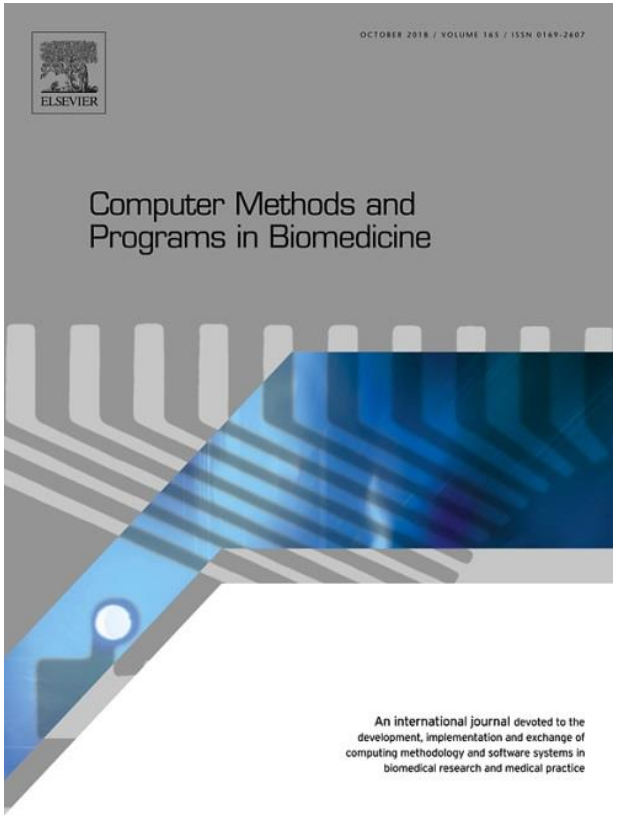
compound.Cox: 単変量Cox回帰法にもとづく特徴選択（feature selection）のRパッケージ

発表者；江村剛志（國立中央大學，統計研究所、台湾）
松井茂之（名古屋大学大学院、医学部）
Hsuan-Yu Chen (中央研究院、統計科学研究所、台湾)

参考文献；**Emura T, Matsui S, Chen HY (2019)**
Computer Methods and Programs in Biomedicine
Volume 168, January 2019, Pages 21-37
https://doi.org/10.1016/j.cmpb.2018.10.020



2018年度R研究集会、統計数理研究所、12月8日

R パッケージ *compound.Cox*

Compound.Cox_1.0: **2012年4月16日にCRAN登録**

Emura et al. (2012 *PLoS ONE*); compound shrinkage法を実装

Compound.Cox_1.x: Emura et al. (2012 *PLoS ONE*)の
遺伝子発現量のシミュレーション法を実装

Compound.Cox_1.4: Emura et al. (2016 *SMMR*)、
Copulaに基づく遺伝子選択（Feature selection）法を実装

Compound.Cox_3.0: Chen et al. (2007 NEJM)の肺がんデータ公開

Compound.Cox_3.4: Matsui (2006 Bioinformatics)のCVL法実装

Compound.Cox_3.x: Witten&Tibshirani (2010 SMMR)のFDR法実装

Compound.Cox_3.13: **最新版2018年7月17日**
遺伝子選択＆生存予測を実行する包括的Rパッケージとして
2019年CMPB誌1月号に掲載

本日発表の部分

問題設定

生存期間 = 応答変数（Response）
遺伝子発現量 = 予後因子（Features）

大量の遺伝子の中から、
生存期間（予後）に関連する遺伝子を選択したい

- 肺がん患者の関連遺伝子 (Chen et al., 2006 NEJM)
- 乳がん患者の関連遺伝子（ Wang et al., 2005 Lancet ）
- 卵巣がん患者の関連遺伝子
  (Yoshihara et al., 2012, PLoS ONE; Waldron et al. 2014
  J Natl Cancer Inst; Emura et al. 2018 SMMR)

肺がん患者の実例

ERBB3, *LCK*, *DUSP6*, *STAT2*
など16の遺伝子が生存期間に関連
(**Chen** et al., 2007 NEJM)
本研究の共同研究者
単変量Cox回帰で16の関連遺伝子を選択

使用したデータ;
n＝125人の肺がん患者（台湾の大学病院）
（死亡38人 ＋ 打ち切りによる生存87人）
p＝485個の遺伝子
（選択された関連遺伝子は16個）

肺がんデータは *Lung* に入っている

訓練標本（n＝63）または
テスト標本（n＝62）の指標

```
> library(compound.Cox)
> data("Lung")   打ち切り
> Lung                ↓                          ↓p=97個の遺伝子（離散値）
```

生存期間(月)

→→t .vec	d.vec	train	VHL	IHPK1	…	RPL5
1 47.06271	0	FALSE	2	2		4
2 49.27393	0	TRUE	3	4		4
3 20.06601	1	TRUE	2	3		1
4 26.99670	1	TRUE	2	4		2
5 39.90099	0	FALSE	3	4		4
⋮ ⋮	⋮	⋮	⋮	⋮		⋮
125 56.84141	0	FALSE	3	2	…	3

Chen et al.（2007）では
n＝63人の訓練標本で関連遺伝子を選択、
n＝62人のテスト標本で関連遺伝子のバリデーション

T = 死亡時間

x_j = j番目の遺伝子発現量

T と x_j 関連強➡単変量Cox回帰の推定値が大

$$h(t \mid x_j) = \frac{\Pr(t \le T < t + dt \mid T \ge t, x_j)}{dt} = h_0(t)\exp(\beta_j x_j)$$

データの形式

$$\{ (t_i, \delta_i, \mathbf{x}_i); i = 1, \ldots, n \},$$

- t_i : survival time or censoring time,
- δ_i : censoring indicator ($\delta_i = 1$ if t_i is survival time, or $\delta_i = 0$ if t_i is censoring time),
- $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$: p -dimensional features (genes).

単変量Cox回帰にもとづく遺伝子選択法

Step1: 単変量Coxモデルで全遺伝子を個々に検定（多重検定）

$$H_{oj}: \beta_j = 0 \quad \text{vs.} \quad H_{1j}: \beta_j \neq 0$$

$$h_{0j}(t)\exp(\beta_j x_{ij}), \quad j = 1,\ldots, p$$

（1）ワルド検定；Z値＝ $\hat{\beta}_j / SE(\hat{\beta}_j)$

$\hat{\beta}_j$ = 単変量部分尤度推定値

（2）スコア検定； Z値＝ $S_j / V_j^{1/2}$ ＝単変量スコア統計量÷SD

Step2：棄却された遺伝子を選択

例； P値が0.05以下である遺伝子を選択

Step3：選択された遺伝子の

（1）FDR値を計算（偽陽性率）

（2）CVL値を計算（予測能力）

Step4：選択された複数の遺伝子で生存期間の予後予測（後述）

Steps1-4は関数 *uni.selection( )* で一括して実行可能

関数 *uni.selection( )*

遺伝子発現量
入力（$n \times p$行列）

スコア検定（TRUE）
ワルド検定（FALSE）

生存期間入力
（n成分ベクトル）

打ち切り指標入力
（n成分ベクトル）

P値閾値入力

FDR値をPermutation法
で計算（TRUE/FALSE）

```
> uni.selection(t.vec,d.vec,X.mat,K=20,P.value=0.05,score=FALSE,permutation=TRUE) ## Wald test
$beta
ANXA5       DLG2       ZNF264      DUSP6      CPEB4       LCK         STAT1
-1.0876762  1.3215044  0.5473276   0.7524497  0.5891676   -0.8447389  -0.5844262
 RNF4        IRF4       STAT2       HGF        ERBB3       NF1         FRAP1
 0.6463635   0.5176704  0.5849869   0.5086750  0.5509026   0.4715235   -0.7696768
 MMD         HMMR
 0.9151541   0.5156711

$Z
 .....   省略        ⇐Z値
$P
 .....   省略        ⇐P値

$CVL
-98.66365      ⇐CVL値

$FDR
P.value * (No. of genes)          Permutation
0.3031250                          0.3128125      ⇐FDR値
```

K分割クロスバリデーションでCVL値を計算

$\hat{\beta}_j$ 回帰係数

P値0.05で *uni.selection( )* を ｎ＝６３の訓練標本に適用し遺伝子選択

```
> uni.selection(t.vec,d.vec,X.mat,K=20,P.value=0.05,score=FALSE,permutation=TRUE)
$beta
ANXA5       DLG2       ZNF264      DUSP6      CPEB4       LCK          STAT1
-1.0876762  1.3215044  0.5473276  0.7524497  0.5891676  -0.8447389  -0.5844262
 RNF4        IRF4       STAT2       HGF        ERBB3       NF1          FRAP1
 0.6463635  0.5176704  0.5849869  0.5086750  0.5509026  0.4715235  -0.7696768
 MMD         HMMR
 0.9151541  0.5156711
```

↑ Chen et al. (2007 NEJM)と同じ１６遺伝子が選択された

- **統計的考察１**、この16の遺伝子の中に、偽陽性はいつくあるか？
  ⇒FDR（False discovery rate）値を計算
- **統計的考察２**、P値の閾値０．０５は適当か？
  ⇒CVL（Cross-validated likelihood)値を計算

False Discovery Rate (FDR)とは

FDR=選択された遺伝子中の無関係な遺伝子の割合

	選択された遺伝子	選択されなかった遺伝子	全遺伝子
関連する遺伝子			
無関係な遺伝子	*f*		
	q=16		***p=97***

FDR = $f/16$ 、 ここで未知数 f を
Permutation法で推定（詳細はEmura et al. 2019）

多重検定に関する注意

- **第一種の過誤（0．05）は1つの検定に対するもの**
  - **多重検定全体としての第一種の過誤は0．05より大**
  - 有意水準5%で100回検定すると、
    0．05×100＝平均5回くらいは誤って棄却される。
- **多重検定では、第一種の過誤でなく、**
  FDRを制御する場合がある
  (例；FDRを0．2以下にする)
- **FDRは選択された遺伝子の予測能力ではない**
  （後述のCVL値を予測能力として使用）

CVL（Cross-validated likelihood）とは

P値の閾値を与えたもとで選択された遺伝子の予測能力の指標。
部分尤度の*K*分割クロスバリデーションを用いて、次の式で定義。

$$CVL = \sum_{k=1}^{K} \{ \ell(\hat{\gamma}_{-k}) - \ell_{-k}(\hat{\gamma}_{-k}) \},$$

$$\text{where} \quad \hat{\gamma}_{-k} = \arg\max_{\gamma} \ell_{-k}(\gamma),$$

線形モデルにおける
二乗予測誤差（PRESS）に相当

$$\ell(\gamma) = \sum_i \delta_i \left[\gamma \mathrm{CC}_{i,-k} - \log \left\{ \sum_{\ell \in R_i} \exp\left(\gamma \mathrm{CC}_{\ell,-k} \right) \right\} \right],$$

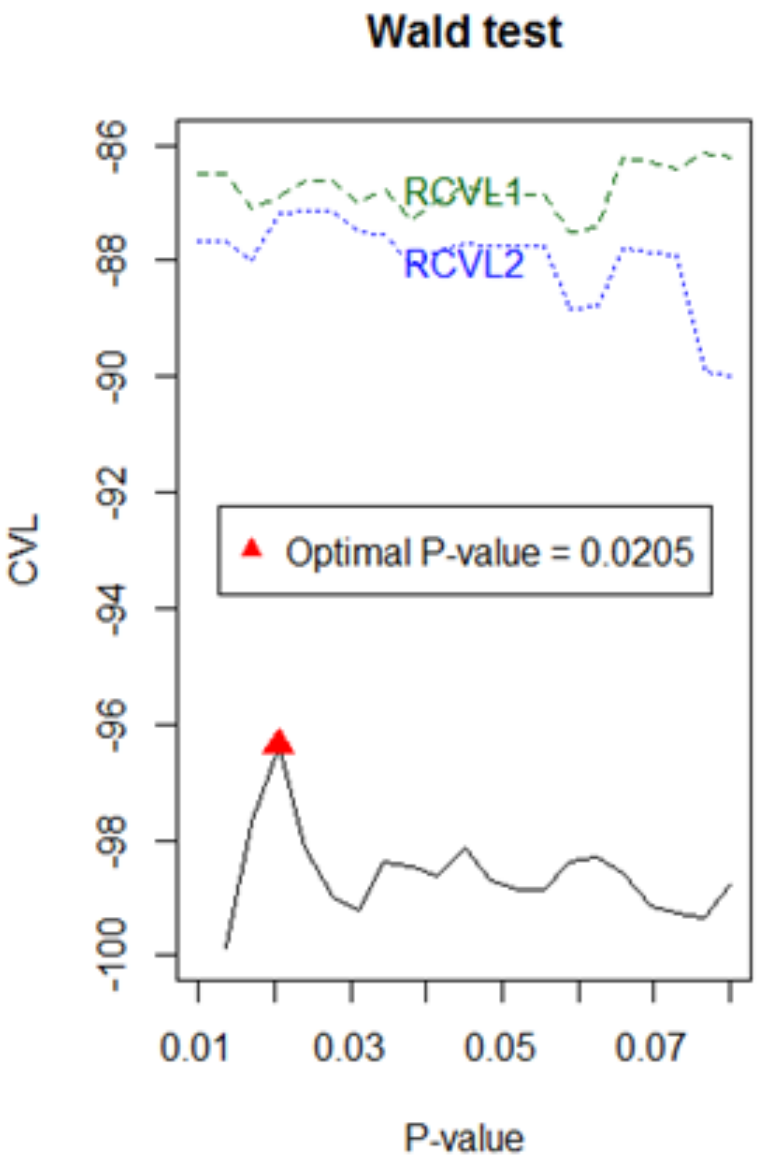
$$\mathrm{CC}_{i,-k} = \sum\nolimits_{j \in \Omega_{-k}} w_{j,-k} x_{ij}$$

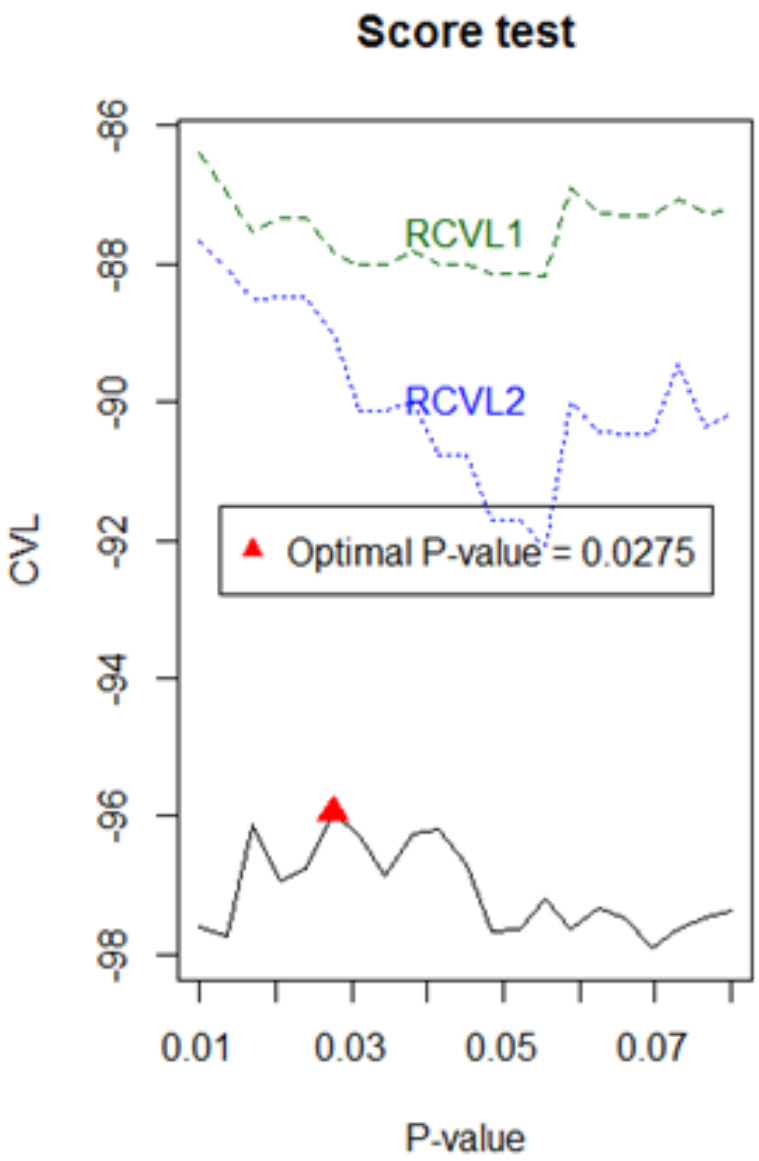
$$\ell_{-k}(\gamma) = \sum_{i \in \Im_{-k}} \delta_i \left[\gamma \mathrm{CC}_{i,-k} - \log \left\{ \sum_{\ell \in R_i \cap \Im_{-k}} \exp\left(\gamma \mathrm{CC}_{\ell,-k} \right) \right\} \right],$$

高CVL値➡高予測能力;　Matsui 2006；　Emura, Matsui and Chen 2019

※CVL値は訓練標本のみ（テスト標本を使わず）で予測能力を推定

CVL値を最適化するP値の閾値を求める





ワルド検定の最適閾値（P＝0.0205）で7個の遺伝子を選択

```
> uni.selection(t.vec,d.vec,X.mat,K=20, P.value=0.0205,score=FALSE)
$beta
ANXA5       DLG2       ZNF264      DUSP6      CPEB4       LCK          STAT1
-1.0876762  1.3215044  0.5473276   0.7524497  0.5891676   -0.8447389   -0.5844262

$CVL   -96.37303
```

↑CVL値

FDR値=0.0205×97/7=0.29（29%）

スコア検定の最適閾値（P＝0.0275）で10個の遺伝子を選択

```
> uni.selection(t.vec,d.vec,X.mat,K=20, P.value=0.0275,score=TRUE)
$Z
ANXA5      DLG2       ZNF264     DUSP6      CPEB4      LCK          STAT1      STAT2
-3.363578  3.111772   2.814363   2.710854   2.538888   -2.511423    -2.445038  2.369334
RNF4       IRF4
2.345912   2.231286
⋮
$CVL     -95.95690
```

↑CVL値

FDR値=0.0275×97/10=0.30（30%）

生存期間の予後予測

- **選択された複数遺伝子の発現量**; $(x_1, \ldots, x_q)$

  スコア検定の最適閾値では　q＝10

- **複合共変量（Compound Covariate）**:

$$\mathrm{CC} = w_1 x_1 + \cdots + w_p x_q$$

β値; $(w_1, \ldots, w_q) = (\hat{\beta}_1, \ldots, \hat{\beta}_q)$

z値; $(w_1, \ldots, w_q) = (z_1, \ldots, z_q)$

- 予後分類:

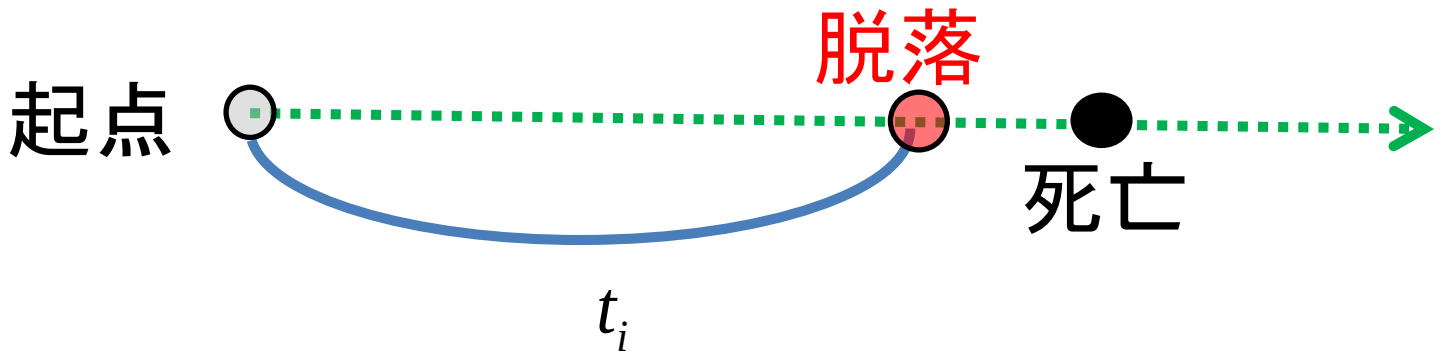
CC < c ➡ 予後は良い

CC > c ➡ 予後は悪い

ここで、c ＝CCの閾値

従属打ち切り（Dependent censoring）問題

死亡の直前に脱落する患者
　➡打ち切りと死亡に相関あり



独立打ち切りの仮定が崩れ、
Cox回帰の推定値にバイアスが生じる
（Emura & Chen 2016 *SMMR*）
⇒予後予測にもバイアスが生じる

従属打ち切りのモデルをコピュラ関数で特定すると、バイアス補正できる

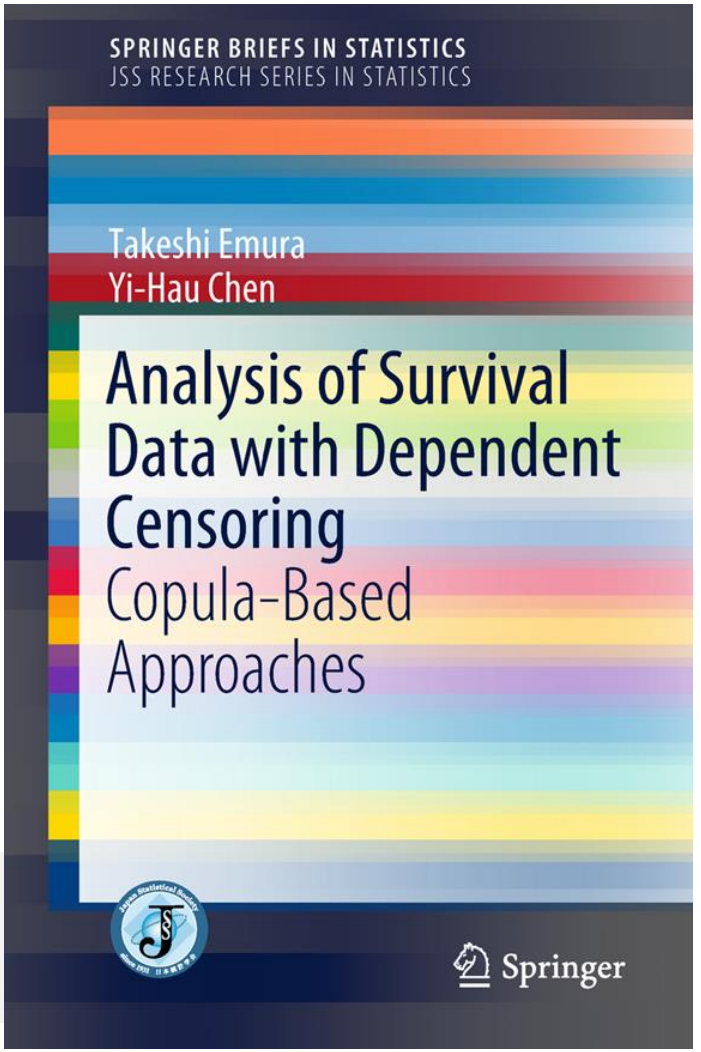
T = 生存期間（死亡時間）

U = 打ち切り（脱落）時間

x_j = j番目の遺伝子発現量

C_α = コピュラ関数

（αは相関パラメータ）



↓ 条件付き2次元生存関数

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

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

遺伝子 j が T に与える影響

コピュラによる従属打ち切りへの対応

セミパラ最尤推定 (Emura & Chen 2016, SMMR)

$$\ell(\beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha)$$

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

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

$$- \sum_i \Phi_\alpha[\exp\{-\Lambda_{0j}(t_i)e^{\beta_j x_{ij}}\}, \exp\{-\Gamma_{0j}(t_i)e^{\gamma_j x_{ij}}\}],$$

↓ compound.Cox パッケージで計算

$$(\hat{\beta}_j(\alpha), \hat{\gamma}_j(\alpha), \hat{\Lambda}_{0j}(\alpha), \hat{\Gamma}_{0j}(\alpha))$$

↘ 予測に使う j 番目の遺伝子の重み w_j を計算

複数遺伝子を使った生存期間の予測法

1. **Optimal Wald（7個の遺伝子）:**

$$CC = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_7 x_{i7}$$

2. **Optimal score（10個の遺伝子）:**

$$CC = z_1 x_{i1} + \cdots + z_{10} x_{i10}$$

3. **Optimal Wald + copula（7個の遺伝子）:**

$$CC = \hat{\beta}_1(\hat{\alpha}) x_{i1} + \cdots + \hat{\beta}_7(\hat{\alpha}) x_{i7}$$

4. **Optimal score + copula（10個の遺伝子）**

$$CC = \hat{\beta}_1(\hat{\alpha}) x_{i1} + \cdots + \hat{\beta}_{10}(\hat{\alpha}) x_{i10}$$

$CC < c$ ➡ 予後は良い（高生存率）

$CC > c$ ➡ 予後は悪い（低生存率）

予後分類法を、n＝62のテスト標本でバリデーション

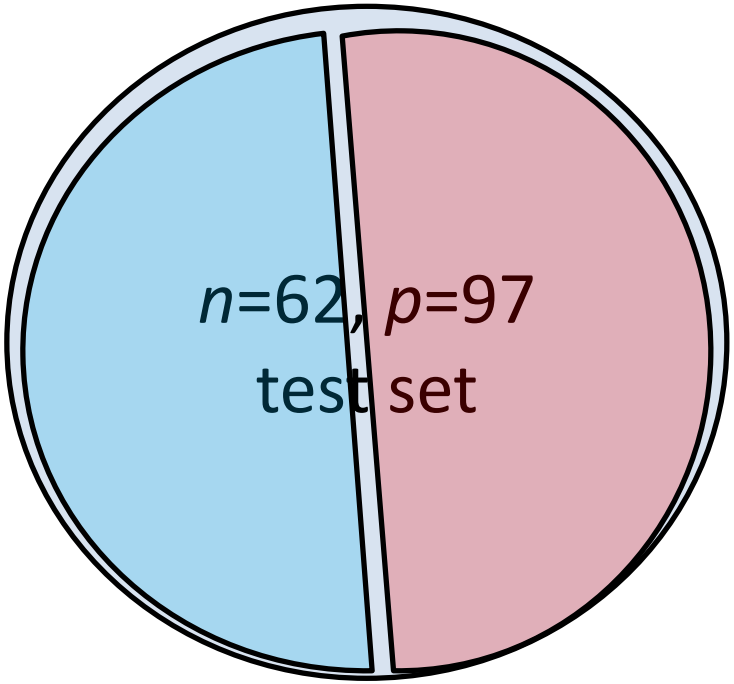
予後分類；予後が良（低CC値）、予後が悪（高CC値）

Good prognosis
予後が良いと
分類された
グループ

⇓

⇓

実際の生存率を計算

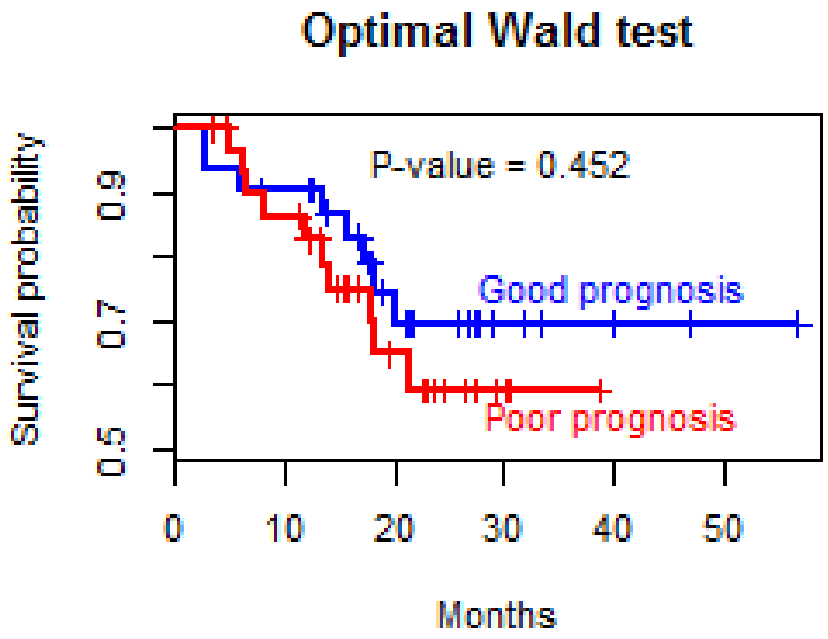


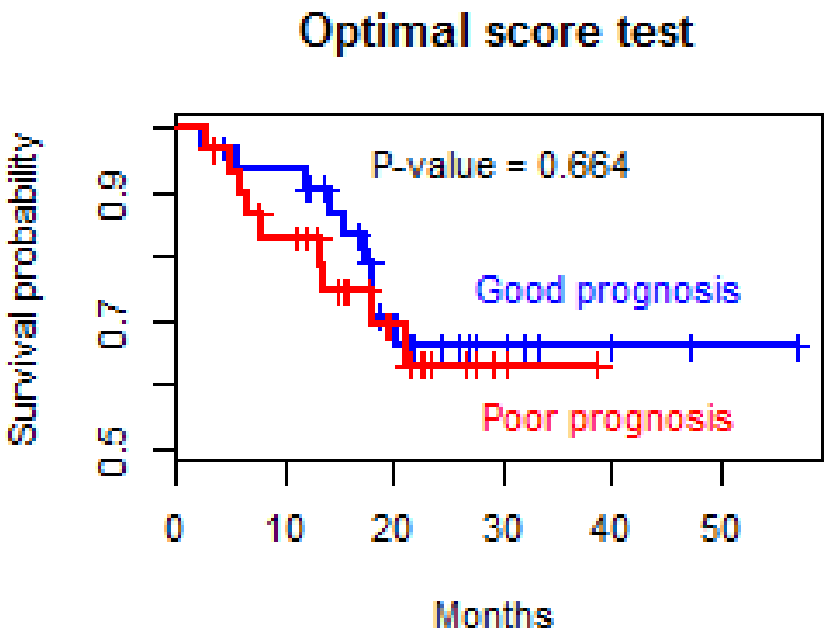
Poor prognosis
予後が悪いと
分類された
グループ

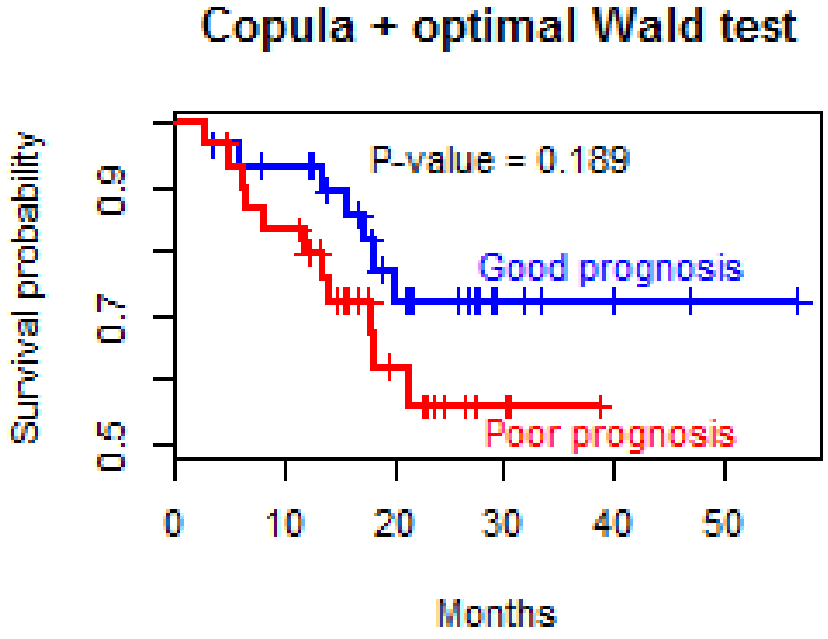
⇓

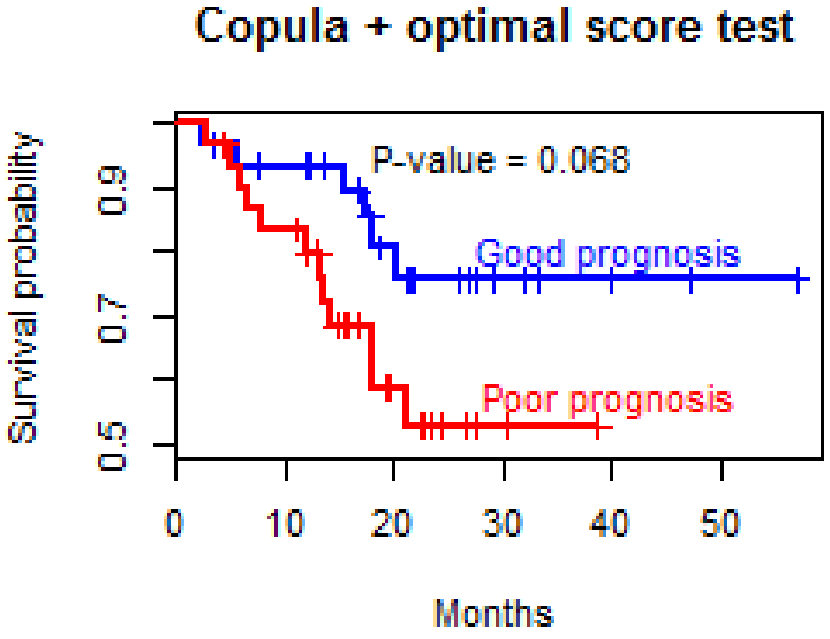
⇓

実際の生存率
を計算


Optimal Wald test
P-value = 0.452
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months


Optimal score test
P-value = 0.664
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months


Copula + optimal Wald test
P-value = 0.189
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months


Copula + optimal score test
P-value = 0.068
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months

本研究のまとめ

Rパッケージ；compound.Cox　を提案

- 単変量Cox回帰法による多重検定で遺伝子選択を実行

- **選択された遺伝子の評価法を実装**
  - 無関係な遺伝子の割合（FDR値）
    Witten and Tibshirani（2010 *SMMR*）
  - 予測能力評価（CVL値）　Matsui (2006 *BMC Bioinformatics*)

- **肺がん患者データ（125患者、97遺伝子）を公開**
  - n=63患者で遺伝子選択＆予後分類法の導出
  - n=62患者で予後分類法をバリデーション

- **従属打ち切りによるバイアス問題への対処法を実装**
  - 回帰推定値のバイアス補正　Emura and Chen（2016 *SMMR*）
  - 予測精度向上（患者の予後分類がより正確に）

ご清聴ありがとうございました

参考文献

[1] Witten M, Tibshirani R. Survival analysis with high-dimensional covariates. Statist Method Med Res 2010; 19: 29-51.

[2] Wang Y, Klijn JG, Zhang Y, Sieuwerts, AM, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet 2005; 365(9460): 671-79.

[3] Matsui S. Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. BMC Bioinformatics 2006; 7:156.

[4] Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007; 356: 11-20.

[5] Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H et al (2010). Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. PloS One 5(3), e9615

[6] Emura T, Nakatochi M, Matsui S, Michimae H, Rondeau V, Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors; meta-analysis with a joint model, Statist Methods Med Res 2018; 27:2842-2858.

[7] Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst 2014 106(5): dju049

[8] Emura T, Chen YH, Chen HY. Survival prediction based on compound covariate under Cox proportional hazard models. PLoS One 2012; 7(10): e47627. DOI:10.1371/journal.pone.0047627.

[9] Emura T, Chen YH, Gene selection for survival data under dependent censoring, a copula-based approach, Statist Method Med Res 2016; 25(6): 2840-2857.

[10] Emura T, Chen YH, Analysis of Survival Data with Dependent Censoring, Copula-Based Approaches, JSS Research Series in Statistics, Springer, Singapore; 2018.