Gene selection for survival data under dependent censoring: a copula-based approach

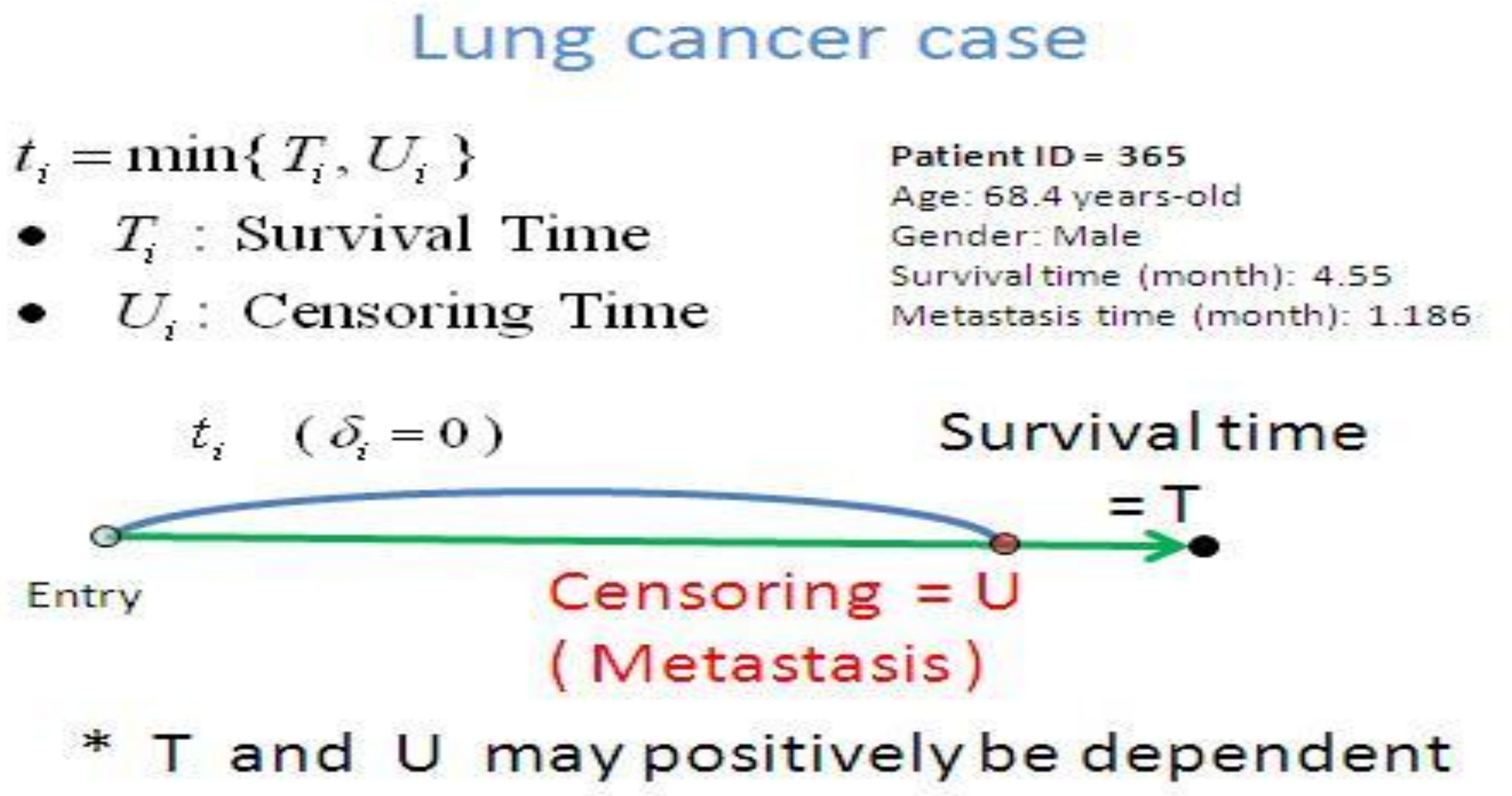
Takeshi Emura (presenter): Graduate Institute of Statistics, National Central University, Taiwan
Yi-Hau Chen: Institute of Statistical Science, Academia Sinica, Taiwan

Abstract: Dependent censoring arises in biomedical studies when the survival outcome of interest is censored by competing risks. In survival data with microarray gene expressions, gene selection based on the univariate Cox regression analyses has been used extensively in medical research, which however, is only valid under the independent censoring assumption. In this work, we utilize copula-based dependence models to develop an alternative gene selection procedure. Simulations show that the proposed procedure adjusts for the effect of dependent censoring and thus outperforms the existing method when dependent censoring is indeed present. The non-small-cell lung cancer data is analyzed to demonstrate the usefulness of our proposal. We implemented the proposed method in an R "compound.Cox" package.

1. Introduction

For survival data with microarrays, the primary task is selecting a small fraction of genes that are relevant to survival. The simplest approach is to select subsets of genes by using univariate selection based on Cox regression analyses which are used extensively in medical research[1-3].

The aforementioned univariate selection critically relies on the independent censoring assumption[4]; survival time and censoring time need to be statistically independent at a given gene. For our motivating example of non-small-cell lung cancer data of Chen *et al.*[3], some patients die soon after metastasis occurs (Figure below). Therefore, their censoring and survival times may be positively dependent.



If the independent censoring assumption is violated, univariate Cox regression analyses may not correctly identify the effect of each gene and thus may fail to select truly effective genes.

2. Univariate Selection

The approach called *univariate selection* is performed using a univariate Cox regression for each gene, one-by-one. Then a subset of genes that have low P-values is selected from the univariate analysis.

Specifically, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ be genes from individual *i*. We observe $(t_i, \delta_i, \mathbf{x}_i)$, where $t_i = \min\{T_i, U_i\}$ and $\delta_i = \mathbf{I}\{T_i \le U_i\}$. Univariate Cox regression on proportional hazard models

$$h(t \mid x_{ij}) = h_{0j}(t) e^{\beta_j x_{ij}}, \quad j = 1, \ldots, p$$

is performed one-by-one for each j. The resultant estimator $\hat{\beta}_j$ is used to obtain the P-value for the Wald test for $H_{0j}: \beta_j = 0$. One selects genes that exhibit smaller P-values than a threshold.

The estimator $\hat{\beta}_j$ can correctly identify the true β_j under the so-called "independent censoring" assumption[4]:

***Assumption I (Independent censoring)**:*
The survival time T_i and censoring time U_i are conditionally independent given a gene x_{ij} for each $j = 1, \ldots, p$.

3. Proposed method

3.1 Copula-based model

We propose adjusting for the effect of dependent censoring by modeling the dependency with a survival copula[5-8]:

$$\Pr(T_i > t, U_i > u \mid x_{ij}) = C_\alpha\{\Pr(T_i > t \mid x_{ij}), \Pr(U_i > u \mid x_{ij})\}$$

where copula is assumed to be the same across all *j* and indexed by a single parameter α. The analytically most convenient example is the Clayton copula,

$$C_\alpha(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \qquad \alpha \ge 0.$$

In this way, Assumption I is relaxed by the association parameter. For marginal distributions, we assume the proportional hazard models

$$\Pr(T_i > t \mid x_{ij}) = \exp\{-\Lambda_{0j}(t) e^{\beta_j x_{ij}}\},$$

$$\Pr(U_i > u \mid x_{ij}) = \exp\{-\Gamma_{0j}(u) e^{\gamma_j x_{ij}}\},$$

where β_j and γ_j are regression coefficients and Λ_{0j} and Γ_{0j} are baseline cumulative hazard functions.

July 8-9, 2013, Rennes, France

3.2 Semiparametric estimation

We adopt the semiparametric MLE of Chen[8] in which the forms of Λ_{0j} and Γ_{0j} are unspecified. For any given α, we maximize the full likelihood

$$\ell(\beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha) =$$
$$\sum_i \delta_i [\beta_j x_{ij} + \log \eta_{1ij}(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha) + \log d\Lambda_{0j}(t_i)]$$
$$+ \sum_i (1 - \delta_i)[\gamma_j x_{ij} + \log \eta_{2ij}(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha) + \log d\Gamma_{0j}(t_i)]$$
$$- \sum_i \Phi_\alpha[\exp\{-\Lambda_{0j}(t_i) e^{\beta_j x_{ij}}\}, \exp\{-\Gamma_{0j}(t_i) e^{\gamma_j x_{ij}}\}],$$

where, for $\Phi_\alpha = -\log C_\alpha$ and $D_{\alpha,k}(u_1, u_2) = -\partial \Phi_\alpha(u_1, u_2) / \partial u_k$,

$$\eta_{kij}(t; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha) = D_{\alpha,k}[\exp\{-\Lambda_{0j}(t) e^{\beta_j x_{ij}}\}, \exp\{-\Gamma_{0j}(t) e^{\gamma_j x_{ij}}\}]$$
$$\times \exp\{-\Lambda_{0j}(t) e^{\beta_j x_{ij}}\}^{\mathbf{I}(k=1)} \exp\{-\Gamma_{0j}(t) e^{\gamma_j x_{ij}}\}^{\mathbf{I}(k=2)}.$$

The first component of the MLE is denoted by $\hat{\beta}_j(\alpha)$. The standard error $se\{\hat{\beta}_j(\alpha)\}$ can be computed from the observed information matrix[8]. The P-value is computed by the Wald test. We implement the computation R compound.Cox package[9].

For a future subject with covariate $\mathbf{x} = (x_1, \ldots, x_p)'$, the survival prediction can be made by the prognostic index (PI) defined as $\mathrm{PI}_i(\alpha) = \hat{\boldsymbol{\beta}}'(\alpha) x_i$, where $\hat{\boldsymbol{\beta}}(\alpha)' = (\hat{\beta}_1(\alpha), \cdots, \hat{\beta}_p(\alpha))$.

3.3 Choice of association parameter

Due to the nonidentifiability of competing risks data[10], the likelihood may provide little information on α. Our approach maximizing the Harrell's concordance measure (*c*-index) defined as

$$CV(\alpha) = \frac{\sum_{i<j} \{\mathbf{I}(t_i < t_j)\mathbf{I}(\mathrm{PI}_i(\alpha) > \mathrm{PI}_j(\alpha))\delta_i + \mathbf{I}(t_j < t_i)\mathbf{I}(\mathrm{PI}_j(\alpha) > \mathrm{PI}_i(\alpha))\delta_j\}}{\sum_{i<j} \{\mathbf{I}(t_i < t_j)\delta_i + \mathbf{I}(t_j < t_i)\delta_j\}}.$$

Therefore, we set $\hat{\alpha} = \arg\max CV(\alpha)$.

4. Simulations

We compare the performance of the proposed method with the univariate selection via simulations. Data are generated from the Clayton copula with exponential margins:

$$\Pr(T_i > t, U_i > u \mid \mathbf{x}_i) = (\exp\{-te^{\boldsymbol{\beta}'\mathbf{x}_i}\}^{-\alpha} + \exp\{-ue^{\boldsymbol{\gamma}'\mathbf{x}_i}\}^{-\alpha} - 1)^{-1/\alpha}.$$

We set $\boldsymbol{\beta} = \boldsymbol{\gamma} \in \mathbf{R}^{100}$, which yields approximately 50% censoring.

We compare the performance of gene selection in terms of sensitivity and specificity. Let $(P_1, \cdots, P_p)$ be a vector of P-values obtained by a gene selection method (univariate selection or proposed method) and let $P_{(c)}$ be the c^{th} smallest P-value. Then,

$$\text{Sensitivity} = \left\{ \sum_{j=1}^{p} \mathbf{I}(P_j \le P_{(q)}, \beta_j \ne 0) \Big/ \sum_{j=1}^{p} \mathbf{I}(\beta_j \ne 0) \right\} \times 100 \ (\%)$$

is the percentage of selecting truly effective genes while

$$\text{Specificity} = \left\{ \sum_{j=1}^{p} \mathbf{I}(P_j > P_{(q)}, \beta_j = 0) \Big/ \sum_{j=1}^{p} \mathbf{I}(\beta_j = 0) \right\} \times 100 \ (\%)$$

is the percentage of not selecting non-effective genes.

Larger values of sensitivity and specificity correspond to better gene selection ability. We report the results in terms of the average of 50 Monte Carlo replications (see below).

Comparison based on n = 100 samples and 50 replications.

		Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (± SD)	$E[\hat{\alpha}]$
$\boldsymbol{\beta} = (\underbrace{0.4, \ldots, 0.4}_{\times 5}, \underbrace{-0.4, \ldots, -0.4}_{\times 5}, \underbrace{0, \ldots, 0}_{\times 90})$; $\beta_1 = 0.4$				
Univariate selection	$\alpha = 1/2$ (tau = 0.2)	33.20 (92.58)	0.26 (± 0.17)	/
	$\alpha = 2$ (tau = 0.5)	33.80 (92.64)	0.25 (± 0.18)	/
	$\alpha = 8$ (tau = 0.8)	33.60 (92.62)	0.26 (± 0.18)	/
Proposed method	$\alpha = 1/2$ (tau = 0.2)	39.40 (93.27)	0.28 (± 0.14)	4.3
	$\alpha = 2$ (tau = 0.5)	39.60 (93.29)	0.28 (± 0.16)	4.0
	$\alpha = 8$ (tau = 0.8)	41.40 (93.49)	0.27 (± 0.17)	4.6
$\boldsymbol{\beta} = (\underbrace{0.4, \ldots, 0.4}_{\times 10}, \underbrace{0, \ldots, 0}_{\times 90})$; $\beta_1 = 0.4$				
		Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (± SD)	$E[\hat{\alpha}]$
Univariate selection	$\alpha = 1/2$ (tau = 0.2)	32.80 (92.53)	0.23 (± 0.17)	/
	$\alpha = 2$ (tau = 0.5)	32.80 (92.53)	0.23 (± 0.17)	/
	$\alpha = 8$ (tau = 0.8)	33.60 (92.62)	0.25 (± 0.17)	/
Proposed method	$\alpha = 1/2$ (tau = 0.2)	42.60 (93.62)	0.26 (± 0.14)	4.6
	$\alpha = 2$ (tau = 0.5)	42.80 (93.64)	0.26 (± 0.13)	4.5
	$\alpha = 8$ (tau = 0.8)	44.40 (93.82)	0.27 (± 0.16)	5.2
$\boldsymbol{\beta} = (\underbrace{0.2, \ldots, 0.2}_{\times 10}, \underbrace{-0.2, \ldots, -0.2}_{\times 10}, \underbrace{0, \ldots, 0}_{\times 80})$, $\beta_1 = 0.2$				
		Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (± SD)	$E[\hat{\alpha}]$
Univariate selection	$\alpha = 1/2$ (tau = 0.2)	31.00 (82.75)	0.12 (± 0.15)	/
	$\alpha = 2$ (tau = 0.5)	30.60 (82.65)	0.11 (± 0.17)	/
	$\alpha = 8$ (tau = 0.8)	30.70 (82.67)	0.11 (± 0.16)	/
Proposed method	$\alpha = 1/2$ (tau = 0.2)	36.10 (84.03)	0.14 (± 0.15)	4.1
	$\alpha = 2$ (tau = 0.5)	35.30 (83.83)	0.13 (± 0.15)	3.9
	$\alpha = 8$ (tau = 0.8)	37.60 (84.40)	0.13 (± 0.14)	3.9

Higher sensitivity and specificity correspond to better gene selection performance.

5. Data Analysis

We revisit the non-small-cell lung cancer data of Chen *et al.*[3] The data contains 125 lung cancer patients (63 training + 62 testing) in which 38 patients died while the others were censored.

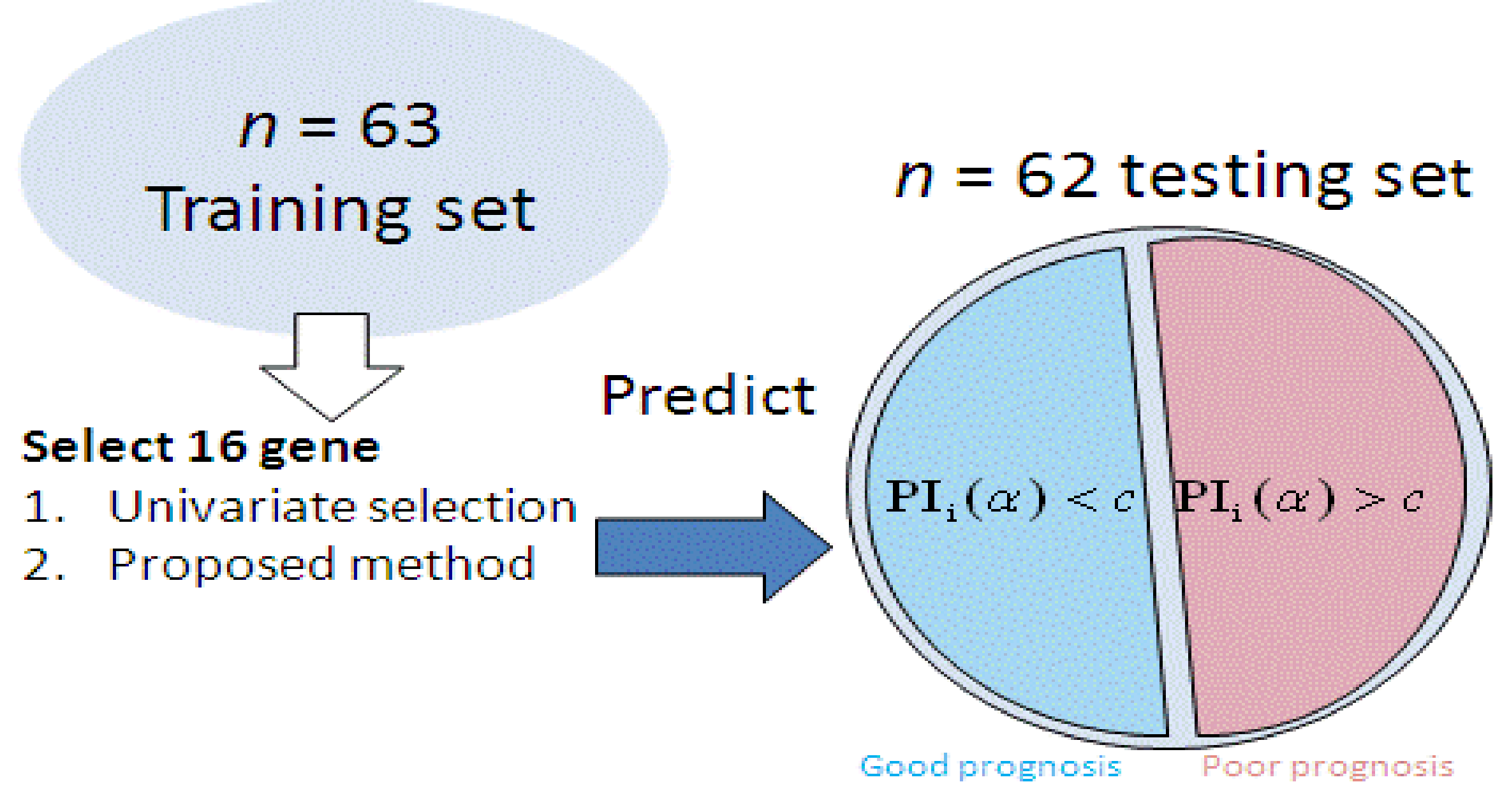
We compare univariate selection with our proposed method in terms of selecting the top = 16 genes among the 485 genes. The two gene selection methods used on the training set resulted in two different lists of the top genes (see below):

The 16 most strongly associated genes based on two methods

	Univariate selection			Proposed method		
No.	Gene	Coefficient	P-value	Gene	Coefficient	P-value
1	ANXA5	-1.09	0.0039	ZNF264	0.51	0.0004
2	DLG2	1.32	0.0041	MMP16	0.50	0.0005
3	ZNF264	0.55	0.0079	HGF	0.50	0.0010
4	DUSP6	0.75	0.0086	HCK	-0.49	0.0012
5	CPEB4	0.59	0.0162	NF1	0.47	0.0016
≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈						
15	MMD	0.92	0.0419	ENG	-0.37	0.0139
16	HMMR	0.52	0.0481	CKMT1A	-0.41	0.0155

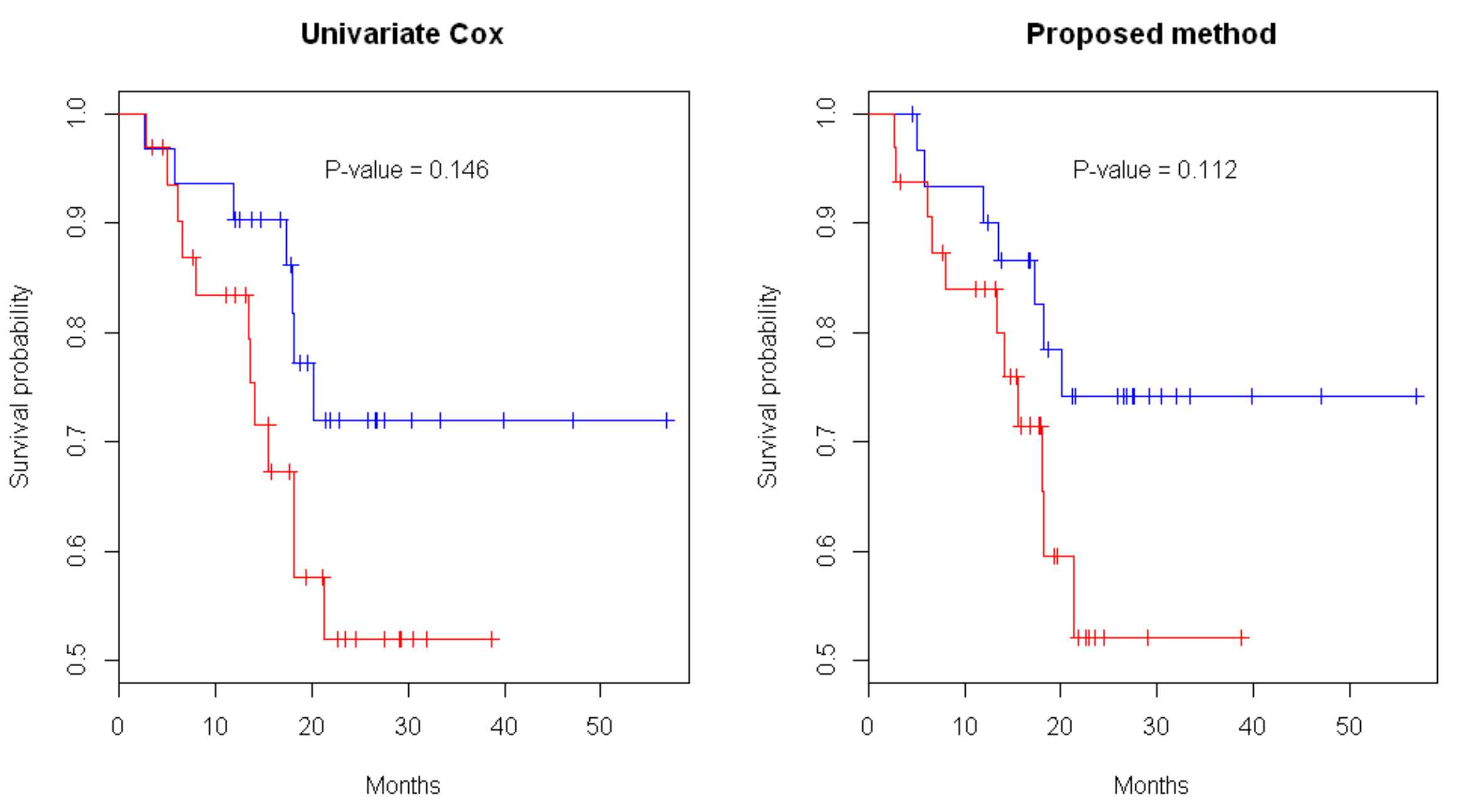
Gray shading = appear in both univariate selection and the proposed method.

We compare the performance of the two methods in terms of the ability to separate the good and poor prognosis groups in the testing sets (see Fig. below):



$$\mathrm{PI}_i(\alpha) = \hat{\beta}_1(\alpha) x_{i1} + \cdots + \hat{\beta}_{16}(\alpha) x_{i16}$$

The proposed method leads to a slightly better separation of the good and poor prognoses (P-value = 0.112) compared to that in the univariate Cox regression method (P-value = 0.146) (see Fig below):



6. Conclusion

1) We propose a copula-based gene selection method.
2) The method improves the chance of selecting truly effective genes over the univariate selection (simulations).
3) Genes selected by the proposed method is more predictive of survival than the univariate selection (data analysis).
4) This work will be published as Emura and Chen (2014)[11]

Reference

1 Jenssen TK, et al. E. *Human Genetics* 2002; **111:** 411-20.
2. Matsui S. *BMC Bioinformatics* 2006; **7**:156.
3 Chen HY, *et al*. *The New England Journal of Medicine* 2007; **356:** 11-20.
4 Andersen PK, et al. *Statistical Models Based on Counting Processes,* 1993.
5 Rivest LP, Wells MT. *Journal of Multivariate Analysis* 2001; **79:** 138-55.
6 Escarela G, Carriere JF. *Stat. Method in Med. Res.* 2003; **12**: 333-349.
7 Braekers R, Veraverbeke. *The Canadian Journal of Statistics* 2005; **33**: 429-447.
8 Chen YH. *Journal of the Royal Statistical Society, Series B* 2010; **72:** 235-51.
9 Emura T, Chen YH. *R compound.Cox package, version 1.1*. 2012.
10 Tsiatis A. *Proc. Natn. Acad. Sci. USA* 1975; **72:** 20-22.
11 Emura T, Chen YH. Gene selection for survival data under dependent censoring, a copula-based approach", to appear in Statistical Methods in Medical Research