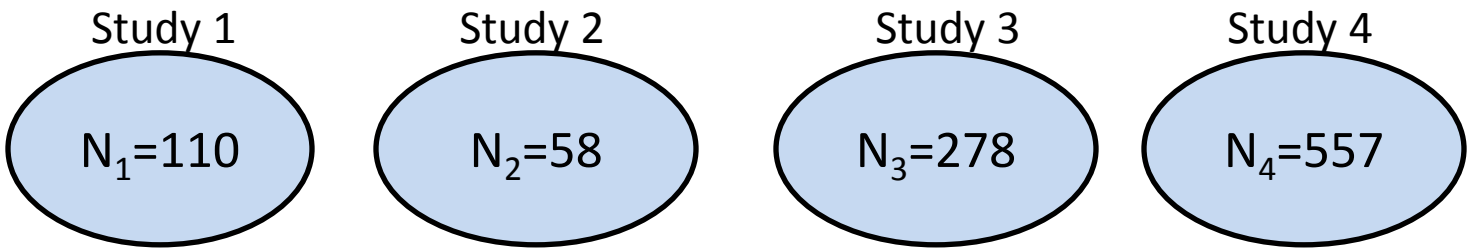
CM Statistics 2016, Seville, Spain, 2016/12/9

Dynamic prediction according to tumour progression and genetic factors

: Meta-analysis with a joint frailty-copula

(Meta-analytic data = Clustered survival data)



Takeshi Emura

Graduate Institute of Statistics,

National Central University, Taiwan

Joint work with

Masahiro Nakatochi, Shigeyuki Matsui,

Hirofumi Michimae, Virginie Rondeau

Outline

Review

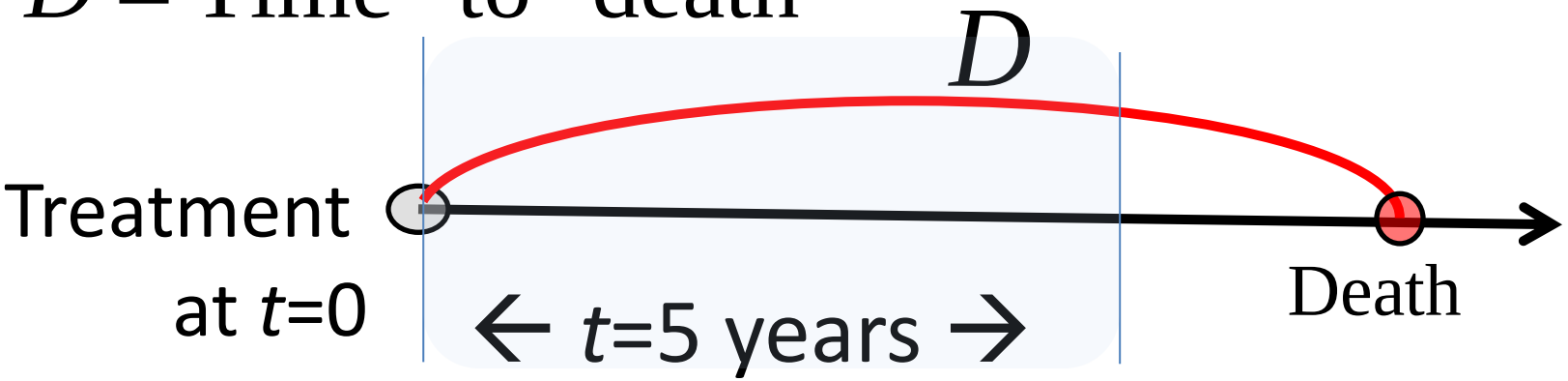
* Dynamic Prediction
* Copula and survival model
* High-dimensional problem

Proposed method

* Tukey's compound covariate
* Proposed dynamic prediction formula
* Ovarian cancer data analysis

Classical Survival Prediction

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



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

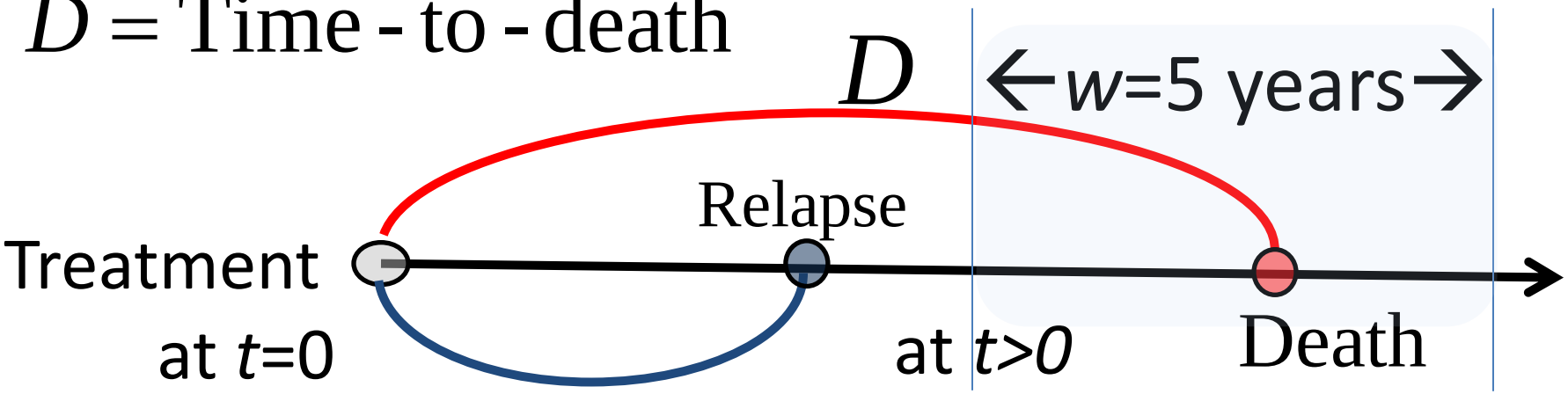
  $\mathbf{Z} =$ (age, sex, stage, tumour size)

  Graf et al. (1999); Gerts and Schumacher (2006)
- Cox proportional hazards model (Cox, 1972)

$$S(t \mid \mathbf{Z}) = S(t \mid \mathbf{0})^{\exp(\boldsymbol{\beta}'\mathbf{Z})}$$

Dynamic Prediction of Death

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



X; time-to-relapse

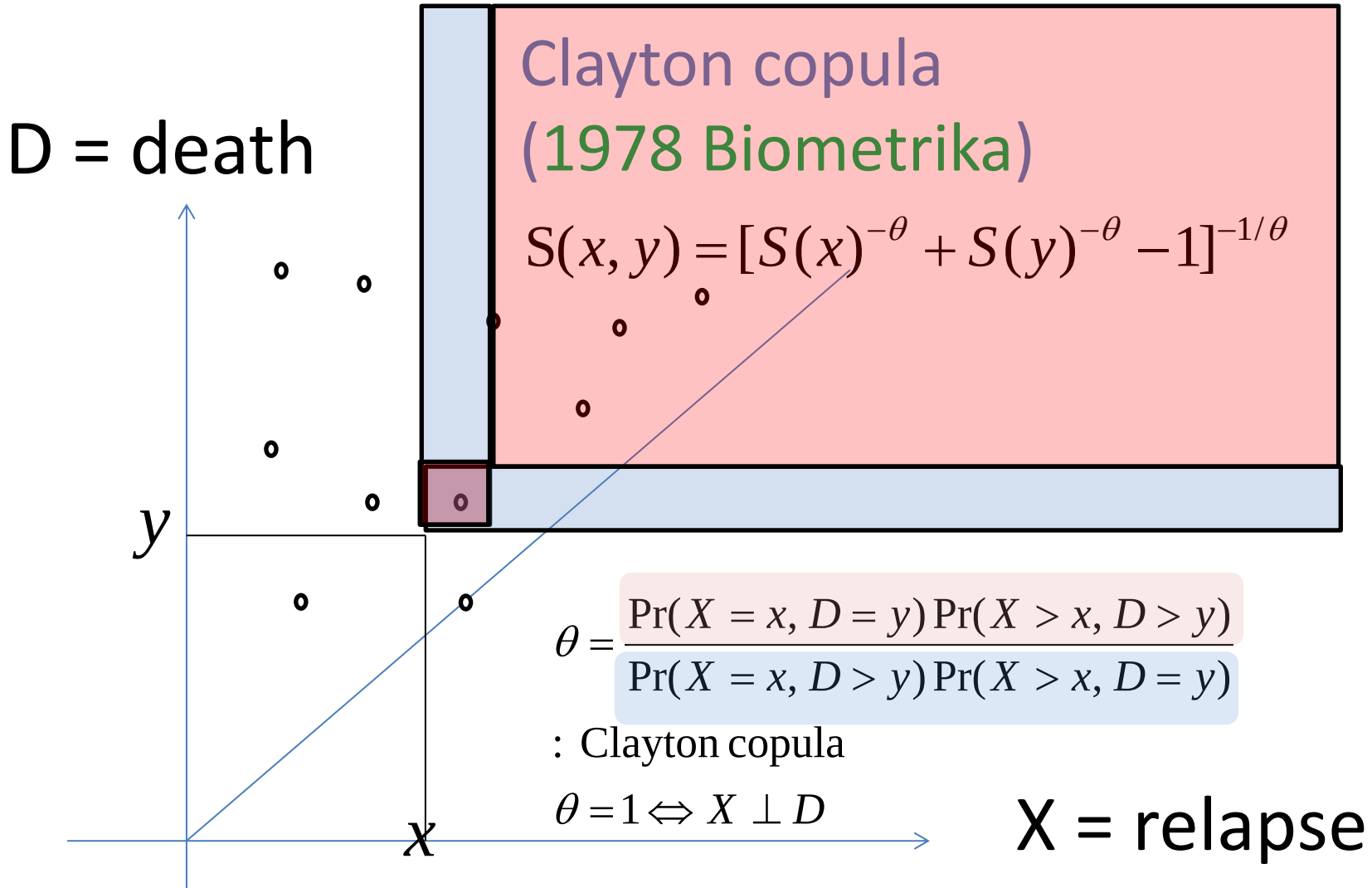
$$F(t, t+w \mid X, \mathbf{Z}) = \Pr(D \le t+w \mid D > t, X, \mathbf{Z})$$

- Predict *death probability* within a time window (5 years) at a certain moment $t>0$:
  (van Houwelingen and Putter 2013)
- Accurate prediction achieved using a joint frailty model between X and D (Mauguen et al., 2013, 2015)

Copulas in survival model

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

$$\Leftrightarrow S(x, y) = C[S_X(x), S_D(y)]$$



Genetic factors

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

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

p can be large ($p > n$)

Genes are informative for survival prediction in

- Breast cancer (Jenssen et al. 2002; Sabatier et al. 2011)
- Diffuse large-B-cell lymphoma
  (Lossos et al. 2004; Binder and Schumacher 2008; Alizadeh 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)

Objective

(Genetic factors) + (Dynamic prediction) = Personalized dynamic prediction

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

X : time-to-tumour progression

$\mathbf{Z}$: clinical & genomic factors

- Landmark approach (van Houwelingen and Putter 2013)

  ※ Conditional model: $(D \mid X, \mathbf{Z})$

  Chap 12: Dynamic prediction when $\mathbf{Z}$ is high-dimension

- Ours: Joint model approach

  ※ joint model: $(D, X \mid \mathbf{Z})$

Dynamic prediction via joint models

	Response	Souse of Dependence	Meta-analysis
Rizopoulos (2011, Biometrics) Taylor et al. (2013, SMMR) Sène et al. (2014, SMMR) Proust-Lima (2014, SMMR)	Longitudinal measurements + Time-to-events	Frailty	No
Mauguen et al. (2013, 2015) Król et al. (2016, Biometrics) Mazroui et al. (2015 LTDA)	Recurrent events + Time-to-death	Frailty	No
This research: **Dynamic prediction using genetic factors**	Time-to-event + Time-to-death	Copula → Subject-level Frailty → Study-level	Yes → frailty

- Existing dynamic predictions do not adapt to "meta-analysis", requiring two sources of dependence (Subject-level dependence and Study-level dependence)
- Existing dynamic predictions do not adapt to "high-dimensional factors"

Motivating example (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
		Relapse	Death	Censoring	of genes
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

Cluster effect (frailty)

Subject-level Dependence (copula)

High-dimensionality (compound covariate)

Data structure

X_{ij} = TTP (Time to tumour progression, e.g., relapse)

D_{ij} = time - to - death

C_{ij} = independent censoring time (e.g., study end)

$\mathbf{Z}_{ij}$ = clinical ovariates (e.g., age, cancer stage)

Under semicompeting risks (Fine et al., 2001) :

* First occuring event time

Indicator of tumour progression

$$T_{ij} = \min(\, X_{ij}, D_{ij}, C_{ij}\,), \qquad \delta_{ij} = \mathbf{I}(T_{ij} = X_{ij})$$

* Terminal event time

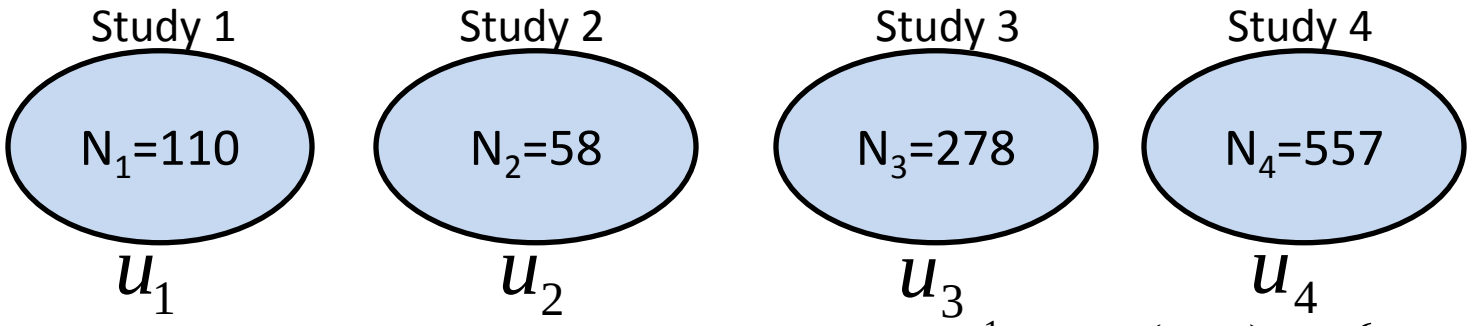
Indicator of death

$$T^*_{ij} = \min(\, D_{ij}, C_{ij}\,), \qquad \delta^*_{ij} = \mathbf{I}(T^*_{ij} = D_{ij})$$

$$(\, T_{ij}\,, T^*_{ij}\,, \delta_{ij}\,, \delta^*_{ij}\,, \mathbf{Z}_{ij}\,), \quad i = 1, 2, ..., G, \quad j = 1, 2, ..., N_i$$

$$(\text{e.g.}, G = 4; \quad N_1 = 84, N_2 = 58, N_3 = 260, N_4 = 510)$$

- Cluster (study) characterized by a frailty



$$\text{Gamma frailty}: \; u_i \sim f_\eta(u) = \frac{1}{\Gamma(1/\eta)\eta^{1/\eta}} u^{\frac{1}{\eta}-1} \exp\left(-\frac{u}{\eta}\right), \quad \begin{cases} E[u_i]=1 \\ Var[u_i]=\eta \end{cases}$$

Joint frailty-copula model (Emura et al., 2015 SMMR)

$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\boldsymbol{\beta}_1' \mathbf{Z}_{1,ij}) & (\text{hazard for } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = u_i^{\alpha} \lambda_0(t) \exp(\boldsymbol{\beta}_2' \mathbf{Z}_{2,ij}) & (\text{hazard for } D_{ij}) \\ \Pr(X_{ij} > x, D_{ij} > y \mid u_i) = C_\theta[R_{ij}(x \mid u_i), \Lambda_{ij}(y \mid u_i)] & \text{Copula} \end{cases}$$

$\boldsymbol{\beta}_1$ = Effect on time-to-progression X_{ij}

$\boldsymbol{\beta}_2$ = Effect on time-to-death D_{ij}

α = Intra-study dependence

θ = Intra-subject dependence

Outline

Review

* Dynamic Prediction
* Copula and survival model
* High-dimensional problem

Proposed method

* Tukey's compound covariate (CC)
* Proposed dynamic prediction formula
* Ovarian cancer data analysis

High-dimensional genetic factors

- **Step 1**: Select genetic factors

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

Univariate Cox regressions:

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

P=0.001 : due to Simon (2003)

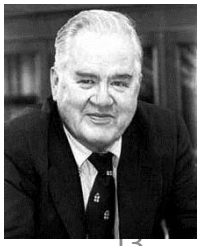
- **Step 2: Tukey's compound covariate (CC)**

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

coefficients from univariate Cox models

CC: Tukey (1993 Controlled Clinical Trial), Matsui (2006, BMC Bioinfomatics),

Simon et al (2011 Boinfo), Matsui etal (2012 Clin Can Res), Emura et al (2012), just name a few



John Tukey

Proposed model with high-dimensional genetic factors

- Joint frailty-copula model

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

- Penalized maximum likelihood estimation under the Clayton copula

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

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, 2016 on CRAN)

Proposed dynamic prediction

Goal: **Predicting the probability of death for a new patient (not in the data)**

i) The patient's covariates measured at time 0

$$\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \mathrm{CC}_1, \mathrm{CC}_2)$$

ii) Tumour progression history at time $t>0$

$$H(t,x) = \begin{cases} X \le t,\ X = x & ;\ \text{tumour progression occurred at } x < t, \\ X > t & ;\ \text{tumour progression did not occurre before } t. \end{cases}$$

The patient's prob. of death between t and $t+w$

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

Prediction formulas under joint frailty-copula model

- Tumour progression does not occur before *t*,

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

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

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

- Tumour progression occurs before *t*,

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

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

where $C_\theta^{[1,0]}(v, w) = \partial C_\theta(v, w) / \partial v$

Assessing Prediction Error

Brier score (Graf et al. 1999, Stat. Med.)

$$Err(t, t+w)$$

$$= E[\{\mathbf{I}(D > t+w) - \hat{S}(t, t+w \mid H(t, X), \mathbf{Z})\}^2 \mid D > t]$$

where

$$\hat{S}(t, t+w \mid X, \mathbf{Z}) = 1 - \hat{F}(t, t+w \mid X, \mathbf{Z})$$

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

- MSE of predicting dichotomous event (death or alive) in [*t*, *t+w*].
- E[ ] is over the new patient (*D*, *X*, ***Z***).
- $\hat{S}$ is given: randomness of $\hat{S}$ is not accounted in E[ ].

Assessing Prediction Error

Estimate of Brier score

$$\hat{Err}(t,t+w)=\frac{1}{Y(t)}\sum_{ij}\mathbf{I}(T_{ij}^{*}>t)\hat{w}_{ij}(t,t+w)\{\mathbf{I}(T_{ij}^{*}>t+w)-\hat{S}(t,t+w\,|\,H(t,T_{ij}),\mathbf{Z}_{ij})\}^{2}$$

$$\text{where}\quad \hat{w}_{ij}(t,t+w)=\frac{\delta_{ij}^{*}\hat{G}(t)}{\hat{G}(T_{ij}^{*})}\mathbf{I}(T_{ij}^{*}\le t+w)+\frac{\hat{G}(t)}{\hat{G}(t+w)}\mathbf{I}(T_{ij}^{*}>t+w)$$

IPCW: Graf et al. (1999); Gerts and Schumacher (2006)

- **Optimism bias**: evaluated by cross-validation

$$Err(t,t+w)=\hat{Err}(t,t+w)+\text{op}>\hat{Err}(t,t+w)$$ due to overfitting

- **Variability**: evaluated by the bootstrap 95% CI:

$$\hat{Err}^{(b)}(t,t+w)=\frac{1}{Y^{(b)}(t)}\sum_{ij}\mathbf{I}(T_{ij}^{*(b)}>t)\hat{w}_{ij}(t,t+w)\{\mathbf{I}(T_{ij}^{*(b)}>t+w)-\hat{S}(t,t+w\,|\,H(t,T_{ij}^{(b)}),\mathbf{Z}_{ij}^{(b)})\}^{2}$$

Random sampling with replacement

$$(T_{ij},T_{ij}^{*},\delta_{ij},\delta_{ij}^{*},\mathbf{Z}_{ij}),\quad T_{ij}^{*}>t\Rightarrow(T_{ij}^{(b)},T_{ij}^{*(b)},\delta_{ij}^{(b)},\delta_{ij}^{*(b)},\mathbf{Z}_{ij}^{(b)}),\quad T_{ij}^{*(b)}>t:\quad b=1,\ldots,1{,}000$$

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

Data Analysis: model fitting

Joint frailty-copula model (after variable selection)

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

Clinical covariate:

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

Compound covariate (CC):

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

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

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

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

Data Analysis: model fitting

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

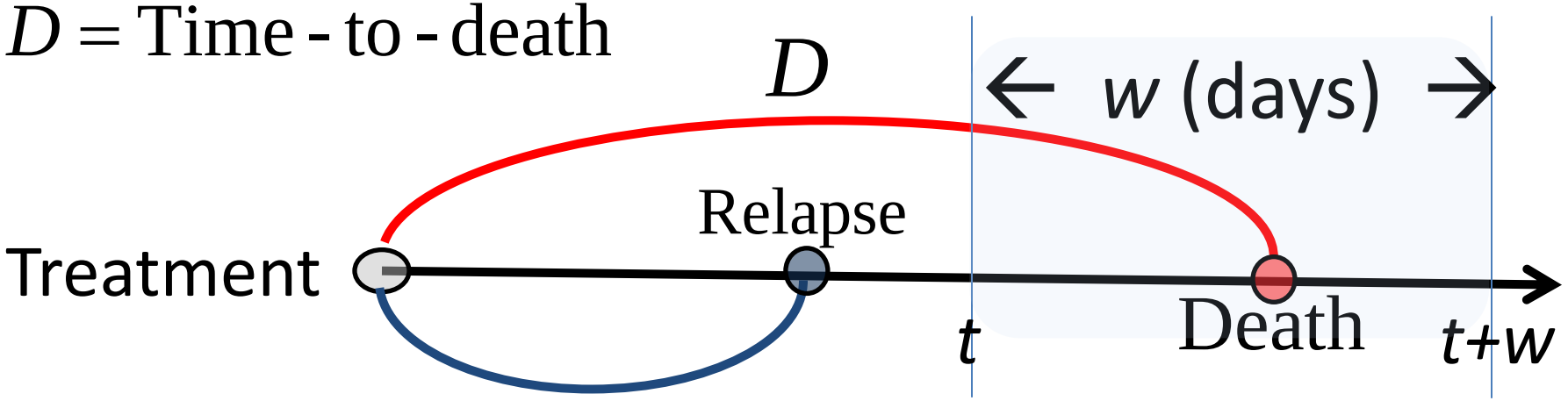
$$\Pr(X_{ij} > x, D_{ij} > y \mid u_i) = C_\theta[S_X(x \mid u_i), S_D(y \mid u_i)]$$

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

Prediction settings

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



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

$$H(t);\ \text{relapse history before } t$$

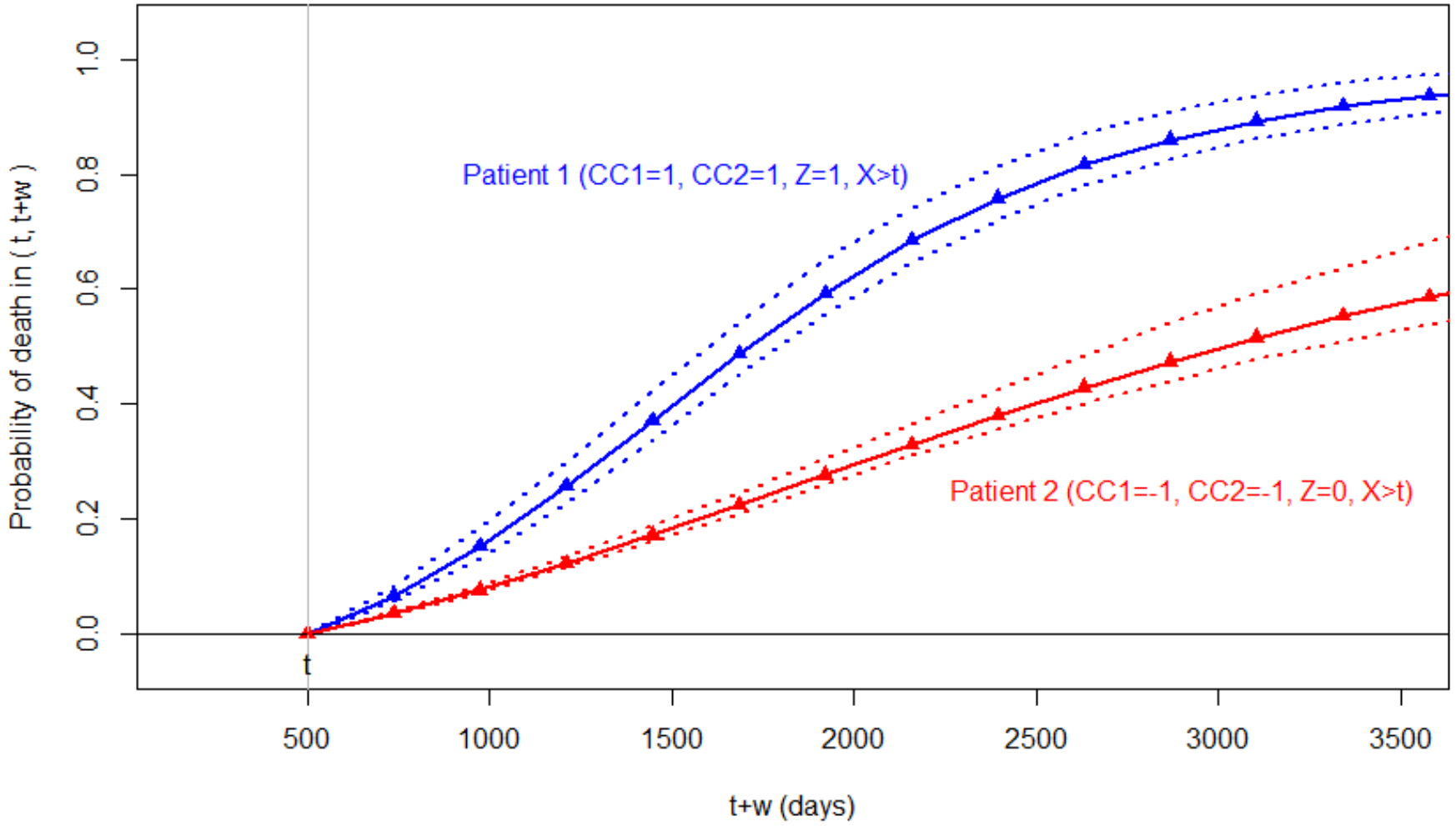
- t=500 days (early prediction time)

  $500 < t+w < 3500$
- t=1000 days (late prediction time)

  $1000 < t+w < 3500$

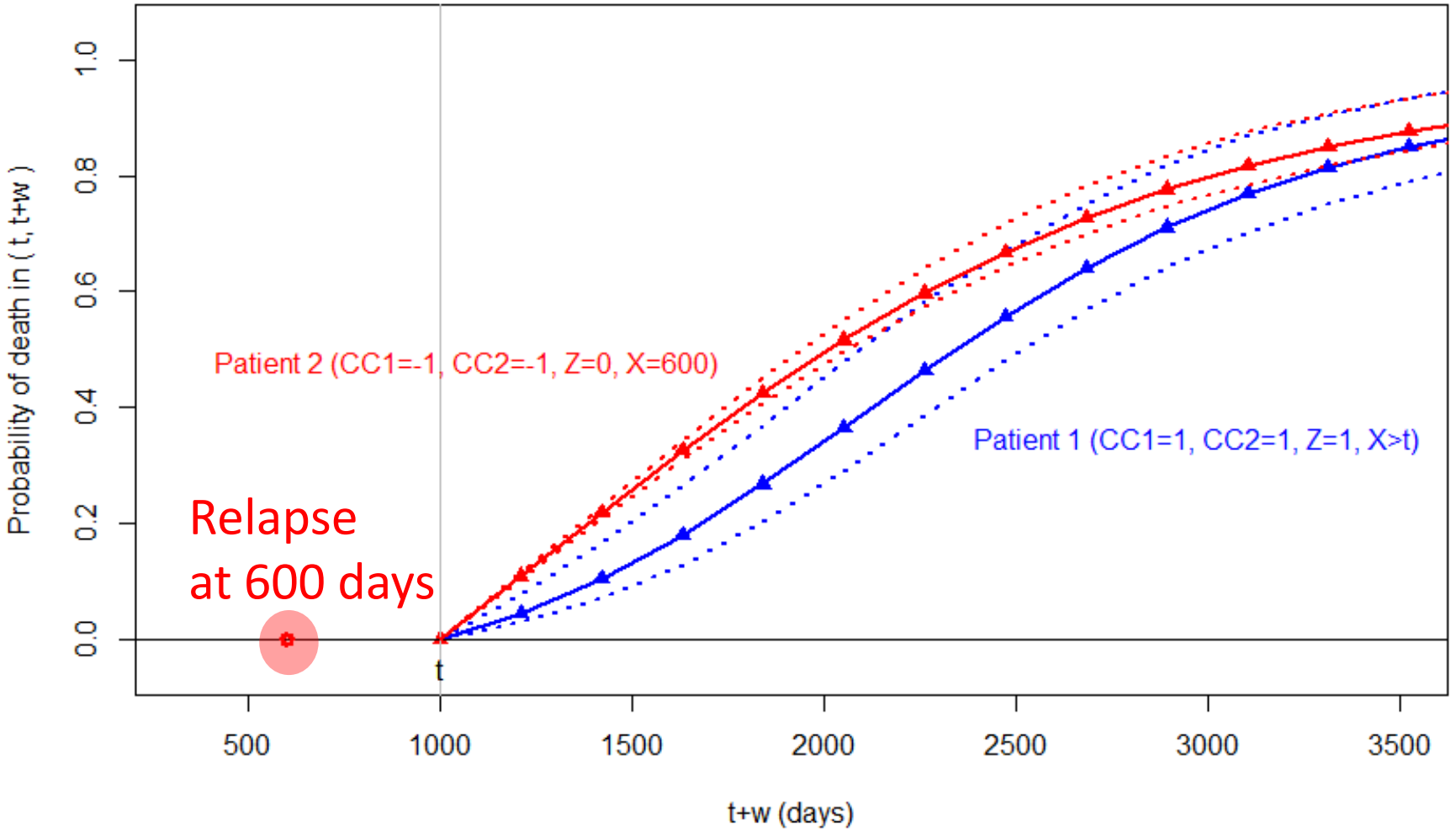
Early prediction time at t = 500 (days)

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



Late prediction time at t = 1000 (days)

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



Prediction error comparison

1. Null model (Kaplan-Meier estimator)

$$\begin{cases} r_{ij}(t \mid u_i) = r_0(t) & (\text{for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = \lambda_0(t) & (\text{for time to death } D_{ij}) \end{cases}$$

2. **Simple model** (*CXCL12* gene alone) considered in Emura et al. (2015 SMMR)

$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\gamma_1 \mathrm{CXCL12}_{ij}) & (\text{for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = \lambda_0(t) \exp(\gamma_2 \mathrm{CXCL12}_{ij}) & (\text{for time to death } D_{ij}) \end{cases}$$

3. Model with high-dimensional genetic factors (proposed)

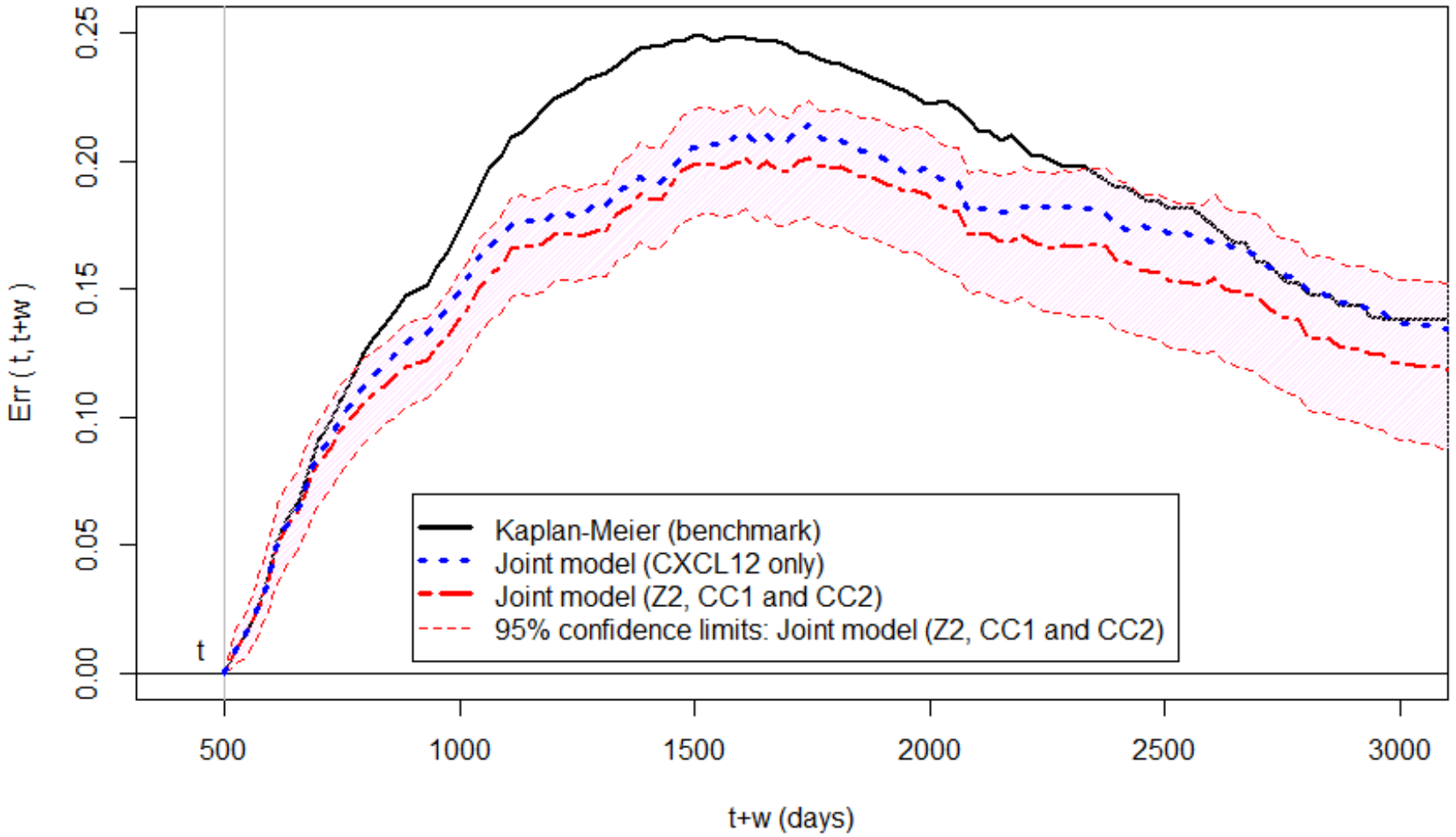
$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\gamma_1 \mathrm{CC}_{1,ij}) & (\text{for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = \lambda_0(t) \exp(\boldsymbol{\beta}_2' \mathbf{Z}_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & (\text{for time to death } D_{ij}) \end{cases}$$

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

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

Prediction error at *t*=500 (days)

$$\hat{Err}(t,t+w)=\frac{1}{Y(t)}\sum_{ij}\mathbf{I}(T_{ij}^{*}>t)\hat{w}_{ij}(t,t+w)\{\mathbf{I}(T_{ij}^{*}>t+w)-\hat{S}(t,t+w\,|\,H(t,T_{ij}),\mathbf{Z}_{ij})\}^{2}$$



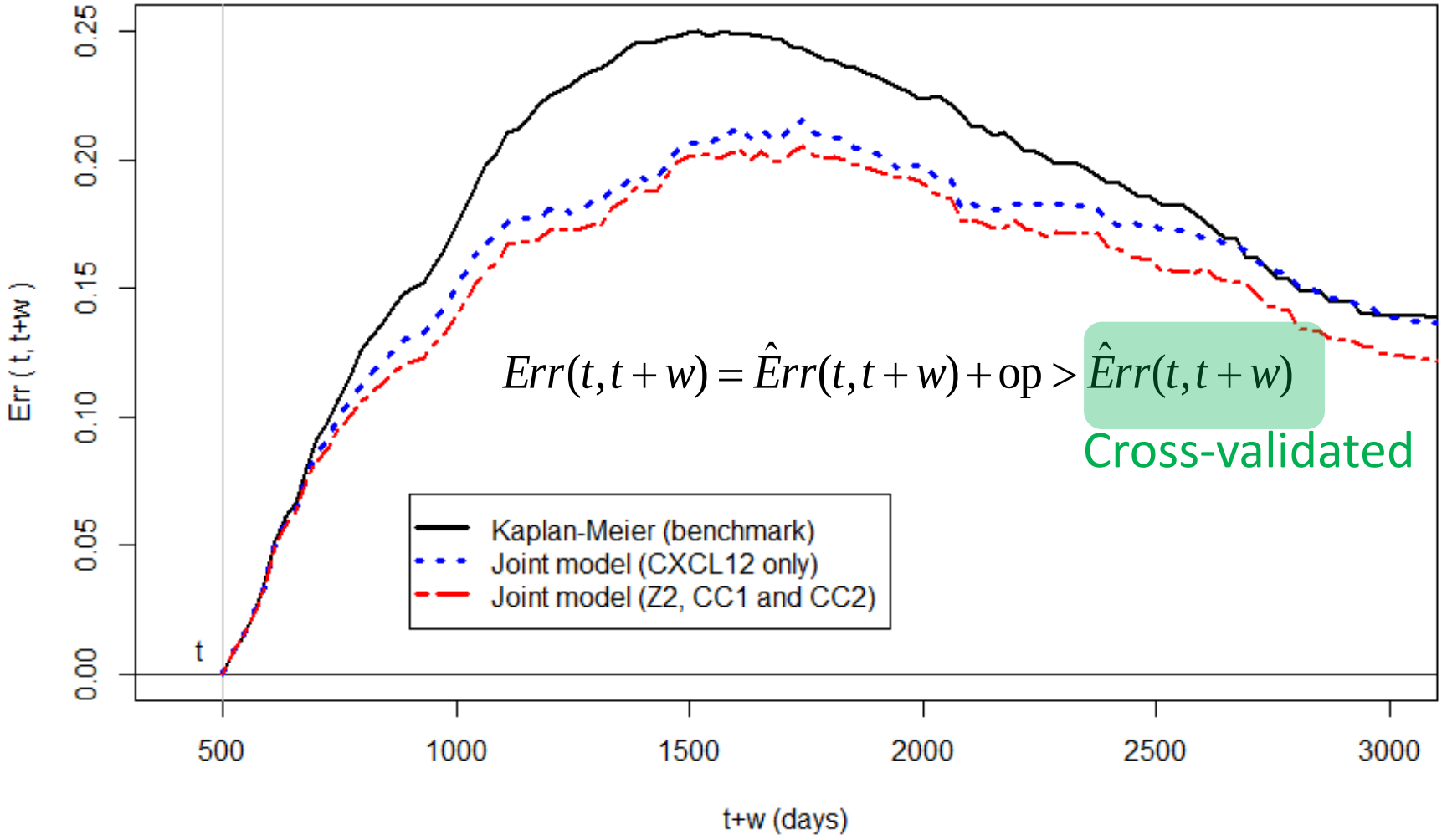
- Joint model with both clinical (Z2) and genetic factors (CC1, CC2) has smallest prediction error

Cross-validated Prediction error at *t*=500 (days)

$$\hat{Err}(t,t+w) =$$

Leave-one-out: Remove one patient

$$\frac{1}{Y(t)}\sum_{ij}\mathbf{I}(T_{ij}^{*}>t)\hat{w}_{ij}(t,t+w)\{\mathbf{I}(T_{ij}^{*}>t+w)-\hat{S}^{-(i,j)}(t,t+w\,|\,H(t,T_{ij}),\mathbf{Z}_{ij})\}^{2}$$



$$Err(t,t+w) = \hat{Err}(t,t+w) + \text{op} > \hat{Err}(t,t+w)$$

Cross-validated

Summary: proposed method

1) Tukey's compound covariate (CC)

followed by univariate selection (P-value<0.001)

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

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

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

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

2) Dynamic prediction formula for a new patient

$$F(t, t+w \mid H(t,x), \mathbf{Z})$$

$$= \Pr(D \le t+w \mid D > t, H(t,x), \mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, CC_1, CC_2))$$

$$H(t,x) = \begin{cases} X > t \\ X = x, \ x \le t \end{cases}$$

Tumour progression

3) Optimism bias was small: Effect of "0.001" cut-off

$$Err(t,t+w) = \hat{E}rr(t,t+w) + \text{op} > \hat{E}rr(t,t+w)$$

Ridge or Lasso-based approach yield ***bigger*** optimism bias (our simulations)