Dynamic prediction involving high-dimensional factors based on the joint frailty-copula model

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

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

Outline

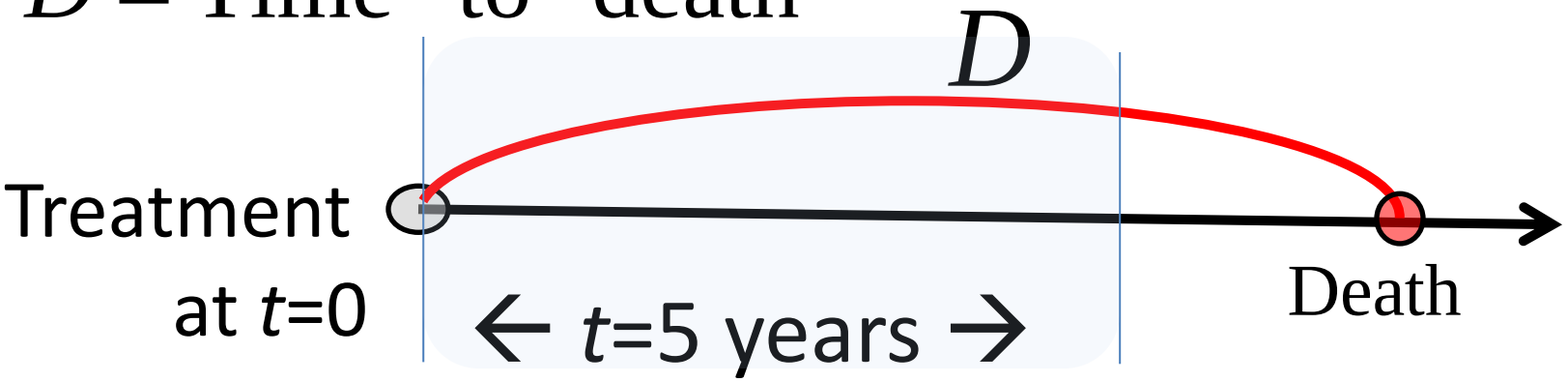
* Survival Prediction
    * Dynamic prediction via copulas
        * High-dimensional genetic factors

Proposed method

*Compound covariate method
    * Prediction formula
        * Prediction error estimation
            *Ovarian cancer data analysis

Classical Survival Prediction

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



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

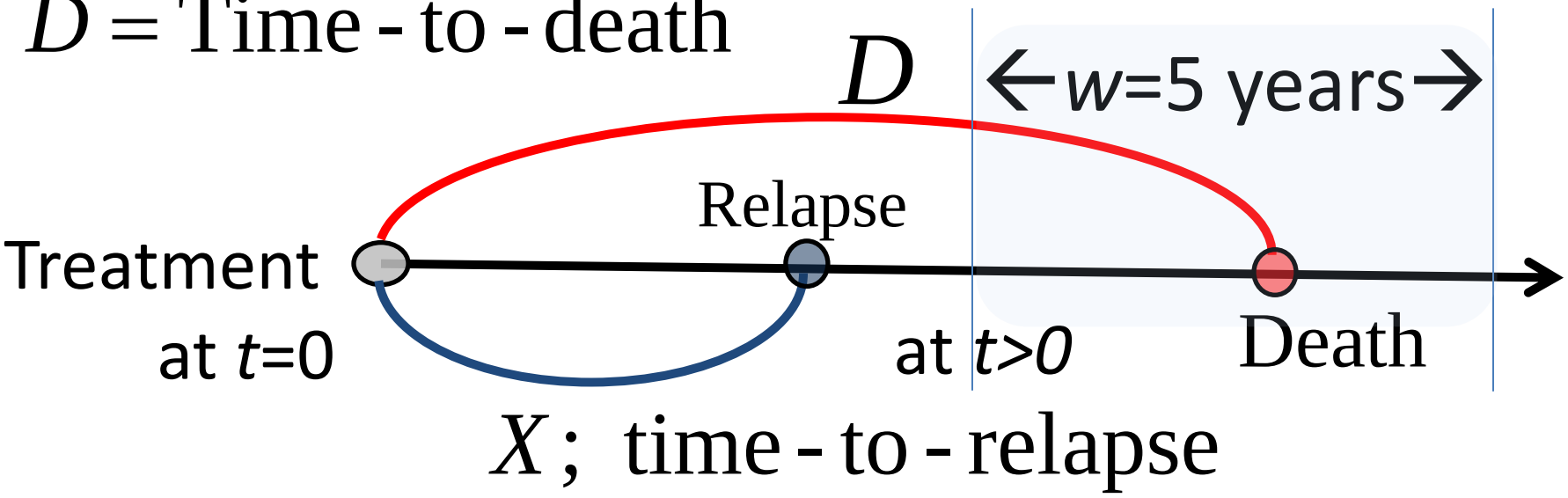
  $\mathbf{Z} =$ (age, sex, stage, tumour size)

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

$$\hat{S}(t \mid \mathbf{Z}) = \exp\{ -\hat{\Lambda}_0(t) e^{\hat{\boldsymbol{\beta}}'\mathbf{Z}} \}$$

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 Cox model (Conditional Cox model at time *t*)
2) Time-dependent covariate ? (Cox model is only for exogenous TDC)
3) Joint model (our approach use a copula on (X, D))

Survival copula model

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

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

(Clayton copula)

$$\theta + 1 = \frac{\Pr(X = x, D = y)\Pr(X > x, D > y)}{\Pr(X = x, D > y)\Pr(X > x, D = y)} = \frac{\text{Concordance}}{\text{Discordance}}$$

$\theta > 0$: Positive dependence

$\theta = 0$: Independence

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

$X=x, D>y$	$X>x, D>y$
$X=x, D=y$	$X>x, D=y$

$$\text{Kendall's tau} = \frac{\theta}{\theta + 2}$$

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)

Methods for high-dimensional factors

- Lasso (Cox-regression with L_1 penalty)

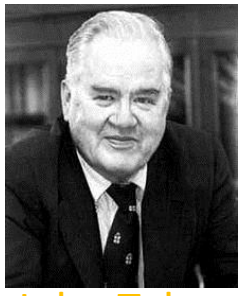
  Tibshirani (1997 Stat Med), Gui & Li (2005 Bioinformatics)

- 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)

Objective

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

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

X : time-to-relapse

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

- Landmark approach (van Houwelingen and Putter 2013)
  - ※ Conditional Cox model: $(D \mid t, X, \mathbf{Z})$
    (multiple models on different *t*'s)
- Ours approach
  - ※ **Joint model**: $(D, X \mid \mathbf{Z})$
    (single model for all *t*'s)

Dynamic prediction via joint models

Method	Response	Souse of Dependence	Meta-analysis	High-dimension
Rizopoulos (2011, Biometrics) Taylor et al. (2013, SMMR) Sène et al. (2014, SMMR) Proust-Lima (2014, SMMR)	Longitudinal measurements + Time-to-events	Frailty	No	No
Mauguen et al. (2013, 2015) Król et al. (2016, Biometrics) Mazroui et al. (2015 LTDA)	Recurrent events + Time-to-death	Frailty	No	No
Our method	Time-to-relapse + Time-to-death	Copula → Subject-level Frailty → Study-level	**Yes** → frailty	**Yes** → CC

- Meta-analysis needs two sources of dependence
  Subject-level dependence + 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

Heterogeneity (frailty)

Dependence (copula)

High-dimensionality (compound covariate)

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 covariates (e.g., age, cancer stage)

$\mathbf{U}_{ij}$ = high-dimensional genetic factors

Semi-competing risks data

* First occuring event time

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

Indicator of tumour progression

* Terminal event time

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

Indicator of death

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

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

Univariate gene selection

Full data:

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

1) Sub-data for time-to-relapse

$$(T_{ij}, \delta_{ij}, U_{ij,k}), \quad i = 1, 2, ..., G, \quad j = 1, 2, ..., N_i$$

Time-to-relapse *k*-th gene

P-value for testing $H_0 : b_k = 0$ via

univariate Cox model: $r_{ij}(t) = r_0(t)\exp(b_k U_{ij,k})$

2) Sub-data for time-to-death

$$(T_{ij}^{*}, \delta_{ij}^{*}, U_{ij,k}), \quad i = 1, 2, ..., G, \quad j = 1, 2, ..., N_i$$

Time-to-death *k*-th gene

P-value for testing $H_0 : c_k = 0$ via

univariate Cox model: $\lambda_{ij}(t) = \lambda_0(t)\exp(c_k U_{ij,k})$

Proposed methods

- **Step 1**: Select genes by **P-value<0.001**

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

$$r_{ij}(t) = r_0(t)\exp(b_k V_{ij,k}), \quad q_1 : \text{the number of gene P-value} < 0.001$$

$$\lambda_{ij}(t) = \lambda_0(t)\exp(c_k W_{ij,k}), \quad q_2 : \text{the number of gene P-value} < 0.001$$

P=0.001 : a criterion in microarray analysis (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 univariate Cox models

Proposed method

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

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

Penalized splines ➔ $r_0(\cdot), \lambda_0(\cdot)$

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, 2017 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) Known 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 $t+w$

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

Proposed prediction formula

- 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*,

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

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

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 $\hat{S} = 1 - \hat{F}$

* $Err(t, t+w) = 0 \quad \Rightarrow$ Perfect prediction

* $Err^{KM}(t, t+w) \quad \Rightarrow$ Null prediction without covariate

where

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

where $\hat{S}^{KM}$ = Kaplan - Meier estimator

Assessing Prediction Error

- Consistent estimation of Brier score

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

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

weight due to IPCW technique:

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

- Variability: evaluate by non-parametric bootstrap:

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

Random sampling with replacement

$$(T_{ij},T_{ij}^{*},\delta_{ij},\delta_{ij}^{*},\mathbf{Z}_{ij}),\quad T_{ij}^{*}>t$$

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

Data analysis (Ganzfried et al., 2013)

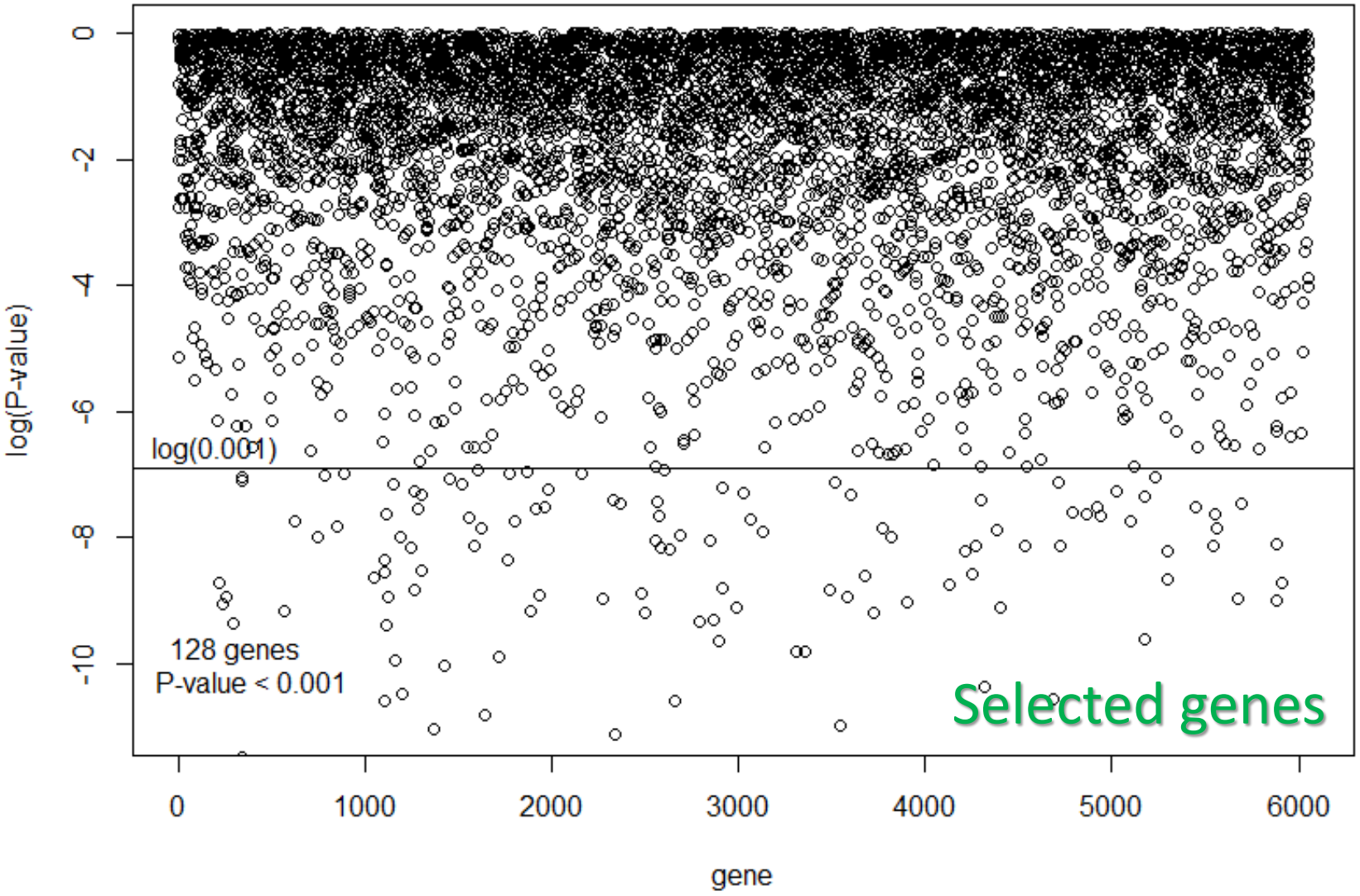
A meta-analytic data combining the four independent studies of ovarian cancer patients

	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

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

Dynamic prediction at t = 500 (early prediction)

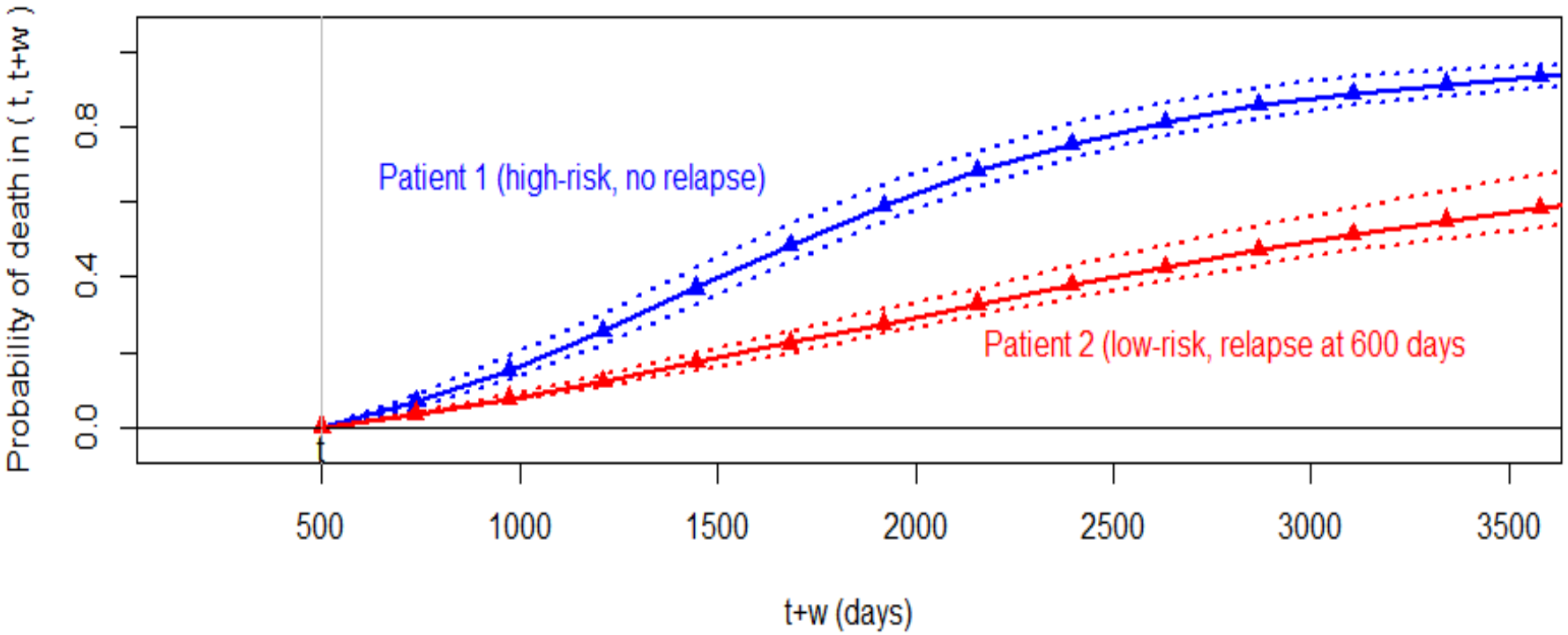
$$F(t, t+w \mid H(t,x), \mathbf{Z}) = \Pr(D \le t+w \mid D > t, H(t,x), \mathbf{Z})$$

Patient 1:

- High-risk factors at $t=0$ ($CC_1=1$, $CC_2=1$, the residual tumour size >=1cm)
- No relapse during the follow-up

Patient 2:

- Low-risk factors at $t=0$ ($CC_1=-1$, $CC_2=-1$, the residual tumour size <1cm)
- Relapse at 600 days after treatment



Dynamic prediction at t = 1000 (late prediction)

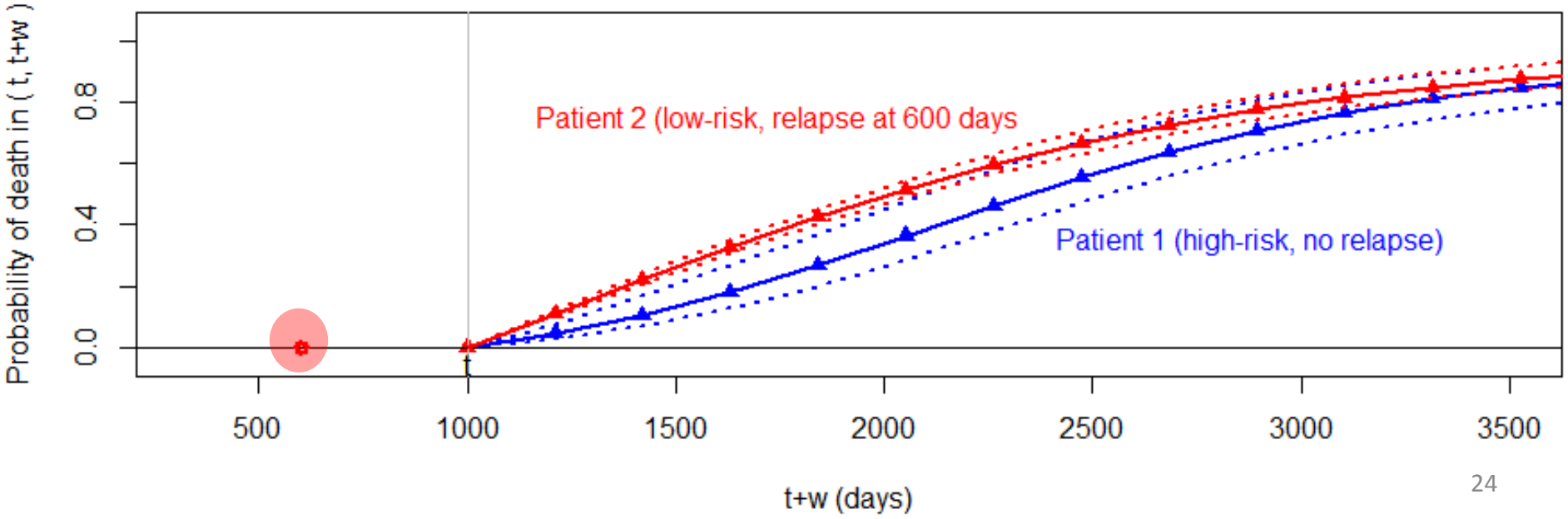
$$F(t, t+w \mid H(t,x), \mathbf{Z}) = \Pr(D \le t+w \mid D > t, H(t,x), \mathbf{Z})$$

Patient 1:

- High-risk factors at $t=0$ ($CC_1=1$, $CC_2=1$, the residual tumour size >=1cm)
- No relapse during the follow-up

Patient 2:

- Low-risk factors at $t=0$ ($CC_1=-1$, $CC_2=-1$, the residual tumour size <1cm)
- Relapse at 600 days after treatment



Prediction error comparison

1. Null model

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

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)

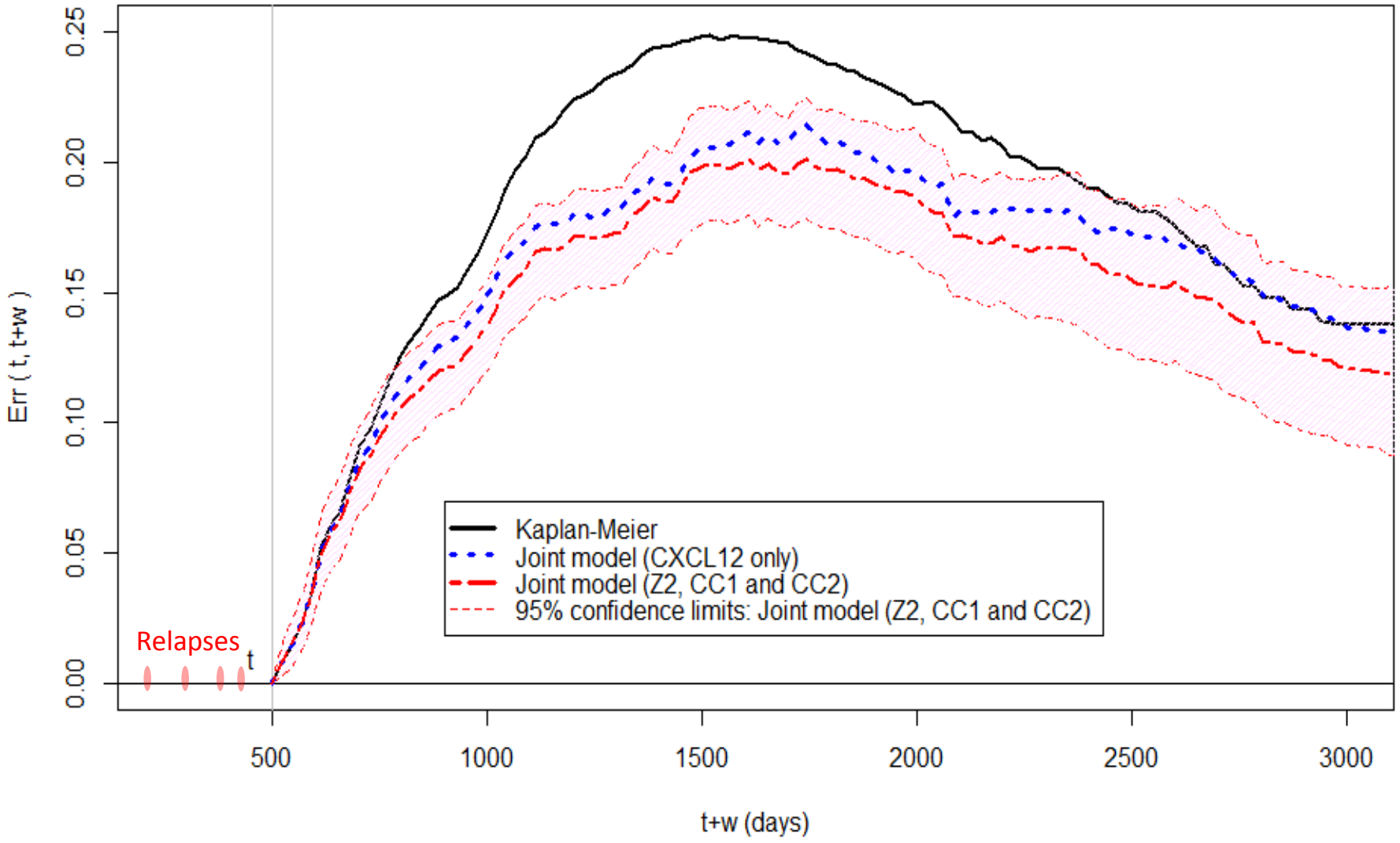
$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\gamma_1 \mathrm{CC}_{1,ij}) & (\text{for time to relapse } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = \lambda_0(t) \exp(\boldsymbol{\beta}_2' \mathbf{Z}_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & (\text{for time to death } D_{ij}) \end{cases}$$

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

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

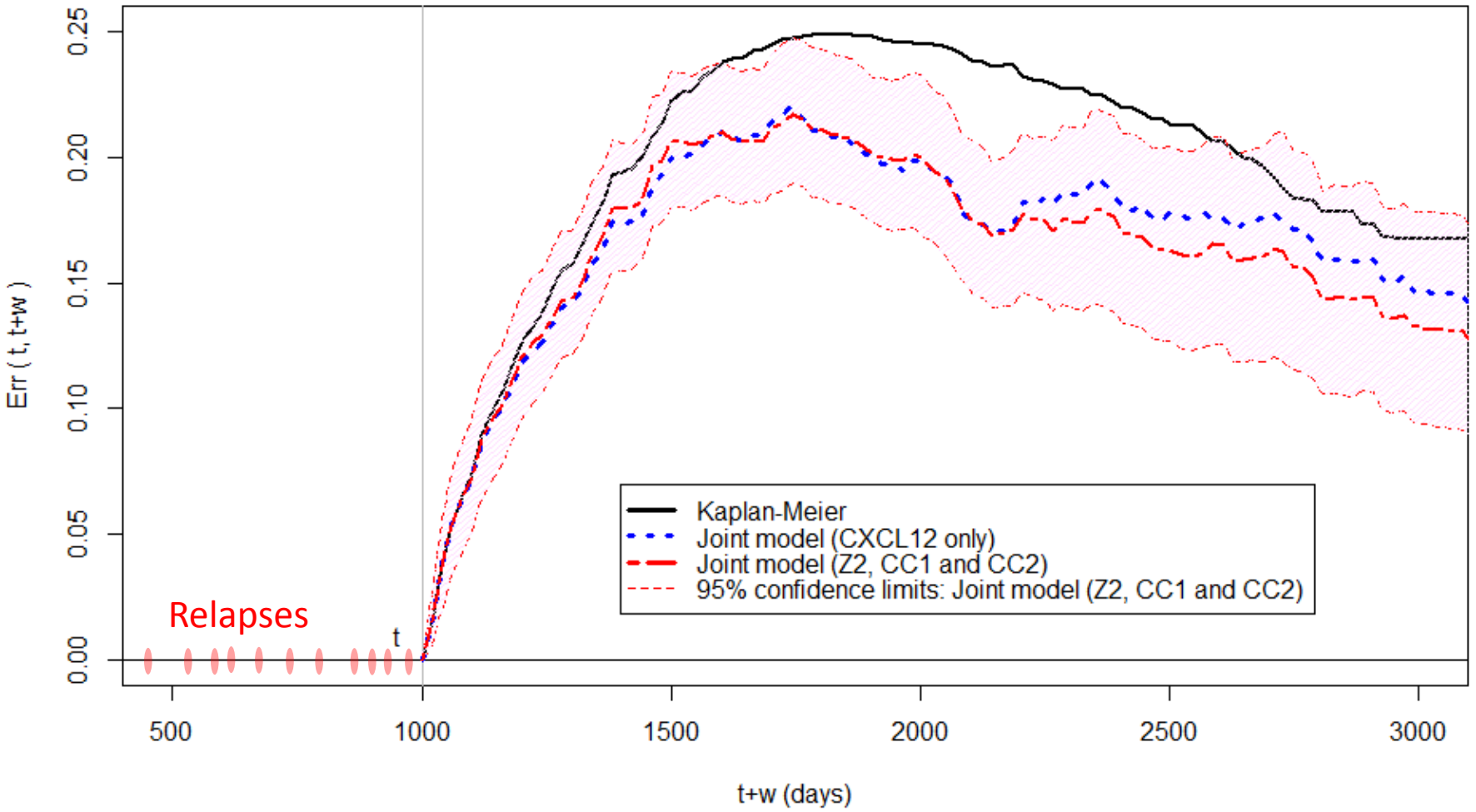
Prediction error at t=500 (early prediction)

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



Prediction error at *t*=1000 (late prediction)

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



- More relapse information is accumulated at 1000 days

Thank you for your attention