which has collected longitudinal data on approximately 90% of cystic fibrosis patients diagnosed in the United States since 1986. They developed multivariate logistic regression models by using data on 5,820 patients randomly selected from 11,630 in the CFFPR in 1993. Models were tested for goodness of fit and were validated for the remaining 5,810 patients for 1993. The validated 5-year survivorship model included age, forced expiratory volume in 1 second as a percentage of predicted normal, gender, weight-for-age *z* score, pancreatic sufficiency, diabetes mellitus, *Staphylococcus aureus* infection, *Burkerholderia cepacia* infection, and annual number of acute pulmonary exacerbations. The model provides insights into the complex nature of cystic fibrosis and supplies a rigorous tool for clinical practice and research. *Am J Epidemiol* 2001;153:345–52.

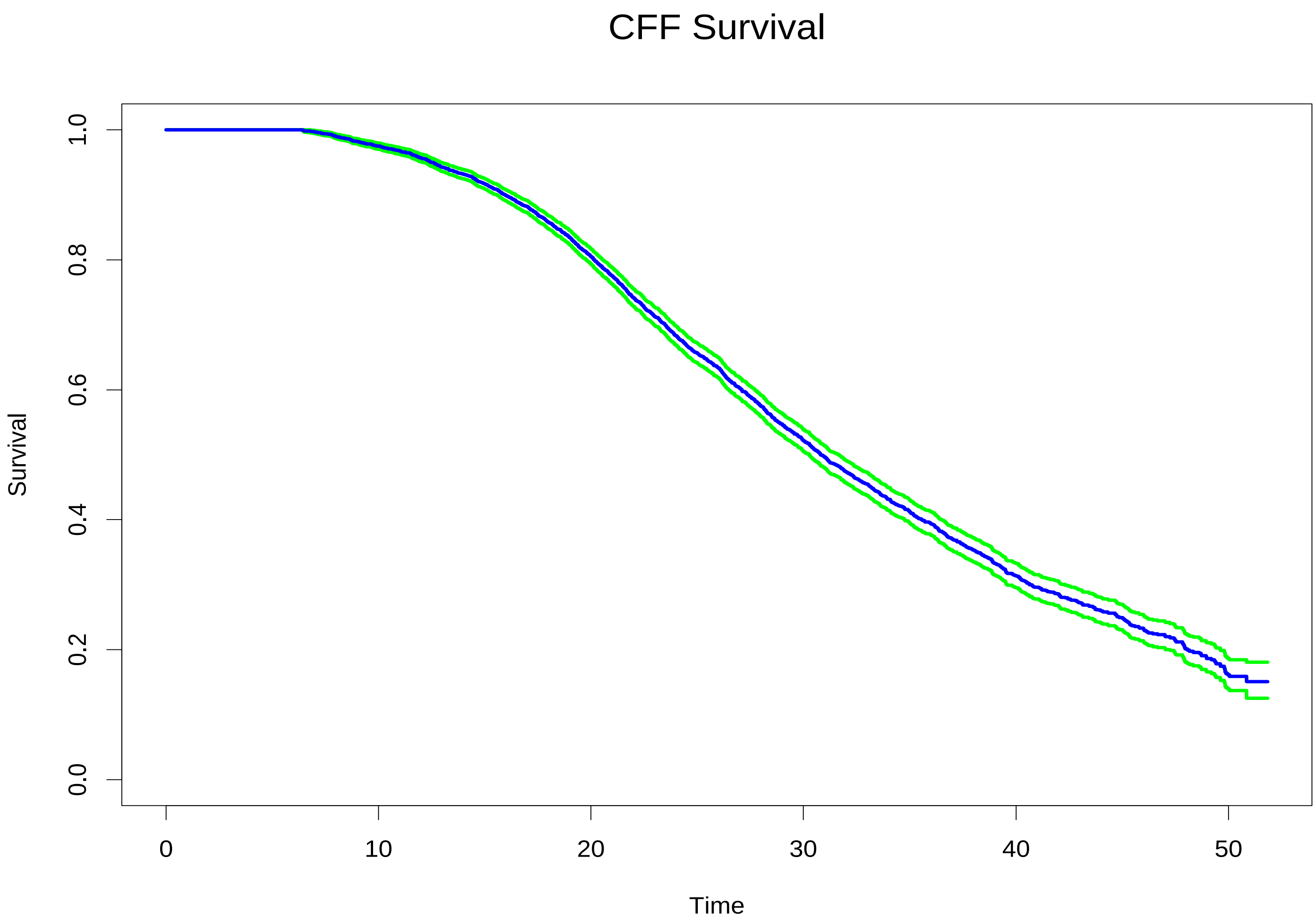
cystic fibrosis; logistic models; models, theoretical; multivariate analysis; proportional hazards models; survival analysis

Illustration: CFF Data and Longitudinal Markers

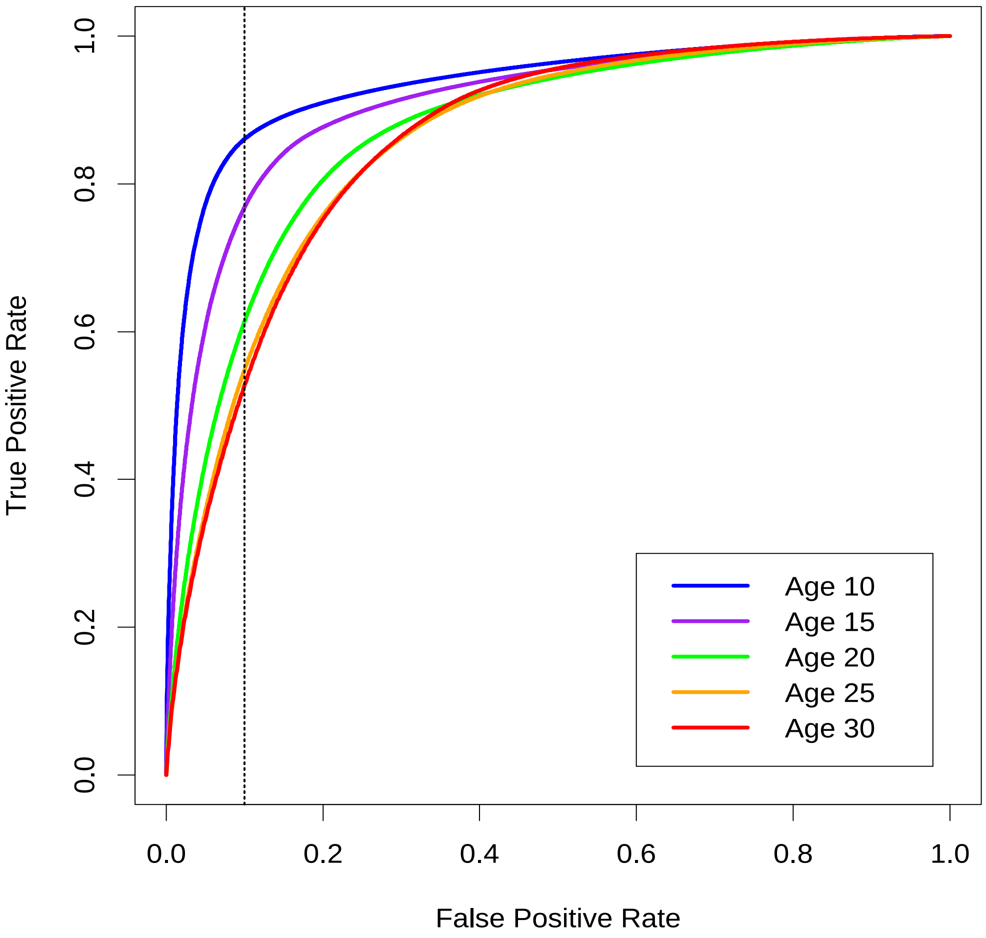
- Split sample (development / validation)
- Time-dependent covariate Cox model

Covariate	estimate	s.e.	Z
fev1	-0.0868	0.0022	-40.03
fev1(T2)	0.0535	0.0062	8.65
fev1(T3)	0.0282	0.0118	2.40
gender	0.1235	0.0459	2.69
weightZ	-0.4334	0.0382	-11.34
heightZ	-0.0603	0.0299	-2.02

- Accuracy evaluated using $\lambda_0(t)\exp[\gamma(t)\cdot M_i(t)]$
- Measurement spacing: median $=$ 1.00, Q1 $=$ 0.87, Q3 $=$ 1.22


CFF Survival
Survival
Time
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

Incident / Dynamic ROC for CFF Data



Summary Survival ROC Curve

Dynamic criterion Saha and Heagerty (manuscript)

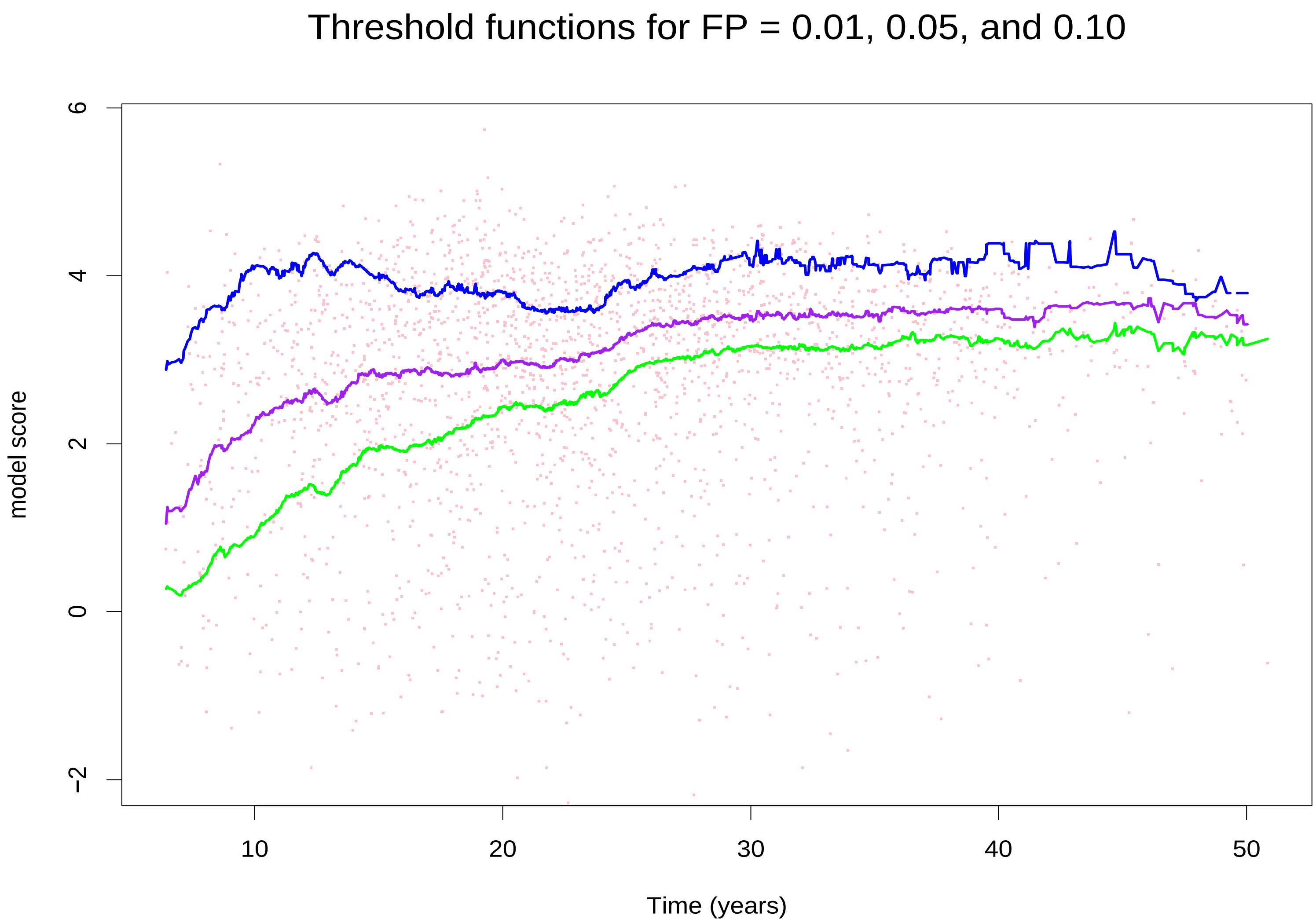
$$c^p(t) \quad : \quad p = P[M(t) > c^p(t) \mid T > t]$$

Q: If a fixed time-dependent FP rate of p is used then what percent of cases will test positive at the appropriate time?

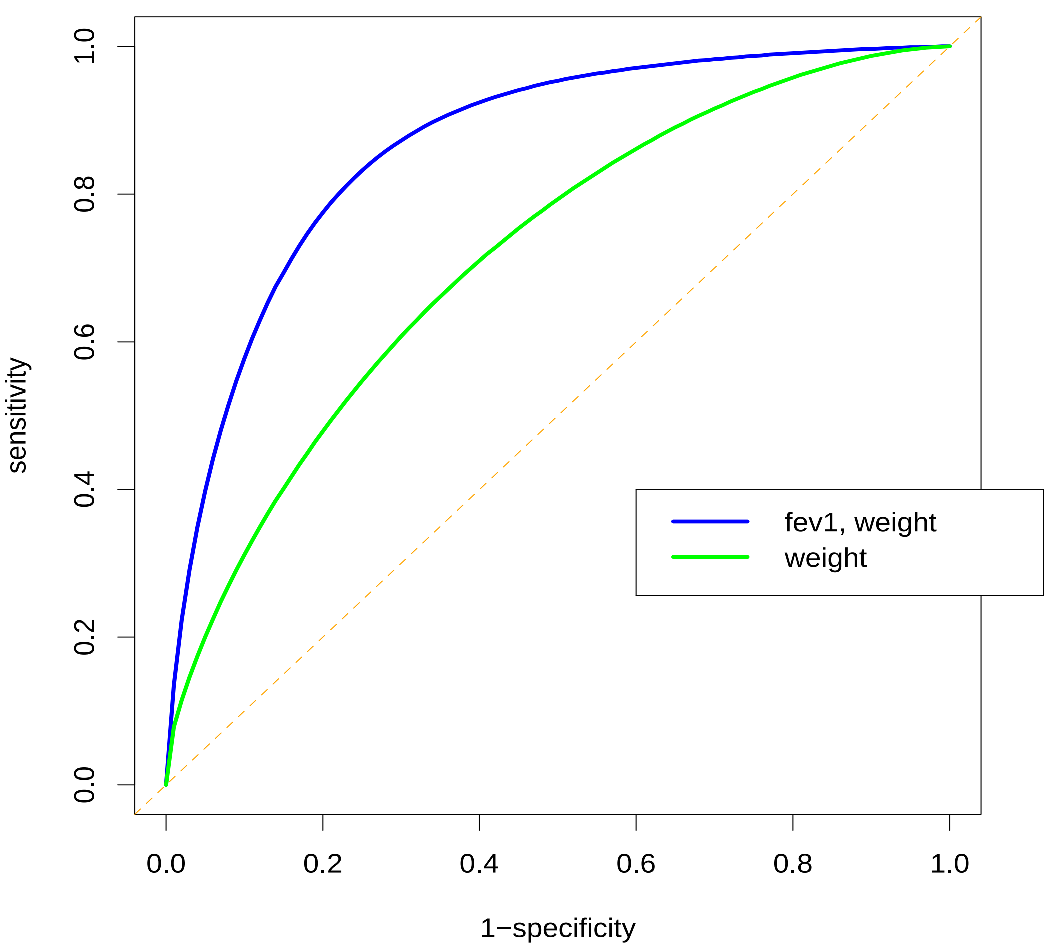
$$\text{Total Sensitivity} \quad = \quad \int_t P[M(t) > c^p(t) \mid T = t] \cdot P[T = t] dt$$

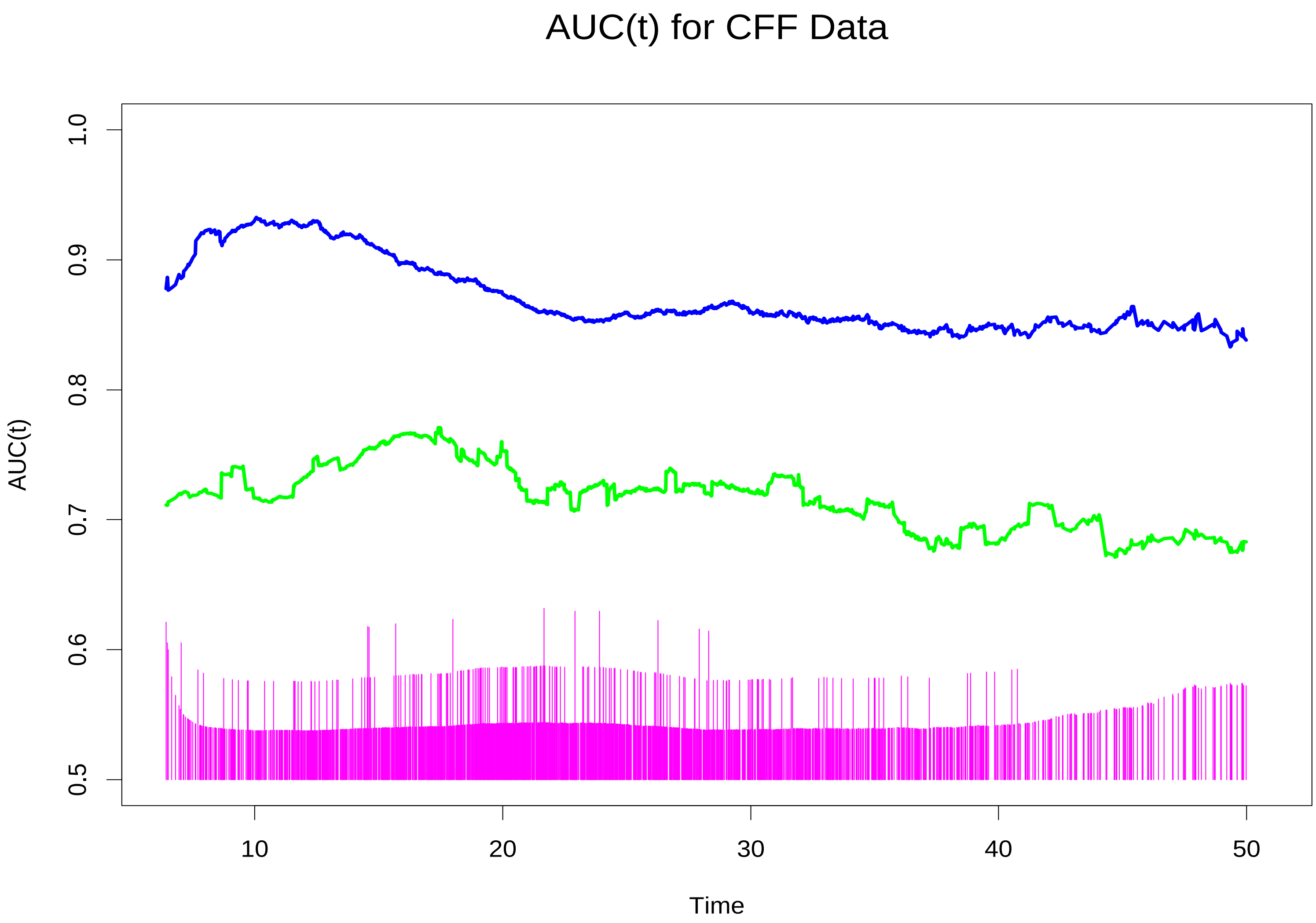
$$TTP(p) \; = \; \overline{ROC}(p) \quad = \quad E_T \left[ROC_T^{\mathbb{I}/\mathbb{D}}(p) \right]$$

Note: Consider a time lag, L, such that the marker used to model the hazard is $M(t-L)$. This leads to "test positive L units before failure."


Threshold functions for FP = 0.01, 0.05, and 0.10
model score
6
4
2
0
−2
10
20
30
40
50
Time (years)

Summary Survival ROC for CFF Data




AUC(t) for CFF Data
AUC(t)
1.0
0.9
0.8
0.7
0.6
0.5
10
20
30
40
50
Time

Estimation for Summary ROC Curve

- $\text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)$ estimated using semi-parametric methods of Heagerty and Zheng (2005)

- Summary curve then averages time-dependent ROC curves with respect to the distribution of times, T_i, estimated using standard non-parametric methods (Kaplan-Meier).

- Inference for $TTP(p)$ using sum over event times and appropriate CLT.

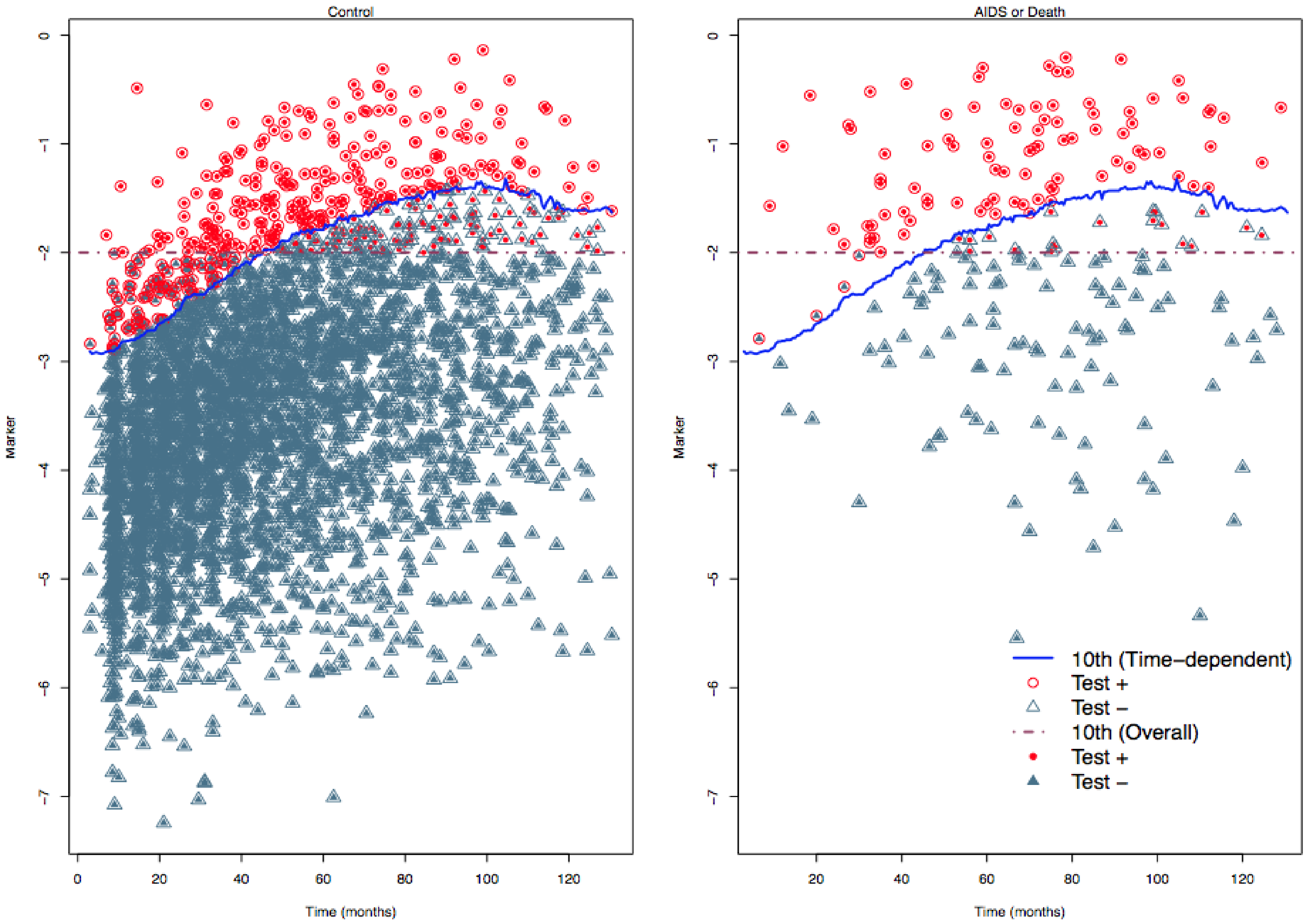
- One approach to comparison of marker A to marker B is:

$$rTTP(p) = \frac{TTP_A(p)}{TTP_B(p)}$$

- **Q**: interpret AUC here?

Illustration using MACS Data

- Kaslow et al. (1987)
- Sero-negative men at baseline: 3,426
- Observed to seroconvert: 479
- Observed events: 176 AIDS, 34 died before AIDS
- Candidate markers: longitudinal CD4 and CD8
- Composite marker: Cox model using sum and diff for CD4 and CD8 (e.g. four predictors)


Control
AIDS or Death
Marker
0
-1
-2
-3
-4
-5
-6
-7
Time (months)
0
20
40
60
80
100
120
10th (Time−dependent)
Test +
Test −
10th (Overall)
Test +
Test −

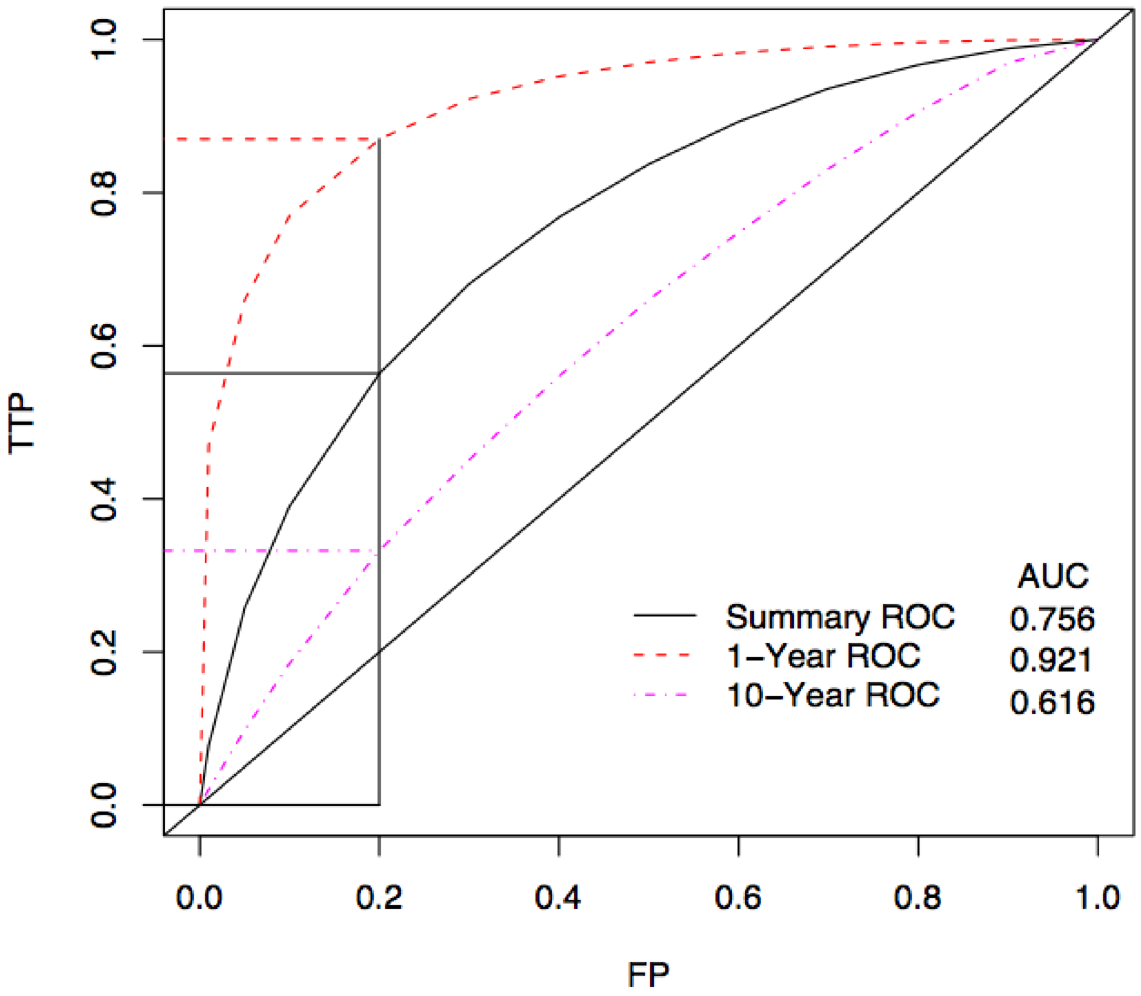


Figure 2. Summary survival ROC curve and 1-year and 10-year ROC curves for a composite marker from MACS data.

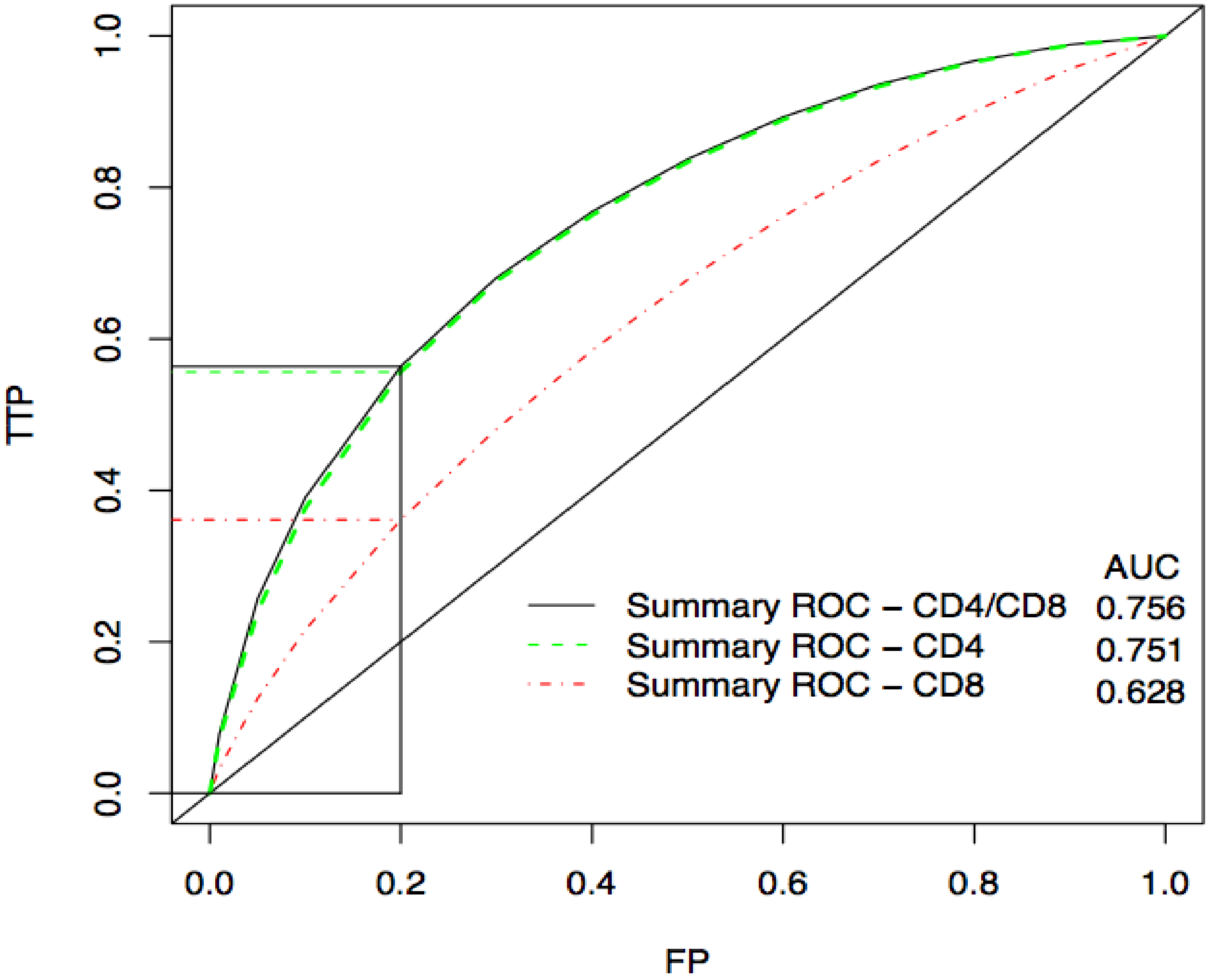
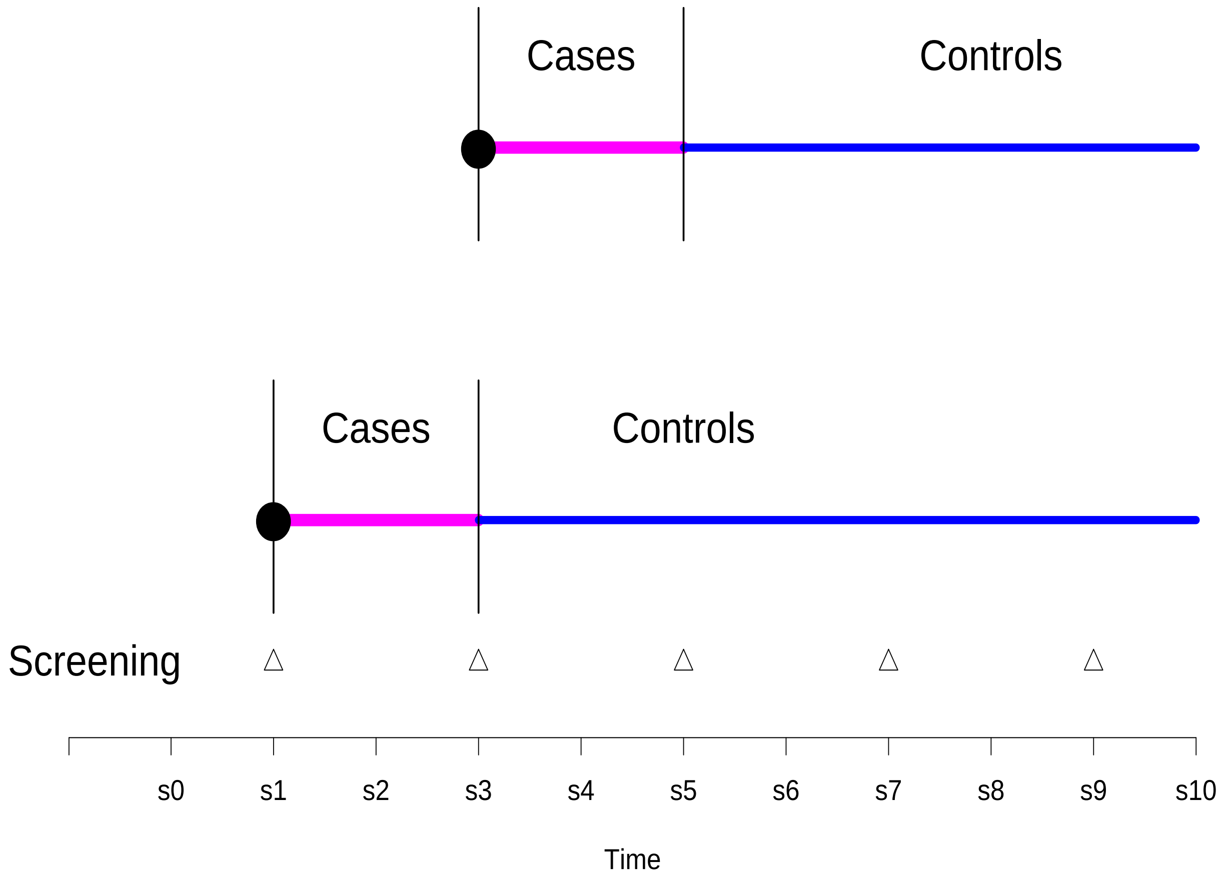


Figure 3. Summary survival ROC curve for three markers from MACS data.

Disease Screening



Disease Screening

$$\begin{aligned} \text{TP}(t,s) &= P[Y(s) > c \mid \text{CASE} \equiv \{T \le t\}, T \ge s] \\ \text{FP}(t,s) &= P[Y(s) > c \mid \text{CONTROL} \equiv \{T > t\}, T \ge s] \end{aligned}$$

$$\boxed{\text{Total Test Positive Proximal}}$$

- Screening times $\mathcal{S} = \{s_1, s_2, s_3, \ldots\}$
- Positive Proximal: $s = s_j$, and $t = s_{j+1}$

$$P[Y(s_j) > c_j \mid T \le s_{j+1}, T \ge s_j]$$

- TTPP($c_1, c_2, \ldots$)

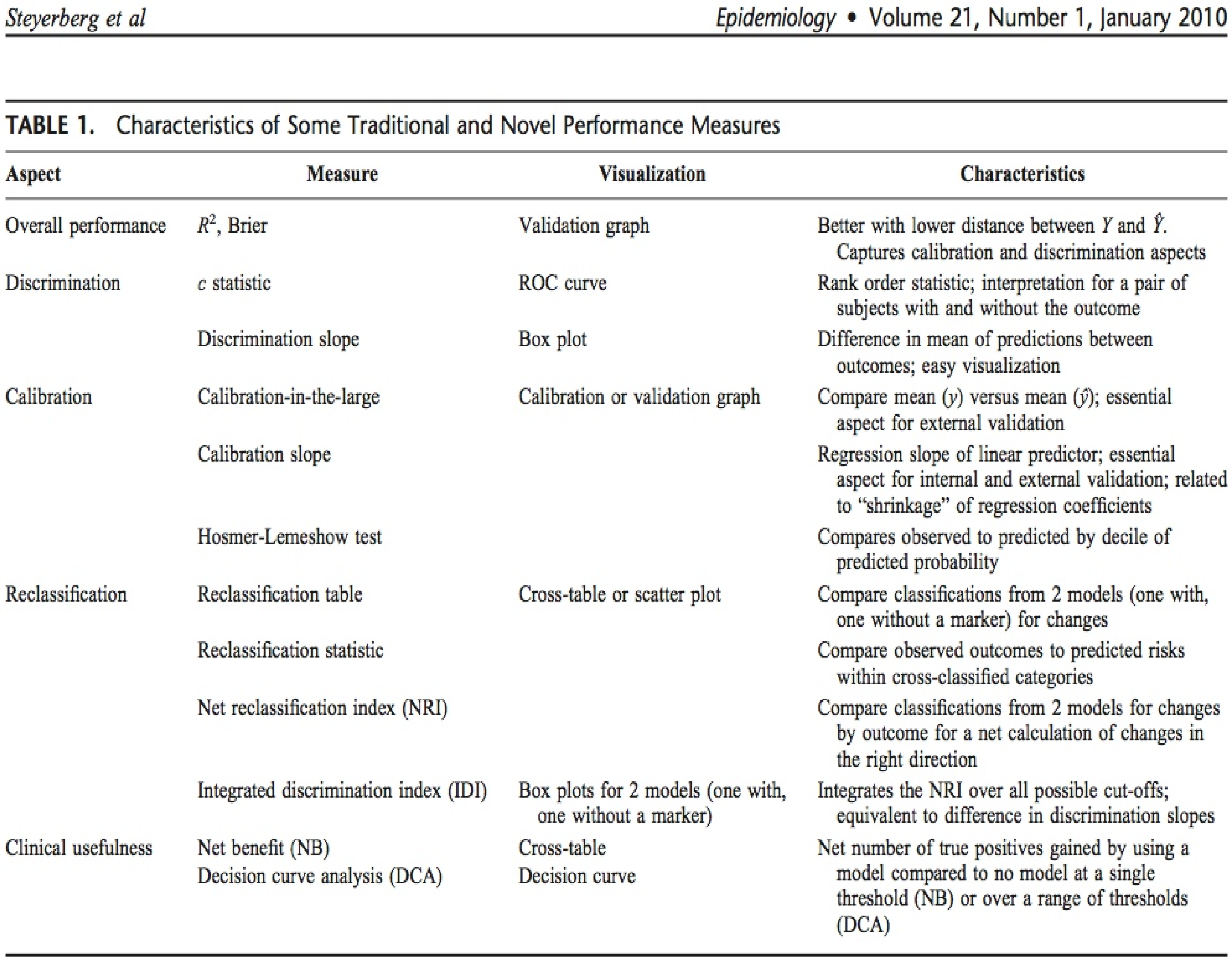
$$\sum_j TP(t = s_{j+1}, s = s_j) \times [S(s_j) - S(s_{j+1})]$$

Summary

Accuracy summary		
$ROC_t^{\mathbb{I}/\mathbb{D}}(p)$	:	vary (M,t)
$AUC(t)$	:	vary (t)
$\overline{ROC}(p)$	:	vary (M)
C	:	global

Summary

- Longitudinal data as predictors of an event time
- Accuracy using sequential binary classification
- Connections to partial likelihood
- Software
  - ▷ R code – `survivalROC`
  - ▷ R code – `risksetROC`
- Now: Inference
- Future: Relax strong censoring assumption



Steyerberg et al Epidemiology • Volume 21, Number 1, January 2010

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	*c* statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB) Decision curve analysis (DCA)	Cross-table Decision curve	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)

Discrimination Slope / IDI

- Pencina et al. (2008)
- **Integrated Discrimination Improvement**
- $\boxed{\text{Idea:}}$
  - ▷ Let $p(M_i) = P[D \mid M_i]$
  - ▷ Contrast mean risk in cases and mean risk in controls

$$\Delta P = \overline{p}_{case} - \overline{p}_{control}$$

  - ▷ IDI $= \Delta P_{new} - \Delta P_{old}$
  - ▷ Note: model-based
- Uno et al. (2012) extend to $D = 1(T \leq t)$
- **Q**: application to incident cases?

Net Reclassification Index

- Consider two marker/models that generate predictions
- Consider whether M_2 "moves" cases and controls relative to M_1
  - ▷ $D(t)$ disease status at time t
  - ▷ $p_1(t)$ and $p_2(t)$ based on M_1, M_2

	Case $D(t) = 1$	Control $D(t) = 0$
$p_2(t) - p_1(t) > 0$	$P(+, 1)$	$P(+, 0)$
$p_2(t) - p_1(t) \leq 0$	$P(-, 1)$	$P(-, 0)$

- **NRI**:

$$NRI(t) = \underbrace{[P(+,1) - P(-,1)]}_{\texttt{case:up/down}} + \underbrace{[P(-,0) - P(+,0)]}_{\texttt{control:down/up}}$$

- Pencina et al. (2008); French et al. (2012)

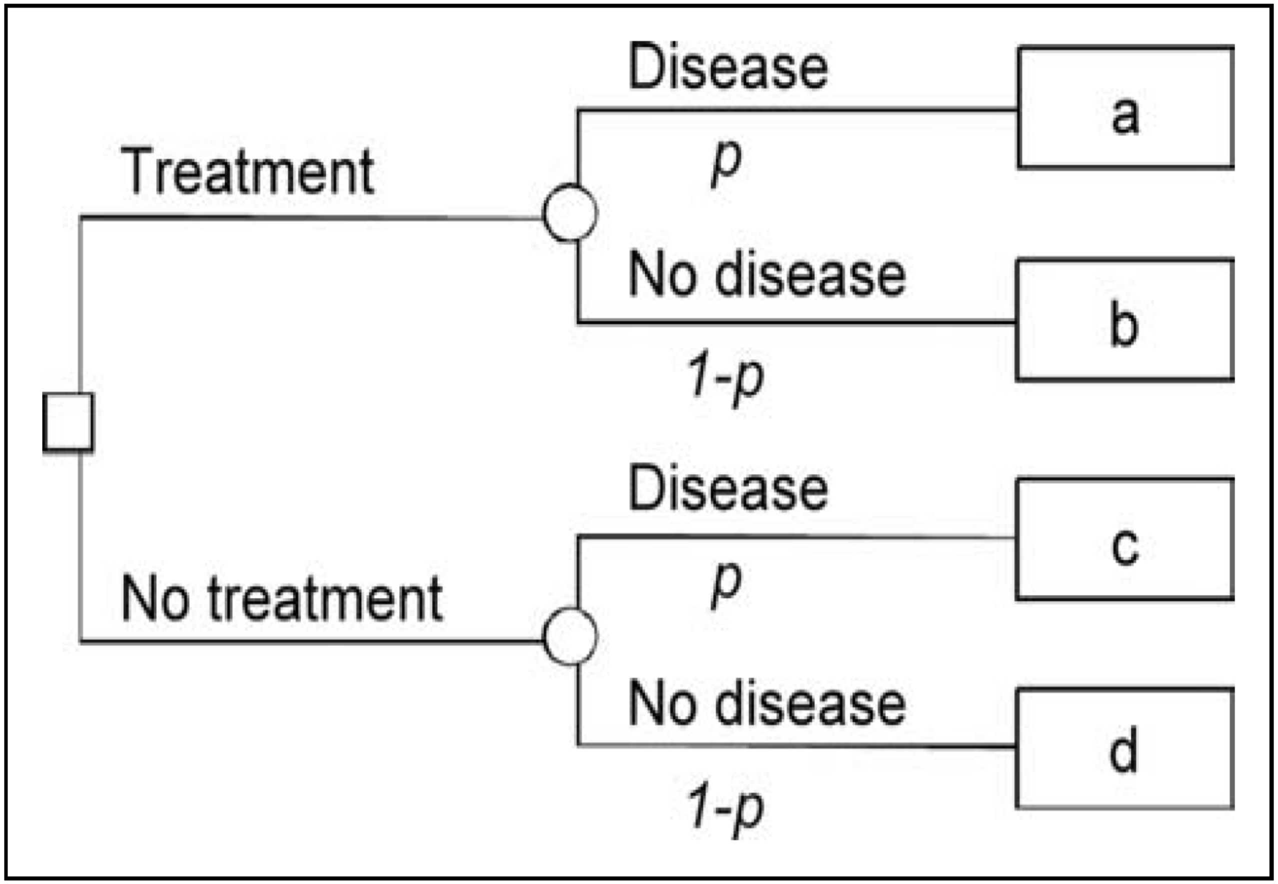


Figure 1 A decision tree for treatment. The probability of disease is given by p; a, b, c, and d give, respectively, the value of true positive, false positive, false negative, and true negative.

Decision Curve Analysis

- Vickers and Elkin (2006)
- Consider the context where patients have **outcomes** (e.g. QALY) that depend on their disease status and the treatment they receive as shown on previous figure.
- Consider a **decision function**

$$A(M, m) \ = \ 1(M > m)$$

- Let $A(M, m) = 1$ denote that treatment is used, and let $A(M, m) = 0$ denote that no treatment is used
- **Q**: What is the population mean if **no treatment** is used?
- **Q**: What is the population mean if **universal treatment** is used?
- **Q**: What is the population mean if **selective treatment** is used?

Decision Curve Analysis

- **No treatment**: $Tx = A(M, +\infty) \equiv 0$

treatment	disease	group size	mean
Tx	D	**0**	a
	$\overline{D}$	**0**	b
$\overline{Tx}$	D	$p(D)$	c
	$\overline{D}$	$p(\overline{D})$	d

Population mean:

$$\mu(0) = c \cdot p(D) + d \cdot p(\overline{D})$$

Decision Curve Analysis

- **Universal treatment**: $Tx = A(M, -\infty) \equiv 1$

treatment	disease	group size	mean
Tx	D	$p(D)$	a
	$\overline{D}$	$p(\overline{D})$	b
$\overline{Tx}$	D	**0**	c
	$\overline{D}$	**0**	d

Population mean:

$$\mu(1) = a \cdot p(D) + b \cdot p(\overline{D})$$

Decision Curve Analysis

- **Selective treatment**: $A(M,m)$

treatment	disease	group size	mean
Tx	D	$p[A(M,m)=1 \mid D] \cdot p(D)$	a
	$\overline{D}$	$p[A(M,m)=1 \mid \overline{D}] \cdot p(\overline{D})$	b
$\overline{Tx}$	D	$p[A(M,m)=0 \mid D] \cdot p(D)$	c
	$\overline{D}$	$p[A(M,m)=0 \mid \overline{D}] \cdot p(\overline{D})$	d

Population mean:

$$\begin{aligned}\mu(m) \quad &= \quad (a-c)\,TP(m)\,p(D) \;+\; c \cdot p(D) \;+ \\ &\quad\; \underbrace{(b-d)\;FP(m)\,p(\overline{D})}_{\Delta(m)} \;+\; \underbrace{d \cdot p(\overline{D})}_{\mu(0)}\end{aligned}$$

Decision Curve Analysis

- **Equipoise**: if $\mu(1)$ was thought to be overall just as beneficial as $\mu(0)$ when $p(D) = p^*$ then we would have:

$$\underbrace{a \cdot p(D) + b \cdot p(\overline{D})}_{\mu(1)} \quad = \quad \underbrace{c \cdot p(D) + d \cdot p(\overline{D})}_{\mu(0)}$$

$$\frac{(a-c)}{(d-b)} \quad = \quad \frac{1-p^*}{p^*}$$

- Interpretation:
  - ▷ $(a-c)$: **benefit** of treating a case
  - ▷ $(b-d)$: **cost** of treating a non-case

Decision Curve Analysis

- If we standardize and set $(a-c)=1$ then we obtain the relative benefit of not treating a non-case (e.g. $-$cost):

$$(d-b) = \frac{p^*}{(1-p^*)}$$

- For any cost/benefit we can then compare use of a model/marker via $A(M,m)$ to $\mu(0)$ (no treatment) to obtain the standardized **net benefit**:

$$\Delta(m) = TP(m)\cdot p(D) \; - \; FP(m)\cdot p(\overline{D})\cdot\frac{p^*}{(1-p^*)}$$

Decision Curve Analysis

- Note that we really have a net benefit surface, $\Delta(m, p^*)$
- One option would be to explore the net benefit of a decision function that chooses treatment if the predicted probability of disease is greater than p^*: (here assume monotone risk function)

$$\begin{aligned} A(M, m^*) &= 1(M > m^*) \\ &= 1[\ P(D \mid M = m) > P(D \mid M = m^*) = p^*\ ] \end{aligned}$$

- **Decision curve**:

$$\Delta(p^*) = TP(m^*) \cdot p(D) \ - \ FP(m^*) \cdot p(\overline{D}) \cdot \frac{p^*}{(1 - p^*)}$$

- Plot: $[\ p^*,\ \Delta(p^*)\ ]$

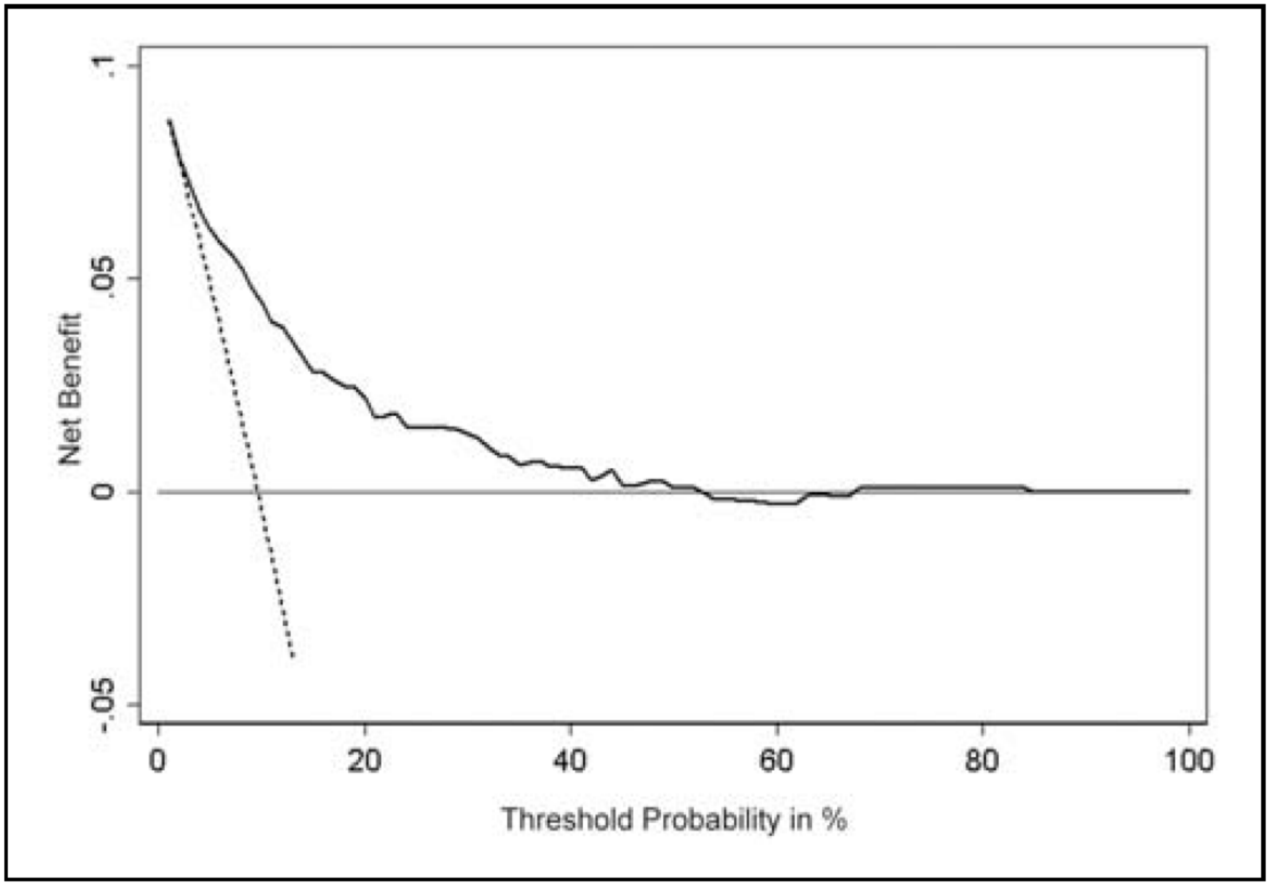


Figure 2 *Decision curve for a model to predict seminal vesicle invasion (SVI) in patients with prostate cancer. Solid line: prediction model. Dotted line: assume all patients have SVI. Thin line: assume no patients have SVI. The graph gives the expected net benefit per patient relative to no seminal vesicle tip removal in any patient ("treat none"). The unit is the benefit associated with 1 SVI patient duly undergoing surgical excision of the seminal vesicle tip.*

Summary

- Prognosis → Action → Yield
- Cost / benefit
- "Value of information"
- **Q**: action based on prognosis?
- **Q**: outcome independent of m?
- `www.decisoncurveanalysis.org`