科研費（基盤S）シンポジウム
2017年3月27-28日（福岡）

遺伝子発現量をCox回帰モデルに取り入れた生存期間の個別化予測

江村剛志
国立中央大学、統計研究所（台湾）
Graduate Institute of Statistics,
National Central University, TAIWAN

遺伝子発現量は、がん患者の予後予測に有用

- 乳がん (Jenssen et al. 2002; Sabatier et al. 2011)
- リンパ腫（Diffuse large-B-cell lymphoma）

  (Lossos et al. 2004; Binder and Schumacher 2008; Alizadeh 2011)
- 肺がん

  (Beer et al. 2002; Chen et al. 2007; Shedden et al. 2008)
- 卵巣がん

(Popple et al. 2012, Ganz fried et al. 2013; Waldron et al 2014)

これら医学研究では、生存時間解析
（Cox回帰、Kaplan-Meier曲線、ログランク検定）
を運用し、*予後予測モデル*を構築

クラシカルな生存時間解析による予測

$D = \text{Time-to-death}$

- 予後因子： $\mathbf{Z}$ =（年齢、ステージ、腫瘍のサイズ、遺伝子情報）
  ＊ t=0時点（予測時点）で記録

1） 予後分類；予後が良い（悪い）

$\text{PI} = \boldsymbol{\beta}'\mathbf{Z}$: Prognostic Index (予測指標)

$\text{PI} < c$ （良）； $\text{PI} > c$ （悪）, $c = \text{cut-off value}$

2） t-年後生存確率；$S(t \mid \mathbf{Z}) = \Pr(D > t \mid \mathbf{Z})$

通常、Coxモデルで予測式を構築： $S(t \mid \mathbf{Z}) = S(t \mid \mathbf{0})^{\exp(\boldsymbol{\beta}'\mathbf{Z})}$

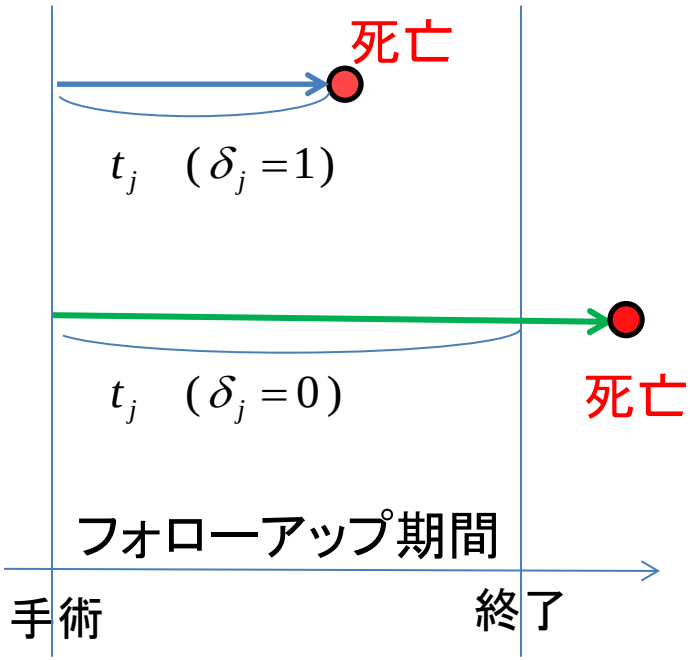
- 高次元生存時間データ

$$t_i : \text{time-to-death or censoring}$$

$$\delta_i = \begin{cases} 1 & \text{if death} \\ 0 & \text{if censoring} \end{cases}$$

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})', \quad \text{possibly} \quad p > n$$

共変量 → 遺伝子発現量



t_j Time-to-event	δ_j Censoring	x_{i1} AP3S1	x_{i2} APMAP	x_{i3} ARHGAP28	x_{i4} CXCL12	……..	x_{i127} ASB7	$x_{i,128}$ B4GALT5
1650	0	-0.52	1.12	-0.37	1.30	……	0.354	-1.015
30	1	-0.18	-0.69	-0.93	1.28	……	0.026	0.38
⋮	⋮					……		
1800	1	-1.08	0.70	-0.29	-0.529	……	-0.50	-1.09

Differentially expressed

←卵巣ガンのデータ
(n=912, p=18,548)
R *Joint.Cox* package
(Ganzfrid et al. 2013
Emura 2016, *CRAN*)

CXCL12 が1単位増加すると、死亡リスクが1．2倍（Ganzfried et al. 2013; Emura et al. 2015）

- $\mathbf{x}$;高次元遺伝子発現量
- ハザード関数

$$h(t \mid \mathbf{x}) = \Pr(t \le D \le t + dt \mid D \ge t, \mathbf{x}) / dt$$

$$D: \text{ time-to-death}$$

- Cox比例ハザードモデル (Cox 1972, *JRSSB*)

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}), \quad \boldsymbol{\beta} \in \mathbf{R}^p, \quad p > n$$

•部分尤度推定量:

$$\hat{\boldsymbol{\beta}} \in \mathbf{R}^p: \quad \text{maximize} \quad L_n(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{t_l \le t_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right)^{\delta_i}$$

$p > n$ のとき、$\hat{\boldsymbol{\beta}} \in \mathbf{R}^p$ は一意に定まらない（無限個ある）

(Witten & Tibshirani 2010, *SMMR*)

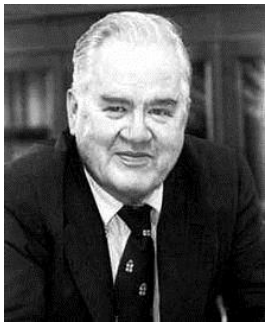
高次元（$p>n$）の生存時間データの解析法

- Lasso法 (Cox回帰の L_1 縮小推定)

  Tibshirani (1997 *Stat Med*), Gui & Li (2005 *Bioinformatics*)
- リッジ回帰法 (Cox回帰の L_2 縮小推定)

  Verveij & van Howelingen(1994 *Stat Med*), Zhao et al. (2011 *PONE*)
- 単変量Cox回帰法による変数選択（最も単純な方法）

  Jenssen et al. (2002 *Nature Med*), Chen et al. (2007 *NEJM*)
- 単変量Cox回帰法による複合共変量（Compound covariate）



Tukey (1993 *Controlled CT*), Wang et al. (2005 Lancet)
Matsui (2006, *BMC Bioinformatics*), Simon et al (2011 *Boinformatics*),
Matsui et al (2012 *Clin Can Res*), Emura et al (2012 *PONE)*,
Emura & Chen (2016 *SMMR*), Emura et al. 2017 SMMR),
←John Tukey

- 複合縮小（Compound shrinkage）法

  Emura et al (2012 *PONE*)
- その他 (PC, supervised PC, partial lease square, Boosting etc.)

縮小推定法のアイデア

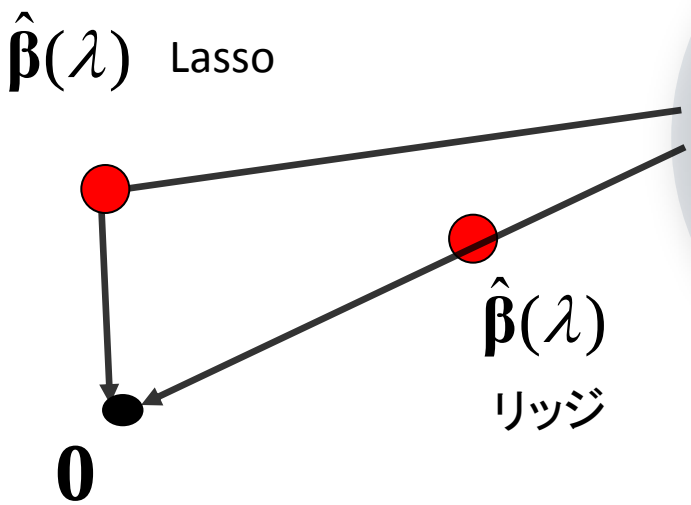
無限個の解空間をゼロに縮小し、

解の一意性を保障（罰則付き尤度法）

$$\hat{\boldsymbol{\beta}}(\lambda): \quad L_\lambda(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right)^{\delta_i} - \lambda \| \boldsymbol{\beta} \|^q$$

$\lambda > 0$ ；縮小の度合いを決める Tuning parameter

- リッジ回帰: $q = 2$
- Lasso: $q = 1$



無限個の解空間

$$\hat{\boldsymbol{\beta}}(0): \quad \mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_n(\boldsymbol{\beta}) = \mathbf{0}$$

単変量Cox回帰（1つ1つの遺伝子ごと）

- 単変量Coxモデル $h(t \mid x_j) = h_0(t)\exp(\beta_j x_j)$
  - $x_j = j\text{th gene expression}$
  - x_k is ignored for every $k \neq j$
- 回帰係数の解釈；
  j 番目の遺伝子発現値が1単位増加したときの、
  相対リスク

$$\exp(\beta_j) = \frac{h(t \mid x_j = 1)}{h(t \mid x_j = 0)}$$

注；*j* 番目の遺伝子の発現値が1単位増加すると、
その遺伝子と相関をもつ遺伝子の発現値も変化
（同Pathway内の他の遺伝子）

回帰係数の解釈；*j* 番目の遺伝子が他の遺伝子に与える影響も
包含した*主効果（Main effect）*

単変量Cox回帰による変数選択

Step1: 単変量Coxモデルを j 番目の遺伝子にあてはめ

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

Step2: **Wald 検定** $H_{oj}:\beta_j=0$ vs. $H_{1j}:\beta_j\neq 0$

$$|\hat{\beta}_j / sd\{\hat{\beta}_j\}| > z_{\alpha/2} \quad \Rightarrow \; j \text{ 番目の遺伝子を選択}$$

P値の選択基準

- P値 < 0.05 (e.g., Chen et al. 2007, *NEJM*)
- P値 < 0.001 (Simon 2003, *book*)

P値は***有意性ではなく***、単に

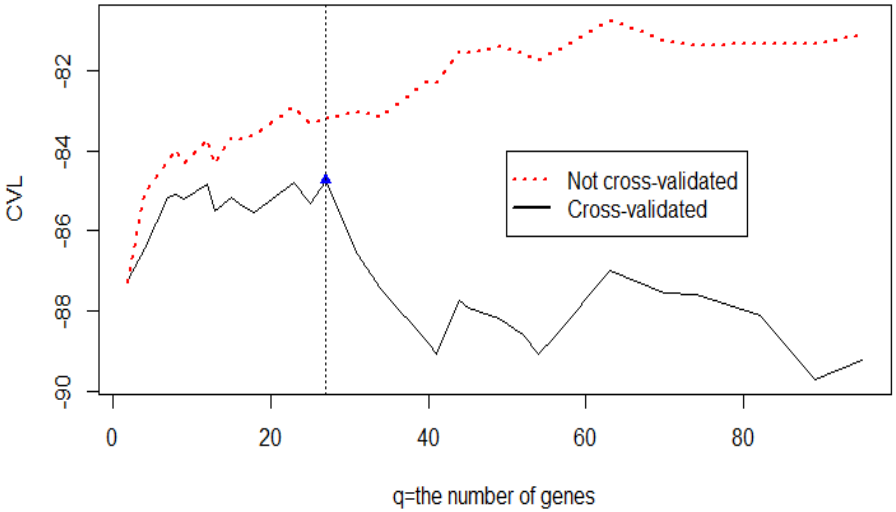
Tuning parameterと解釈すべき

- 部分尤度のクロスバリデーション

  (Matsui 2006 *BMC Bioinformatics*)

  P値<0.075（27genes）が最適➔

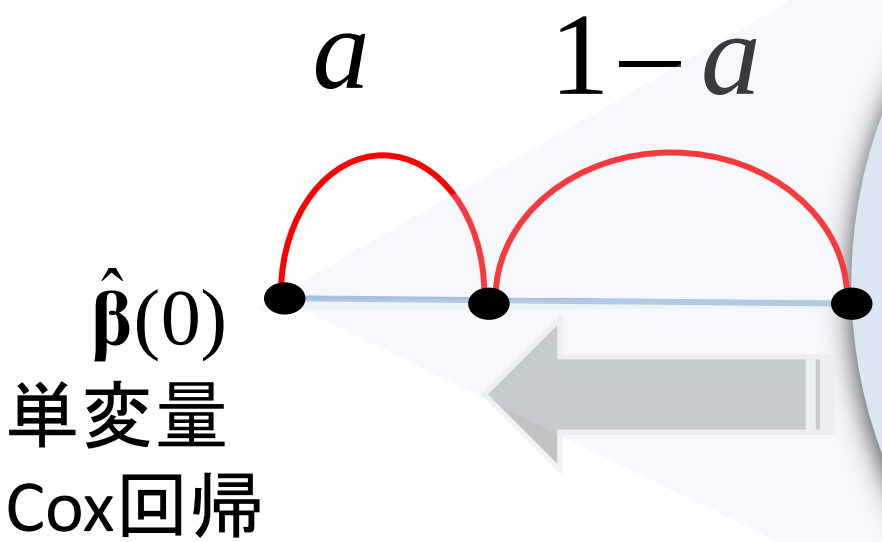
- FDR (Witten & Tibs. 2010 *SMMR*)



複合縮小推定 （多変量と単変量の混合尤度）

$$\hat{\boldsymbol{\beta}}(a) = \mathrm{argmax}\left\{ a \log L_n(\boldsymbol{\beta}) + (1-a) \log L_n^0(\boldsymbol{\beta}) \right\}$$

混合比 a



多変量Cox回帰の
無限個の解空間

$$\hat{\boldsymbol{\beta}}: \quad \mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_n(\boldsymbol{\beta}) = \mathbf{0}$$

注; $\mathbf{0}$ へ縮小するリッジやLassoとは本質的に異なる

４つの手法を数値的に比較（データ解析）

1. 複合共変量Compound covariate (CC)推定

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)', \quad \text{where } \hat{\beta}_j = \text{univariate Cox regression estimators}$$

2. 複合縮小 Compound shrinkage(CS)推定

$$\hat{\boldsymbol{\beta}}(\hat{a}): \quad a \log L_n^1(\boldsymbol{\beta}) + (1-a) \log L_n^0(\boldsymbol{\beta})$$

←R compound.Cox package (Emura et al, 2017, CRAN)

3. リッジ　Ridge estimator

$$\hat{\boldsymbol{\beta}}(\hat{\lambda}): \quad \log L_n^1(\boldsymbol{\beta}) - (\lambda/2) \sum_{j=1}^{p} \beta_j^2$$

←R penalized package (Goeman et al., 2016, CRAN)

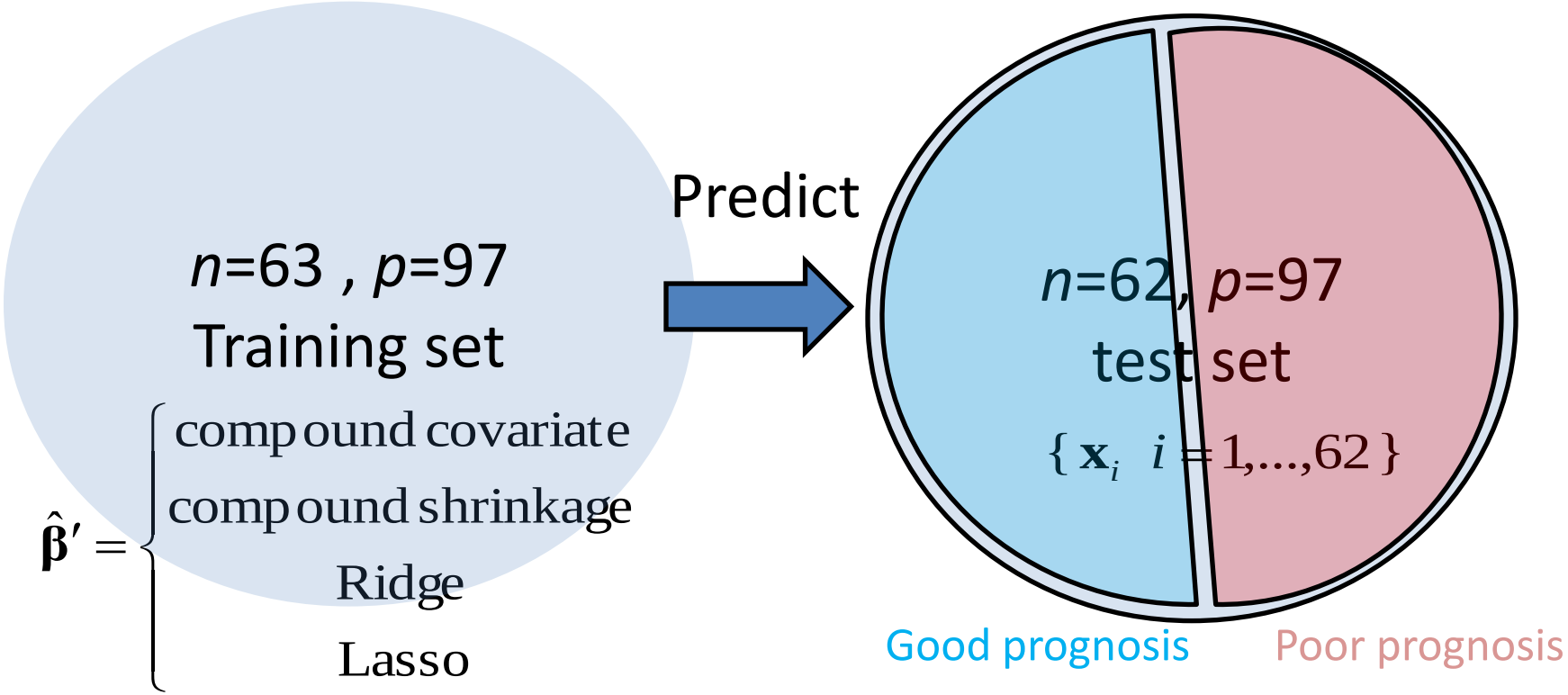
4. Lasso estimator

$$\hat{\boldsymbol{\beta}}(\hat{\lambda}): \quad \log L_n^1(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|$$

← R penalized package

* $\hat{a}$ or $\hat{\lambda}$ is obtained by cross-validation (Verveij & Houwelingen 1993 Stat.Med.)

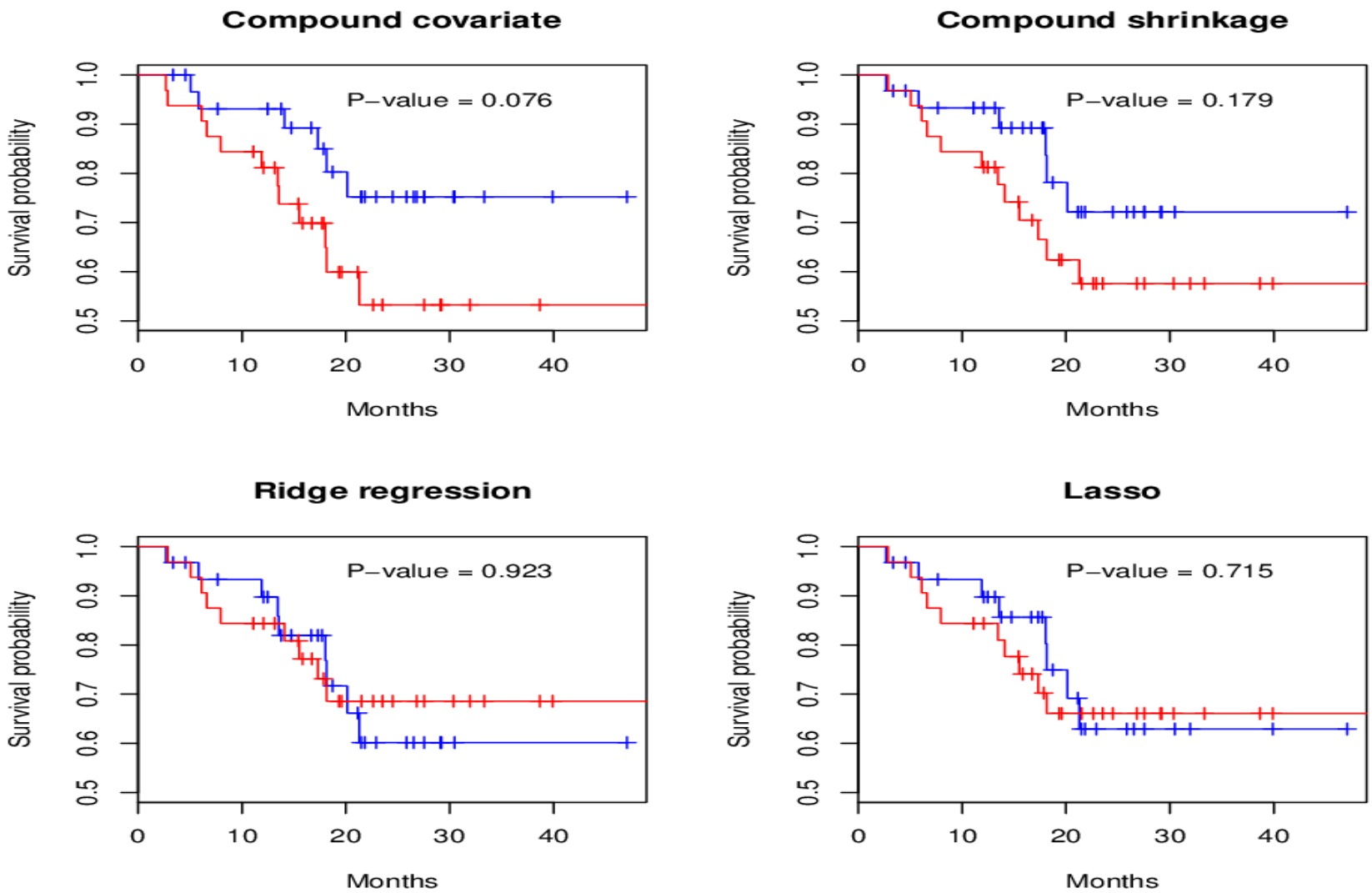
- Data: n=125の肺がん患者 (Chen et al., 2007 NEJM)



$$\hat{\boldsymbol{\beta}}'\mathbf{x}_i < c \ \ (\text{Good prognosis}) \ ; \ \ \hat{\boldsymbol{\beta}}'\mathbf{x}_i > c \ \ (\text{Poor prognosis}),$$

$$\text{where } c \text{ is the median of } \{\hat{\boldsymbol{\beta}}'\mathbf{x}_i, \mathrm{i} = 1,\ldots,\mathrm{n}\}$$

Survival curves for Poor vs. Good prognosis groups in n=62 testing data; p-value for Log-rank test



より高度な個別化生存予測

遺伝子発現量のみでの予測能力は限界あり
（Waldron et al.2014）

いくつかの解決法；

（1）通常の予後因子との複合（複合共変量を使用）
（2）動的予測の利用
（予測タイミングを変化、増悪後に予測をアップデート
（3）IPDメタアナリシス（患者個別データ）
（推定量の安定性の向上、予測モデルの一般化）
（4）コピュラを使った多変量生存モデル
〔死亡〕と〔増悪〕の同時モデル（Joint model）

Table 1. A meta-analytic data combining the four independent studies of ovarian cancer patients of Ganzfried et al. [34].

２変量生存データ（強い相関あり）

4つの研究
のメタアナリシス

ata set[a]	Sample size	The number of observed events (event rates)			The number of genes
		Relapse ($\delta_{ij}=1$)	Death ($\delta^*_{ij}=1$)	Censoring ($\delta^*_{ij}=0$)	
SE17260	$N_1=84$	59 (70%)	38 (45%)	46 (55%)	18,548
SE30161	$N_2=58$	48 (83%)	36 (62%)	22 (38%)	18,524
SE9891	$N_3=260$	185 (71%)	113 (43%)	147 (57%)	18,524
TCGA	$N_4=510$	252 (49%)	278 (55%)	232 (45%)	12,211
Total	$\sum_{i=1}^{4} N_i =$ 912	544 (60%)	465 (51%)	447 (49%)	Common=11,756

高次元遺伝子発現データ

Notes: The data are extracted from the *curatedOvarianData* R Bioconductor package of Ganzfried et al. [34];

Table 1 from Emura et al. (2017 *SMMR*)

同時モデル (Joint frailty Model, Rondeau et al. 2016 *SMMR*)

メタアナリシスのランダムエフェクト＝Frailty

$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t)\exp(\gamma_1 \mathrm{CC}_{1,ij}) & \text{(for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = u_i \lambda_0(t)\exp(\beta_2 Z_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & \text{(for time to death } D_{ij}) \end{cases}$$

予後因子＝術後腫瘍サイズ

$Z_{2,ij}$=the residual tumour size at surgery (<1cm vs. ≥ 1cm)

高次元遺伝因子＝Compound covariate (CC):

- $\mathrm{CC}_{1,ij} = (0.249*CXCL12)+(0.235*TIMP2)+(0.222*PDPN)+\cdots+(-0.152*MMP12)$,

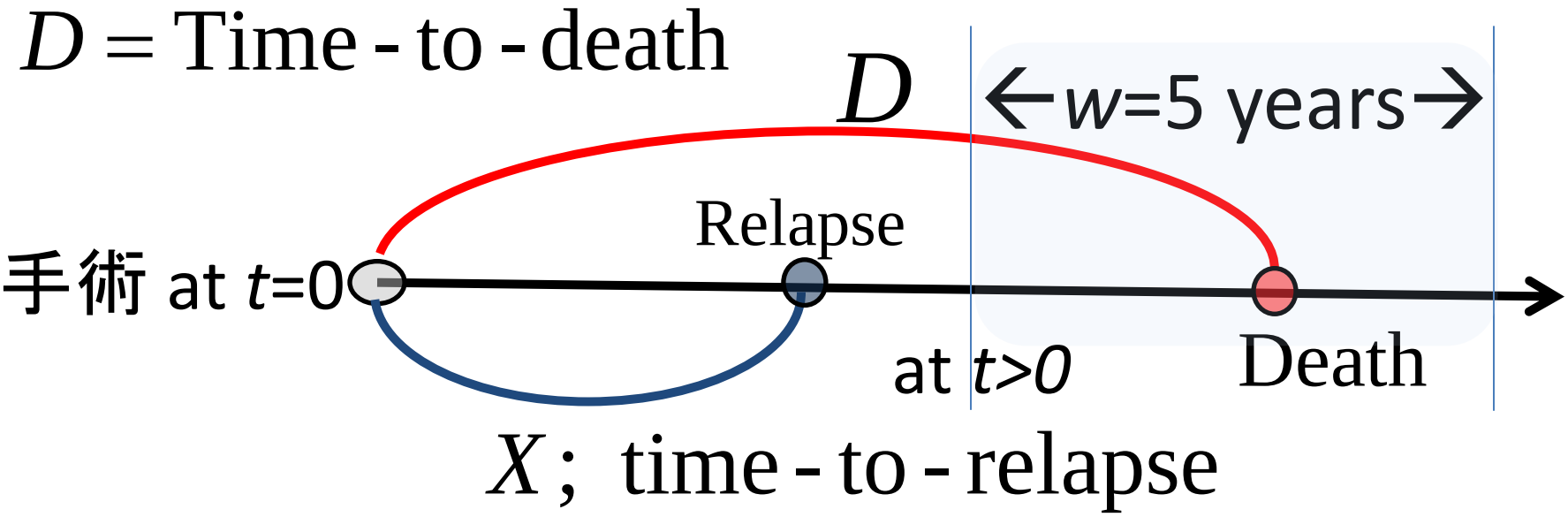
  involving 158 genes (P-value < 0.001 for time-to-relapse)

- $\mathrm{CC}_{2,ij} = (0.237*NCOA3)+(0.223*TEAD1)+(0.263*YWHAB)+\cdots+(-0.157*KCNH4)$,

  invloving 128 genes (P-value < 0.001 for time-to-death).

死亡の動的予測（Dynamic prediction of death）

$$D = \text{Time-to-death}$$



- 区間(t、t＋w)の死亡確率 (van Houwelingen and Putter 2013)

$$F(t, t+w \mid X, \mathbf{Z}) = \Pr(D \le t+w \mid D > t, X, \mathbf{Z})$$

- X と D の間の相関をCopulaでモデル化 (Emura et al. 2015)

$$\Pr(X > x, D > y \mid u) = C_\theta[S_X(x \mid u), S_D(y \mid u)]$$

$$C_\theta(v, w) = (v^{-\theta} + w^{-\theta} - 1)^{-1/\theta}, \qquad \theta \ge 0$$

データ解析;モデルのあてはめ

$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\gamma_1 \mathrm{CC}_{1,ij}) & (\text{for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = u_i^{\alpha} \lambda_0(t) \exp(\beta_2 Z_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & (\text{for time to death } D_{ij}) \end{cases}$$

$$\Pr(X_{ij} > x, D_{ij} > y \mid u_i) = C_\theta[S_X(x \mid u_i), S_D(y \mid u_i)]$$

Results obtained from R *joint.Cox* package (Emura, 2016 on CRAN)

	Parameter	Estimate	95% CI
Relapse	$\exp(\gamma_1)$	1.48	1.37-1.59
Death	$\exp(\beta_2)$	1.18	1.03-1.35
	$\exp(\gamma_2)$	1.56	1.44-1.70
Copula	θ	1.90	1.49-2.42
	$\tau = \theta/(\theta+2)$	0.49	0.32-0.65

患者レベルの予測（術後 500日）

$$F(t, t+w \mid H(t,x), \mathbf{Z}) = \Pr(D \le t+w \mid D > t, H(t,x), \mathbf{Z})$$

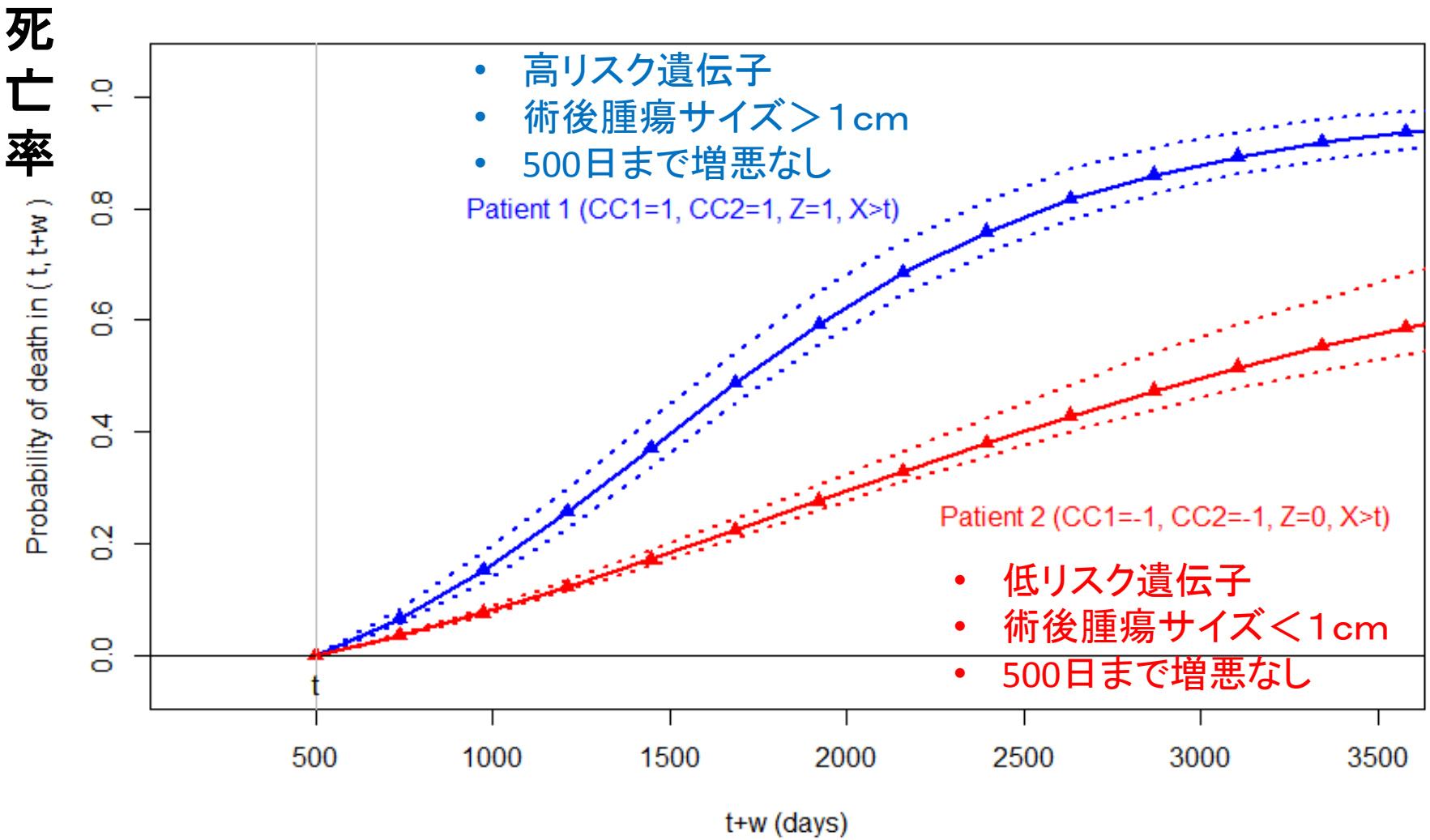


Figure from Emura et al. (2017 *SMMR*)

患者レベルの予測（術後 1000日）

$$F(t, t+w \mid H(t,x), \mathbf{Z}) = \Pr(D \le t+w \mid D > t, H(t,x), \mathbf{Z})$$

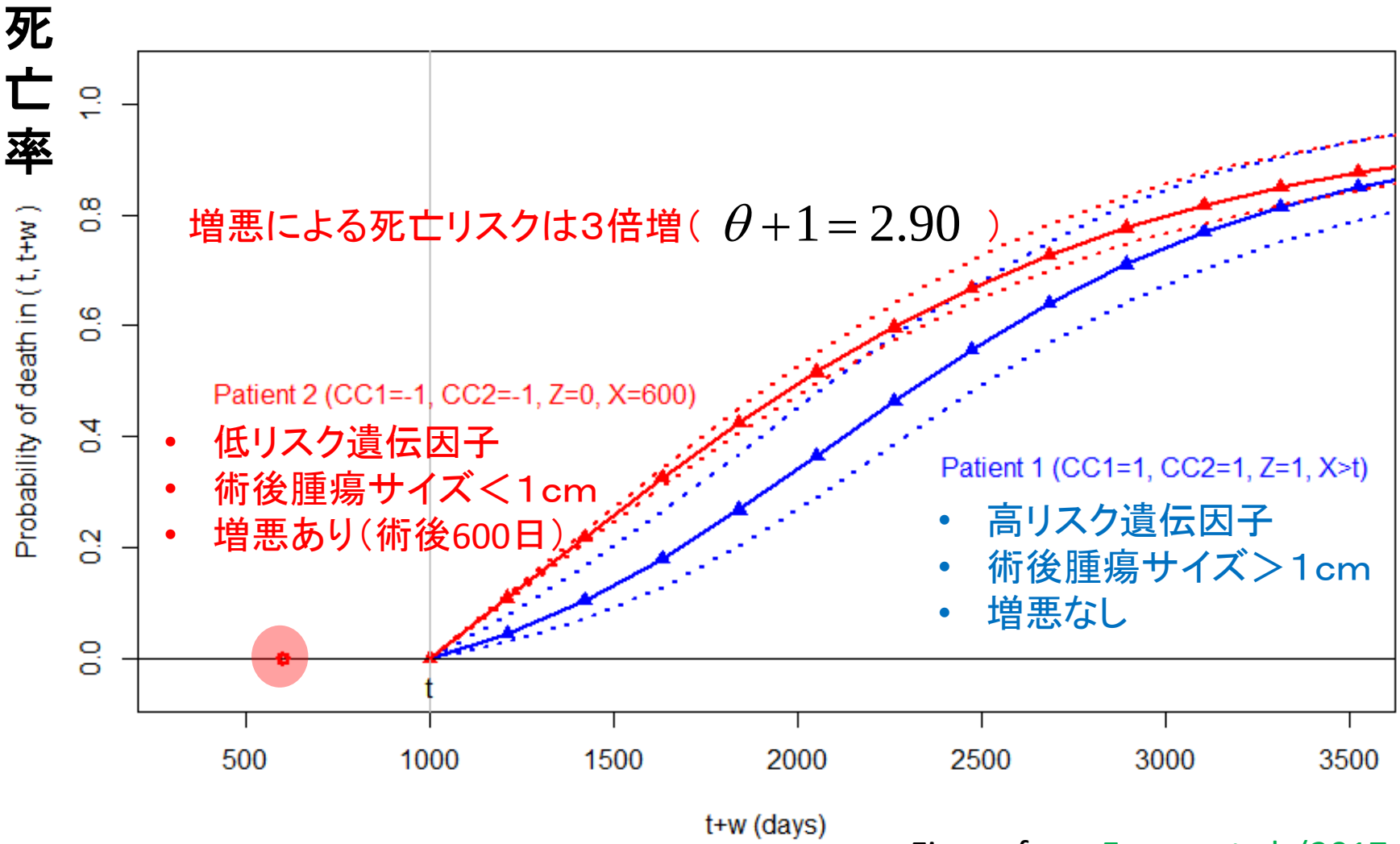


Figure from Emura et al. (2017 *SMMR*)

References

[1] Cox DR. Regression models and life-tables (with discussion), Journal of the Royal Statistical Society, Series B 1972; 34: 187-220.

[2] Jenssen TK, Kuo WP, Stokke T, Hovig E. Association between gene expressions in breast cancer and patient survival. Human Genetics 2002; 111: 411-20.

[3] Sabatier R, Finetti P, Adelaide J, Guille A, Borg JP, Chaffanet M, Bertucci F. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. PLoS One 2011: 6(11); e27656.

[4] Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R, Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. New England Journal of Medicine 2004; 350(18): 1828-1837.

[5] Alizadeh AA, Gentles AJ, Alencar AJ, Liu CL, Kohrt HE, et al. Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. Blood 2011; 118(5): 1350-1358.

[6] Beer DG, Kardia SLR., Huang CC., Giordano TJ, Levin AM, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine 2002; 8: 816-824.

[7] 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.

[8] Shedden K, Taylor JMG, Enkemann SA, Tsao MS, Yeatman TJ, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature Medicine 2008; 14: 822-827.

[9] Popple A, Durrant LG, Spendlove I, Scott PRI, Deen S, Ramage JM. The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. British Journal of Cancer 2012; 106: 1306-1313.

[10] 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. Journal of the National Cancer Institute 2014; 106(5): dju049.

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

[12] Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics; 2008; 9(1): 14

[13] Goeman J, Meijer R, Chaturvedi N. R penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model, CRAN 2016; version 0.9-47.

[14] Rondeau V, Pignon JP, Michiels S. A joint model for dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. Statistical Method in Medical Research 2015; 24(6): 711-729.

[15] Emura T, Nakatochi M, Murotani K, Rondeau V, A joint frailty-copula model between tumour progression and death for meta-analysis, Statistical Methods in Medical Research 2015; DOI: 10.1177/0962280215604510.

[16] Ganzfried BF, Riester M, Haibe-Kains B, Risch T, Tyekucheva S, et al. Curated ovarian data: clinically annotated data for the ovarian cancer transcriptome, Database 2013; Article ID bat013: DOI:10.1093/database/bat013.

[17] Emura T, Chen YH, Gene selection for survival data under dependent censoring, a copula-based approach, Statistical Methods in Medical Research 2016; 25(6): 2840-2857.

[18] Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997; 16: 385-395.

[19] Van Houwelingen HC, Bruinsma T, Hart AA, van't Veer LJ, Wessels LF, Cross-validated Cox regression on microarray gene expression data. Statistics in Medicine 2006; 25(18): 3201-3216.

[20] Tukey JW, Tightening the clinical trial. Controlled Clinical Trials 1993; 14: 266-285.

[21] 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.

[22] Radmacher MD, Mcshane LM, Simon RM. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 2002; 9:505-511.

[23] Matsui S, Simon RM, Qu P, Shaughnessy JD, Barlogie B, Crowley J. Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. Clinical Cancer Research 2012; 18(21): 6065-6073.

[24] Emura T. R joint.Cox: penalized likelihood estimation and dynamic prediction under the joint frailty-copula models between tumour progression and death for meta-analysis, CRAN; version 2.10 2016-10-30.

[25] Emura T*, Nakatochi M, Matsui S, Michimae H, Rondeau V (2017) Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model, Stat Methods Med Res, doi:10.1177/0962280216688032