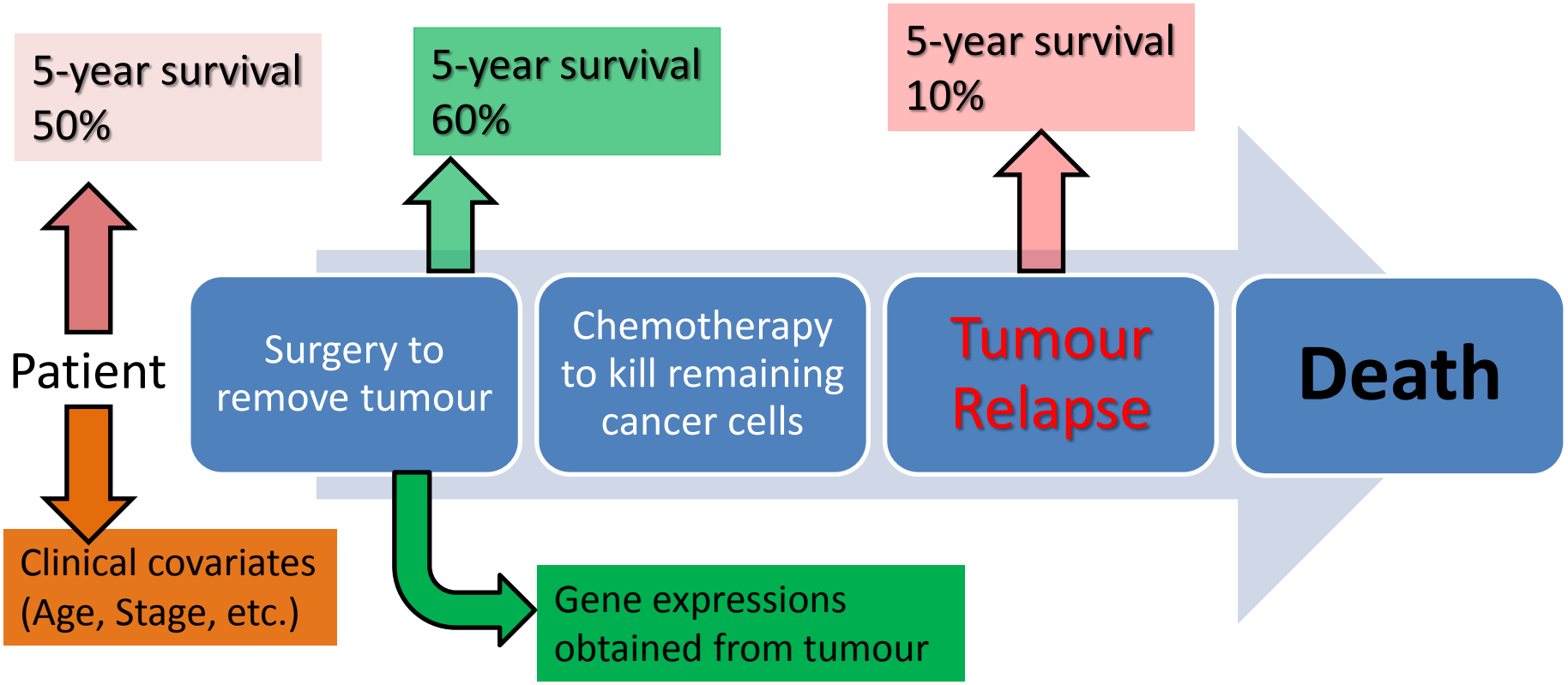
Personalized Prediction of Overall Survival Utilizing Gene Expressions

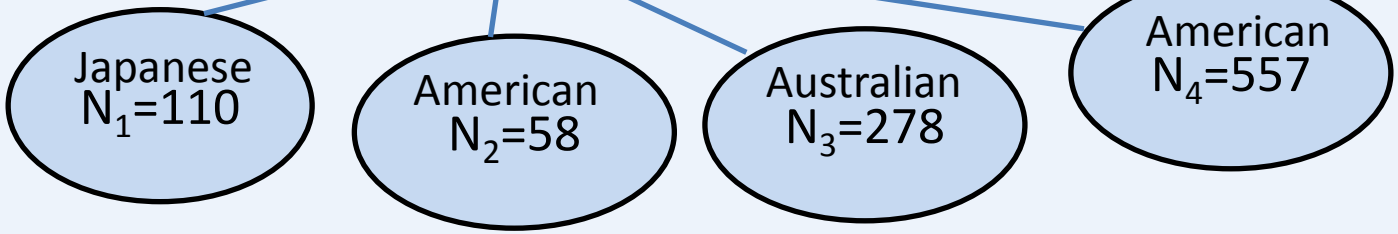
Takeshi Emura
Graduate Institute of Statistics,
National Central University, Taiwan

Joint work with
Masahiro Nakatochi, Shigeyuki Matsui,
Hirofumi Michimae, Virginie Rondeau

Follow-up for a cancer patient



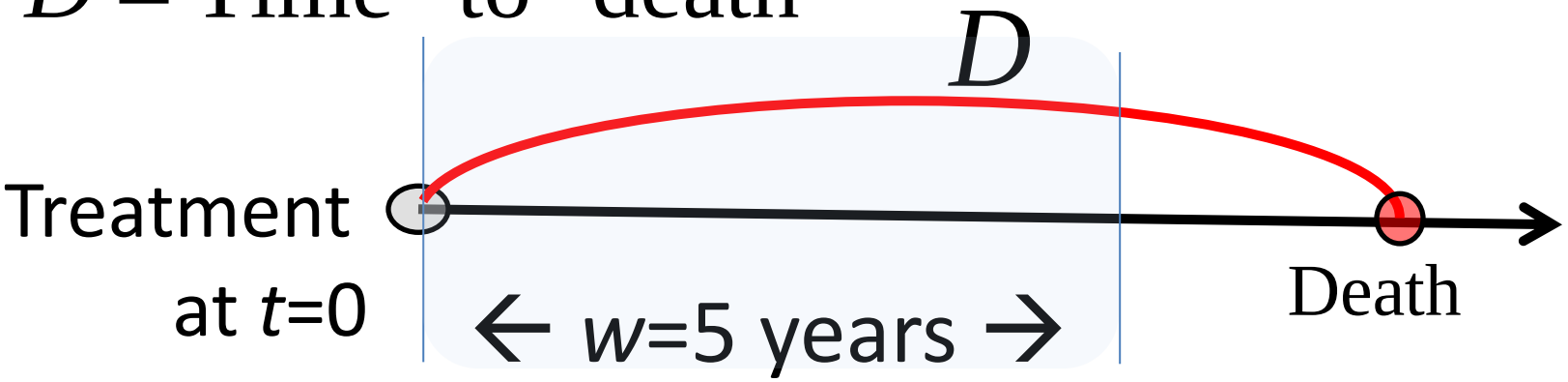
Survival probability = (Clinical, Gene, Relapse, Timing)



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

Classical Survival Prediction

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



- Predict vital status (*death* or *alive)* after 5 years
- *w*-year survival: $S(w \mid \mathbf{Z}) = \Pr(D > w \mid \mathbf{Z})$

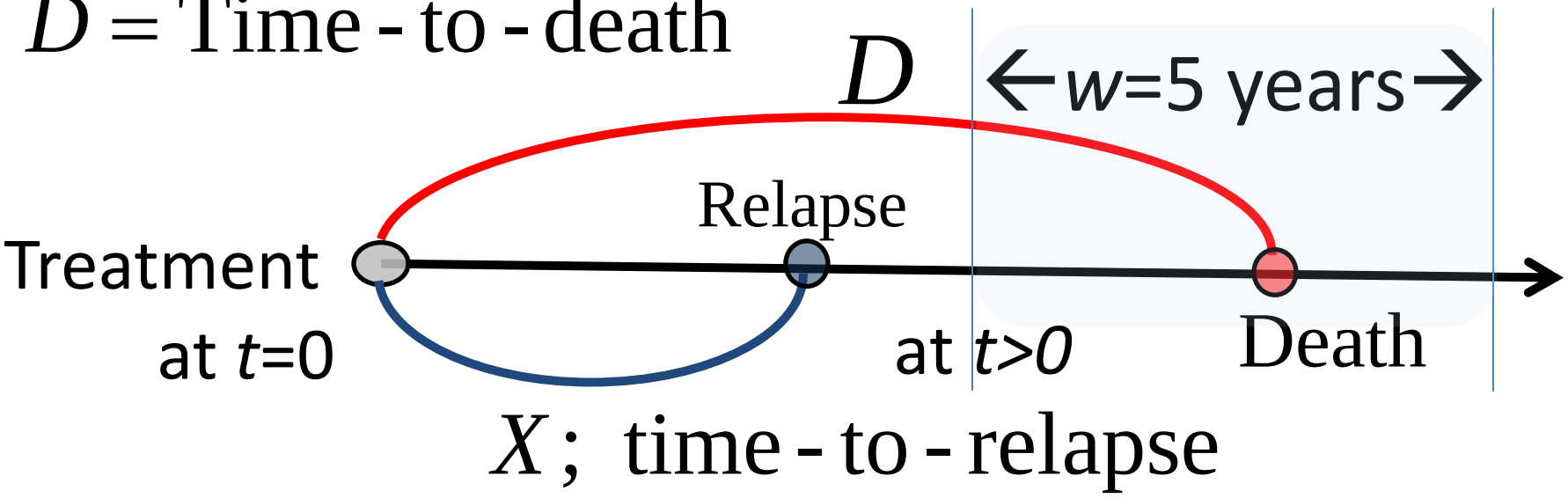
  $\mathbf{Z} =$ (age, stage, tumour size)

- Prediction formula via the Cox model (Cox, 1972)

$$\hat{S}(w \mid \mathbf{Z}) = \exp\{ -\hat{\Lambda}_0(w) e^{\hat{\boldsymbol{\beta}}'\mathbf{Z}} \}$$

Dynamic Prediction

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



X; time-to-relapse

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

↑Conditional failure function (van Houwelingen and Putter 2013)

How to construct the prediction formula?

1) Landmark model (Conditional Cox models fitted at different time points)
2) Time-dependent covariate ? (Cox model is only for exogenous TDC)
3) Joint model (use a copula on (X, D))

Copula model

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

Clayton copula: $C_\theta(v, w) = (v^{-\theta} + w^{-\theta} - 1)^{-1/\theta}$

$$\theta + 1 = \frac{\Pr(X = x, D = y)\Pr(X > x, D > y)}{\Pr(X = x, D > y)\Pr(X > x, D = y)} = \text{Odds ratio in } 2 \times 2 \text{table}$$

$\theta > 0$: Positive dependence

$-1 < \theta < 0$: Negative dependence

- Kendall's tau $= \dfrac{\theta}{\theta + 2}$

	Relapse	Relapse-free
Death	$X=x, D=y$	$X>x, D=y$
Alive	$X=x, D>y$	$X>x, D>y$

Gene expressions predict survival of cancer patients

$$S(t \mid \mathbf{Z}) = \Pr(D > t \mid \mathbf{Z});$$

$$\mathbf{Z} = (Z_1, \ldots, Z_p) : \text{Clinical \& Genetic factors}$$

$$p \text{ can be large } (p > n)$$

- Breast cancer

  (Jenssen et al. 2002; Wang et al 2005; Sabatier et al. 2011)

- Lung cancer

  (Beer et al. 2002; Chen et al. 2007; Shedden et al. 2008)

- Ovarian cancer

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

- **Ovarian cancer data (Ganzfried et al. 2013)**

T_i : time - to - tumour progression or censoring

δ_i : progression status (0 or 1)

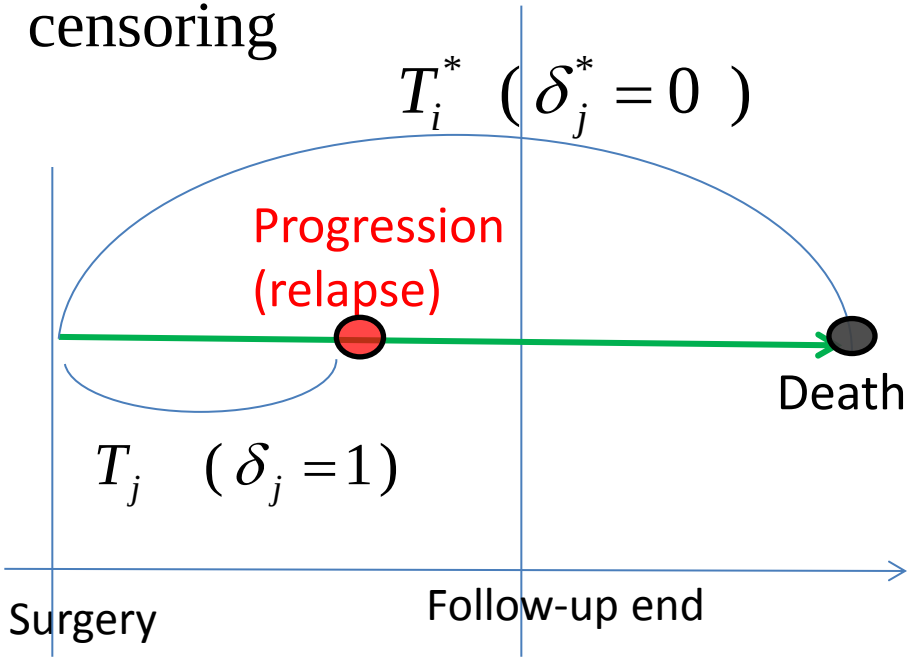
T_i^* : time - to - death or censoring

δ_i^* : vital status (0 or 1)

Z_i : residual tumour size

(1cm < vs. 1cm >=)

$\mathbf{V}_i = (V_{i1}, ..., V_{ip})'$: gene expressions



T_i^* Time-to-death	δ_i^* Vital status	AP3S1	APMAP	ARHGAP28	CXCL12	……..	ASB7	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

Ovarian cancer data (Ganzfried et al., 2013)

Sample size	The number of observed events (event rates)			The number
	Relapse	Death	Censoring	of genes
Japanese $N_1 = 84$	59 (70%)	38 (45%)	46 (55%)	18,548
American $N_2 = 58$	48 (83%)	36 (62%)	22 (38%)	18,524
Australian $N_3 = 260$	185 (71%)	113 (43%)	147 (57%)	18,524
American $N_4 = 510$	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

Notes: The data are extracted from R Bioconductor *curatedOvarianData* package

Heterogeneity (random effects) Dependence (copula) High-dimensionality

Methods for high-dimensional covariates

- Lasso (Cox-regression with L_1 penalty)

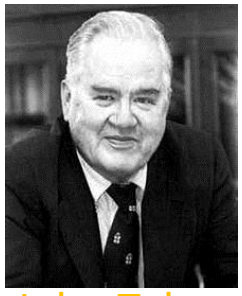
  Tibshirani (1997 Stat Med), Gui & Li (2005 Bioinformatics)

- Ridge regression (Cox-regression with L_2 penalty)

  Verveij & van Howelingen(1994 Stat. Med.), Zhao et al. (2011 PONE)

- Univariate selection (forward selection via univariate Cox – regression Jenssen et al. (2002 Nature Med), Chen et al. (2007 NEJM)

- Compound covariate (adopted for this research)


John Tukey

Tukey (1993 Controlled Clinical Trial), Matsui (2006, BMC Bioinfomatics), Simon et al (2011 Boinfo), Matsui et al (2012 Clin Can Res) Emura et al (2012 PONE)

Proposed method (1/3)

- **Step 1: Univariate Selection**

$$\begin{cases} \mathbf{V}_{ij} = (V_{ij,1}, \ldots, V_{ij,q_1}) & : \text{associated with relapse } X_{ij} \\ \mathbf{W}_{ij} = (W_{ij,1}, \ldots, W_{ij,q_2}) & : \text{associated with death } D_{ij} \end{cases}$$

$$\begin{cases} r_{ij}(t) = r_0(t)\exp(b_k V_{ij,k}), & q_1 : \text{the number of genes with } P < 0.001 \\ \lambda_{ij}(t) = \lambda_0(t)\exp(c_k W_{ij,k}), & q_2 : \text{the number of genes with } P < 0.001 \end{cases}$$

for the *i*-th study, *j*-th patient, and *k*-th gene

P=0.001 : as recommended in Simon (2003)

- **Step 2: compound covariate (CC) predictors**

$$\begin{cases} \mathrm{CC}_{1,ij} = \hat{b}_1 V_{ij,1} + \cdots + \hat{b}_{q_1} V_{ij,q_1} & : \text{associated with relapse } X_{ij} \\ \mathrm{CC}_{2,ij} = \hat{c}_1 W_{ij,1} + \cdots + \hat{c}_{q_2} W_{ij,q_2} & : \text{associated with death } D_{ij} \end{cases}$$

coefficients from the univariate Cox models

Proposed method (2/3)

- **Step 3**: Fit the joint frailty-copula model
  (Emura et al. 2015 *SMMR*)

$$\begin{cases} r_{ij}(t\,|\,u_i) = u_i r_0(t)\exp(\boldsymbol{\beta}_1'\mathbf{Z}_{1,ij} + \gamma_1 \mathrm{CC}_{1,ij}) & \text{for } X_{ij} \\ \lambda_{ij}(t\,|\,u_i) = u_i^{\alpha}\lambda_0(t)\exp(\boldsymbol{\beta}_2'\mathbf{Z}_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & \text{for } D_{ij} \\ \Pr(X_{ij} > x, D_{ij} > y\,|\,u_i) = C_\theta[S_X(x\,|\,u_i), S_D(y\,|\,u_i)] \end{cases}$$

for the ***i***-th study and ***j***-th patient

The Clayton copula: $C_\theta(v, w) = (v^{-\theta} + w^{-\theta} - 1)^{-1/\theta}, \quad \theta \geq 0$

Estimator $(\hat{\theta}, \hat{\eta}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\gamma}_1, \hat{\gamma}_2, \hat{r}_0, \hat{\lambda}_0)$

→ R package *joint.Cox* (Emura, 2017 on CRAN)

,

Proposed method (3/3)

- If the patient does not experience tumour progression before *t*,

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

$$= \frac{\int_0^\infty \left(C_\theta[\, S_X(t \mid u), S_D(t \mid u)\,] - C_\theta[\, S_X(t \mid u), S_D(t+w \mid u)\,] \right) f_\eta(u) du}{\int_0^\infty C_\theta[\, S_X(t \mid u), S_D(t \mid u)\,] f_\eta(u) du}$$

$(\hat{\theta}, \hat{\eta}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\gamma}_1, \hat{\gamma}_2, \hat{r}_0, \hat{\lambda}_0)$

- If the patient experiences tumour progression before *t*,

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

$$= \frac{\int_0^\infty \left(C_\theta^{[1,0]}[\, S_X(x \mid u), S_D(t \mid u)\,] - C_\theta^{[1,0]}[\, S_X(x \mid u), S_D(t+w \mid u)\,] \right) u S_X(x \mid u) f_\eta(u) du}{\int_0^\infty C_\theta^{[1,0]}[\, S_X(x \mid u), S_D(t \mid u)\,] u S_X(x \mid u) f_\eta(u) du}$$

Data analysis (Ganzfried et al., 2013)

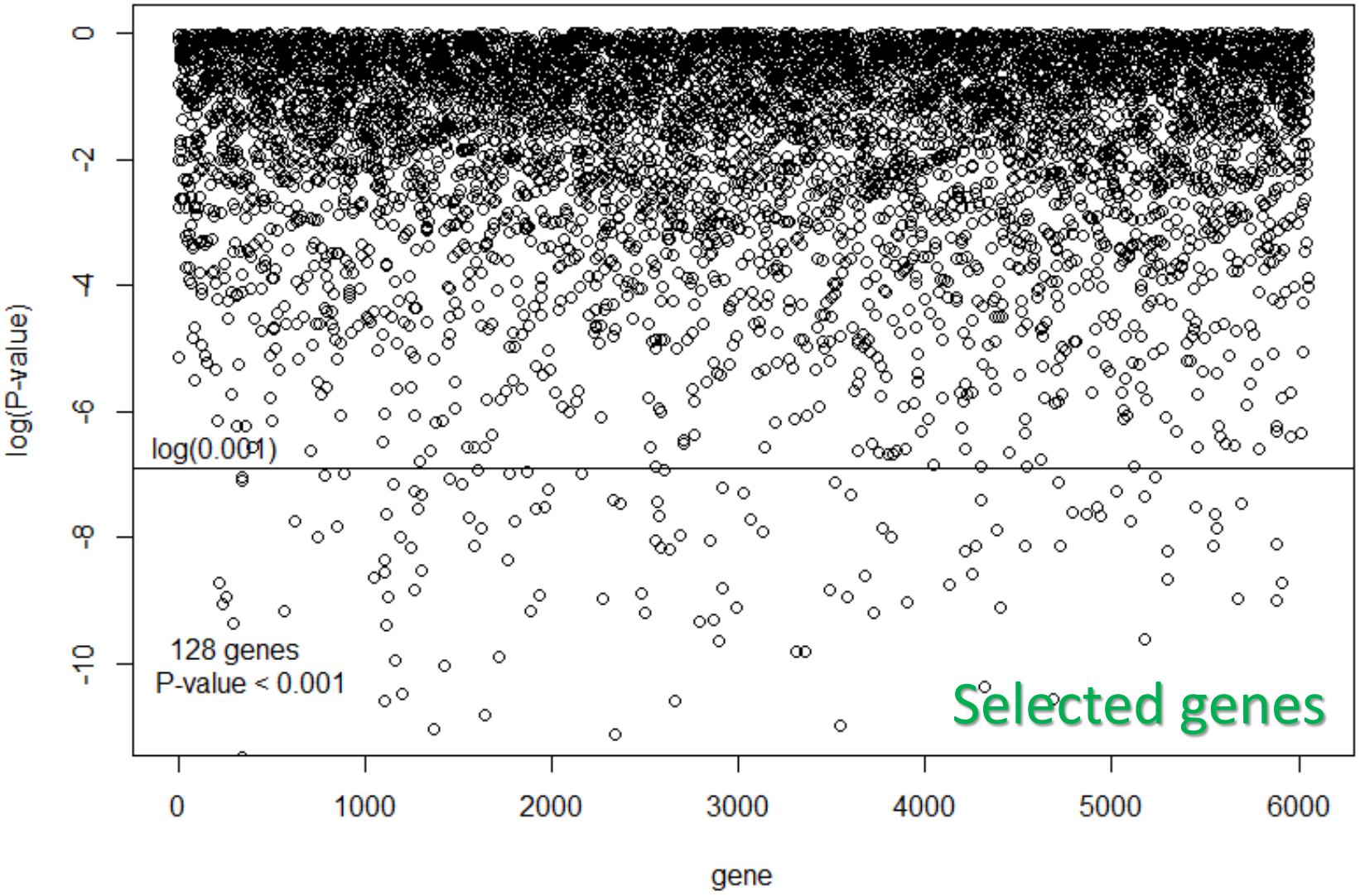
A meta-analytic data combining the four independent studies of ovarian cancer patients

	Sample size	The number of observed events (event rates)			The number of genes
		Relapse	Death	Censoring	
Study 1	$N_1 = 84$	59 (70%)	38 (45%)	46 (55%)	18,548
Study 2	$N_2 = 58$	48 (83%)	36 (62%)	22 (38%)	18,524
Study 3	$N_3 = 260$	185 (71%)	113 (43%)	147 (57%)	18,524
Study 4	$N_4 = 510$	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

Notes: The data are extracted from R Bioconductor *curatedOvarianData* package

Select genes with P-value =0.001

Univariate association between gene and time-to-death



Data Analysis: model fitting

Joint frailty-copula model

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

Clinical covariate:

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

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).

Data Analysis: model fitting

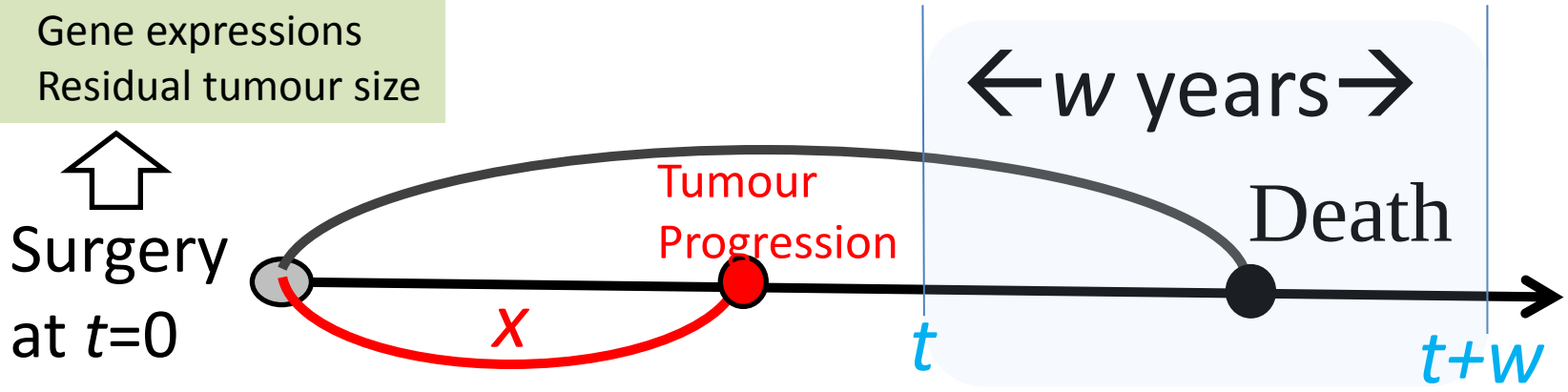
$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\gamma_1 \mathrm{CC}_{1,ij}) & (\text{for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = \lambda_0(t) \exp(\beta_2 Z_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & (\text{for time to death } D_{ij}) \end{cases}$$

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

	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

Estimated prediction formula

- Gene expressions
- Residual tumour size



Estimated conditional failure function

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

$$= \frac{\int_0^\infty \left(C_{\hat{\theta}}^{[1,0]}[\, \hat{S}_X(x \mid u), \hat{S}_D(t \mid u)\,] - C_{\hat{\theta}}^{[1,0]}[\, \hat{S}_X(x \mid u), \hat{S}_D(t+w \mid u)\,] \right) u \hat{S}_X(x \mid u) f_{\hat{\eta}}(u) du}{\int_0^\infty C_{\hat{\theta}}^{[1,0]}[\, \hat{S}_X(x \mid u), \hat{S}_D(t \mid u)\,] u \hat{S}_X(x \mid u) f_{\hat{\eta}}(u) du},$$

$$\hat{S}_X(t \mid u) = \exp\left\{ -u \hat{R}_0(t) \exp(\hat{\gamma}_1 \mathrm{CC}_1) \right\},$$

$$\hat{S}_D(t \mid u_i) = \exp\left\{ -u^{\hat{\alpha}} \hat{\Lambda}_0(t) \exp(\beta_2 Z_2 + \hat{\gamma}_2 \mathrm{CC}_2) \right\},$$

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

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

} Compound covariate

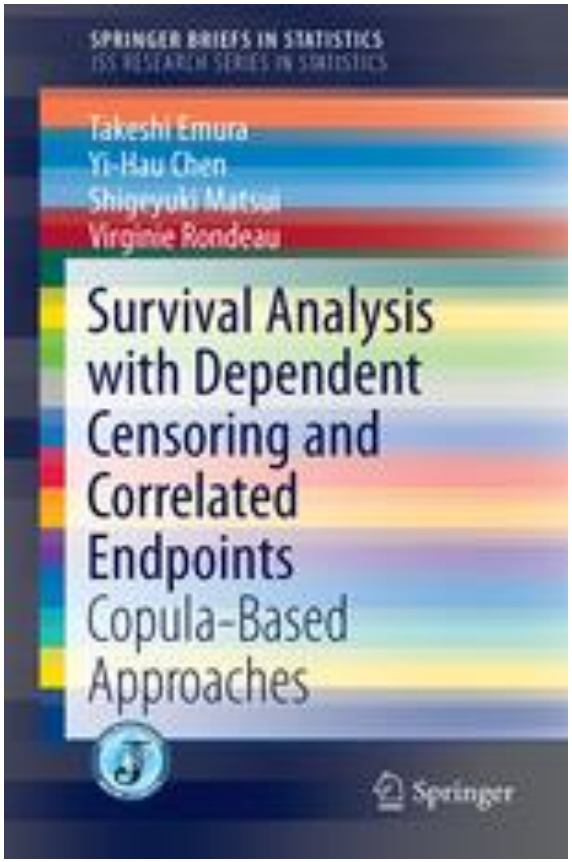
- **Patient 1**: risk genes ($CC_1 = 10$, $CC_2 = 10$); the residual tumour>1cm ($Z_2 = 1$).
- **Patient 2**: protective genes ($CC_1 = -10$, $CC_2 = -10$); the residual tumour≤1cm ($Z_2 = 0$).

```
library(joint.Cox)
gamma1=0.39 # coefficient for CC1
beta2=0.16 # coefficient for residual tumour
gamma2=0.44 # coefficient for CC2
theta=1.9 # copula parameter
eta=0.04 # frailty parameter
g=c(0.85, 2.14, 0, 0.07, 0) # hazard for TTP
h=c(0.17, 1.05, 1.24, 0.27, 0) # hazard for OS
xi1=0 ### lower limit of t ###
xi3=6420 #### upper limit of t+w ###
mu1=0.338 # mean of CC1
SD1=10.468 # SD of CC1
mu2=0.222 # mean of CC2
SD2=7.894 # SD of CC2

time=1000
w_num=20
widths=seq(0,xi3-time,length=w_num)

#### Patient 2 ####
CC1=-10;CC2=-10;Z2=0
X=600 ### relapse at 600 days ###

F.prediction(time=time,width=widths,
             Z1=(CC1-mu1)/SD1,Z2=c((CC2-mu2)/SD2,Z2),X=X,

beta1=gamma1,beta2=c(beta2,gamma2),eta=eta,theta=theta,alpha=0,
             g=g,h=h,xi1=0,xi3=xi3,Fplot=FALSE)
```

Parameters in the joint frailty-copula model

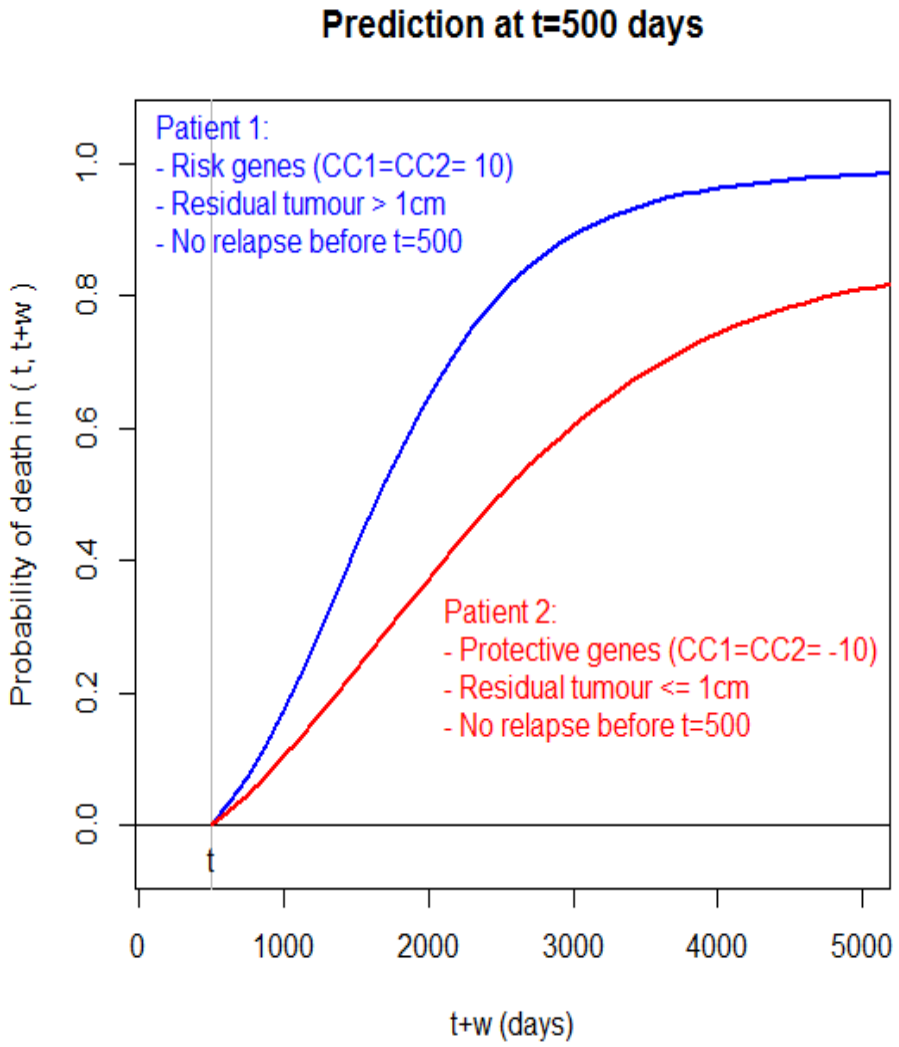
Prediction time

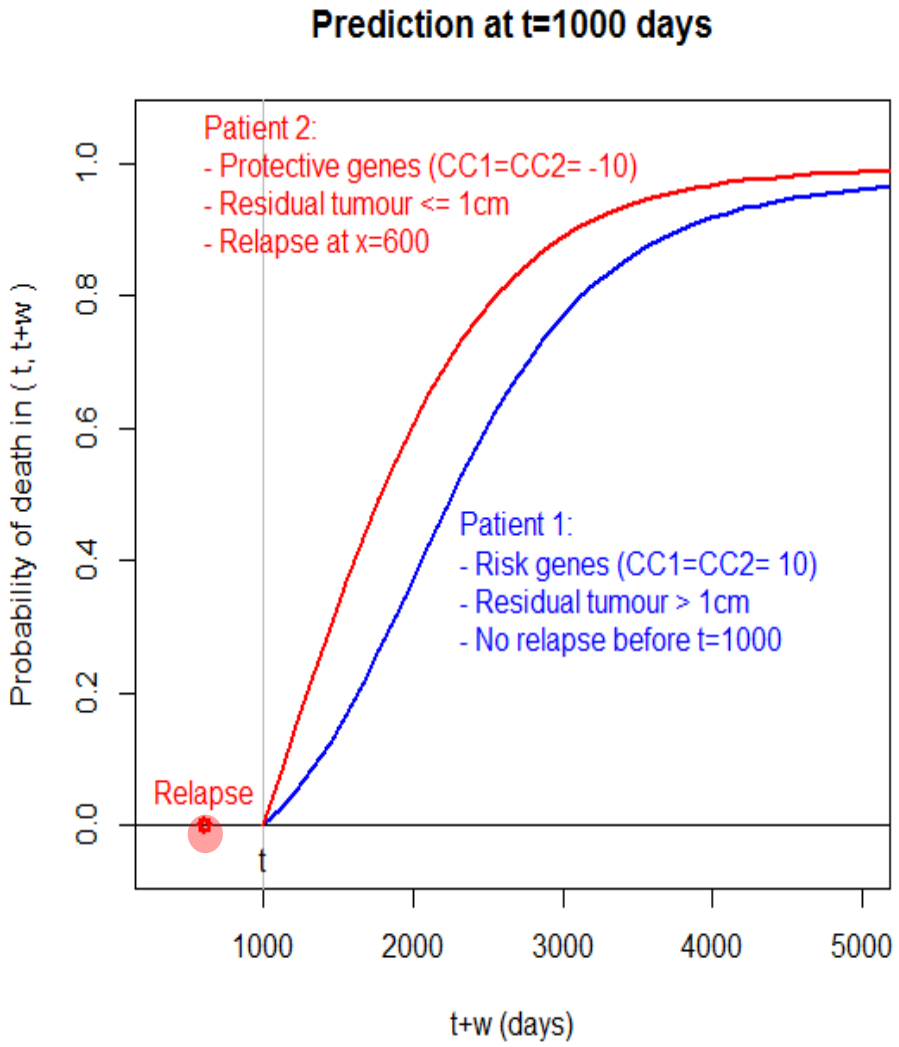
Patient information



Emura T, Chen YH, Matsui S, Rondeau V (2017+)
JSS Research Series in Statistics, Springer

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





Come to close. Thank you !