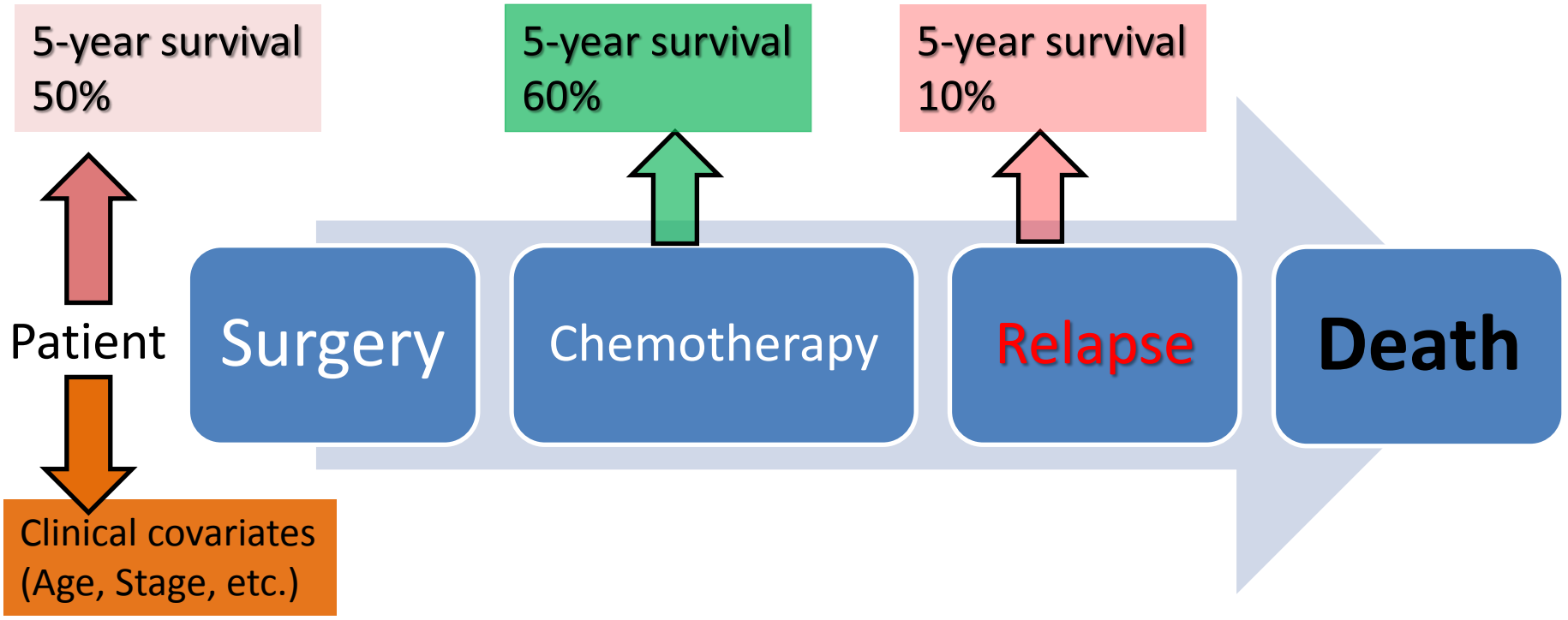
A clinician's guide for dynamic risk prediction of death using an R package *joint.Cox*

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

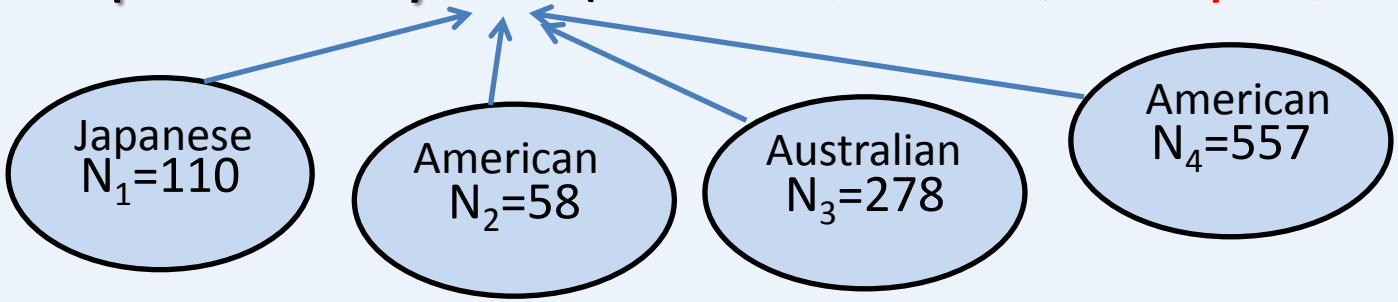
Joint work with
Hirofumi Michimae and Shigeyuki Matsui

2019 WNAR meeting, Portland, Oregon, June 23-26, 2019.

Clinican's prediction for a cancer patient

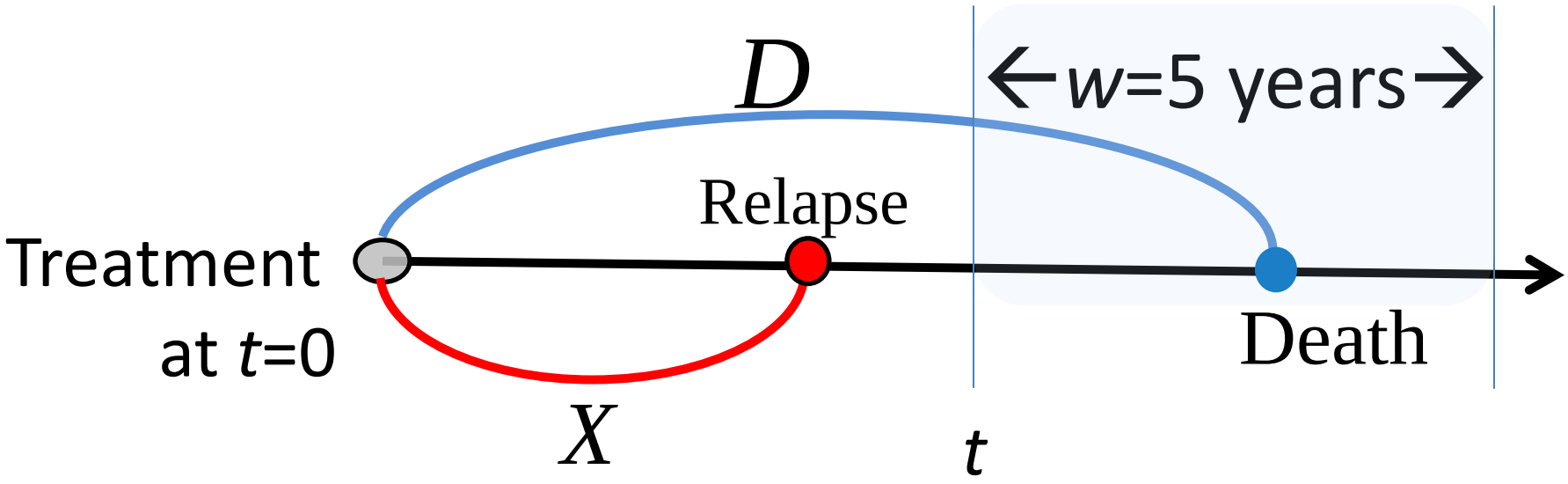


Survival probability = $\hat{F}$ (Clinical, Gene, Relapse, Timing)



Emura T*, Nakatochi M, Matsui S, Michimae H, Rondeau V (2018) Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model, Stat Methods Med Res

Dynamic Prediction

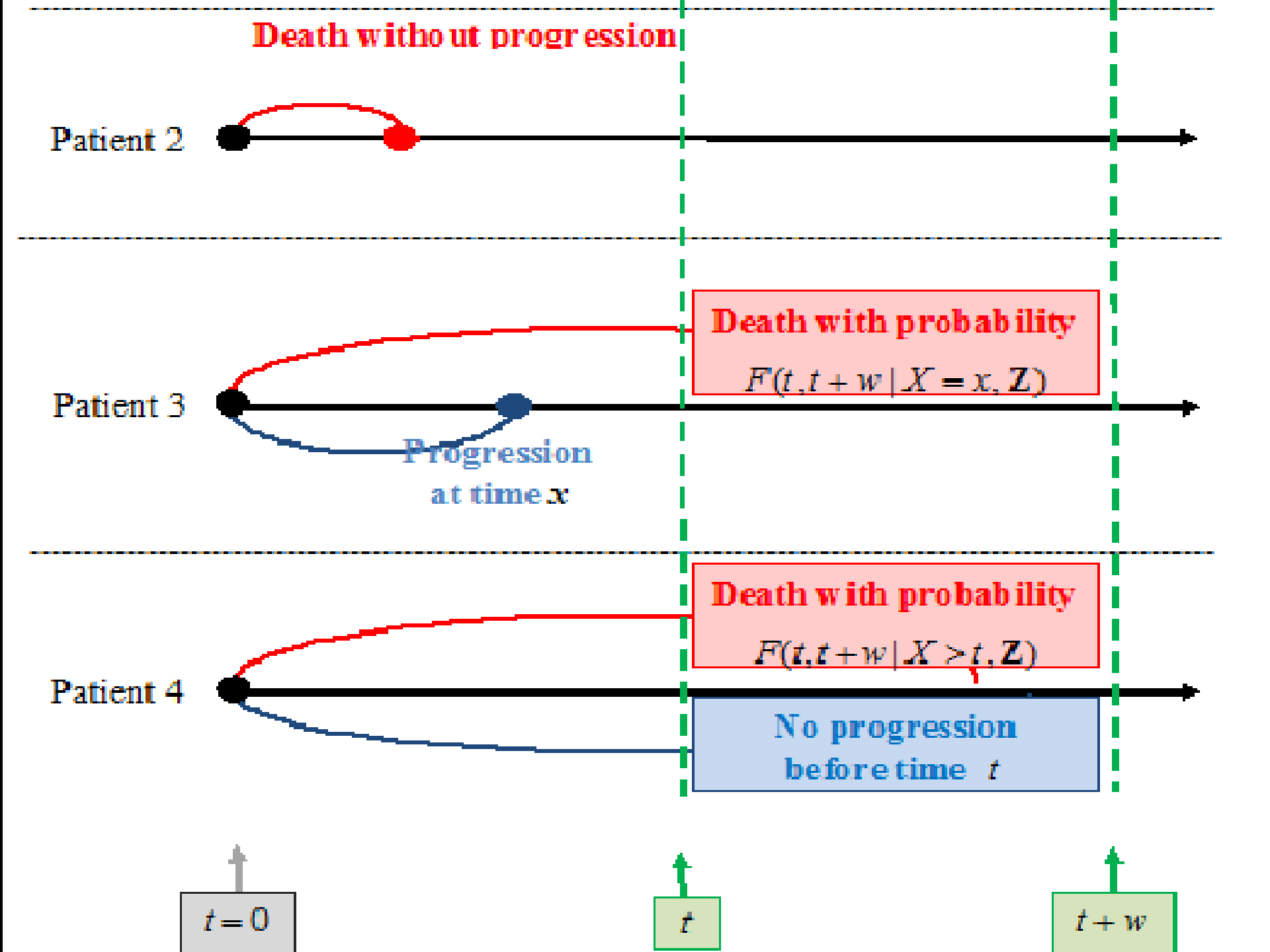


- Conditional failure function (van Houwelingen and Putter 2013)

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

X = time-to-tumour progression (TTP)

D = time-to-death (or OS)


Death without progression
Patient 2
Death with probability
$F(t, t+w \mid X = x, \mathbf{Z})$
Patient 3
Progression
at time x
Death with probability
$F(t, t+w \mid X > t, \mathbf{Z})$
Patient 4
No progression
before time t
$t = 0$
t
$t + w$

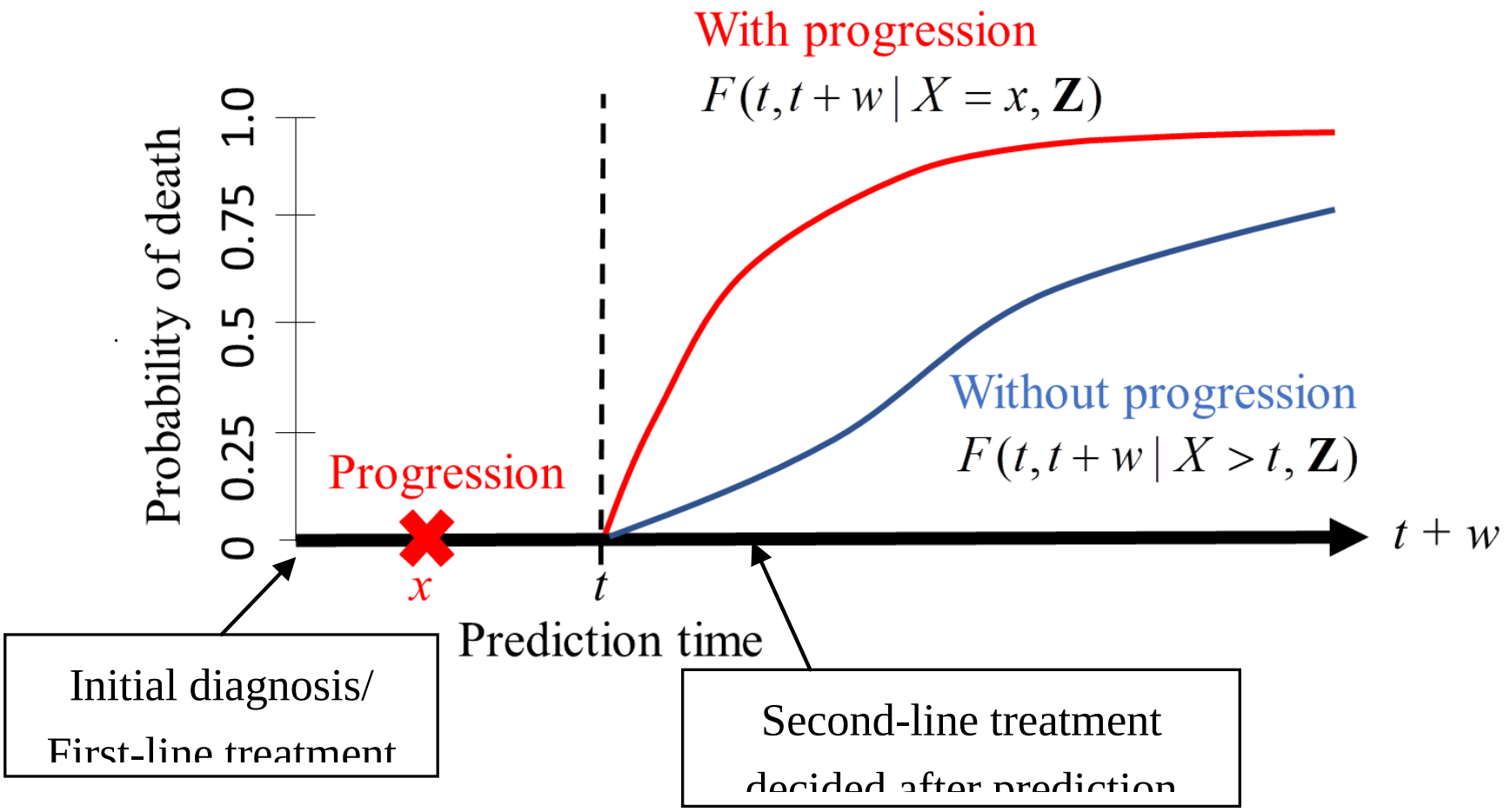


Figure 1. The proposed prediction scheme.

Dynamic prediction via joint models

Method	Response	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
Rondeau et al (2017, SMMR)	Clustered failure events	Frailty	No
Our method	TTP + Time-to-death	**Copula**	Yes

- Joint frailty-copula model (for meta-analysis)

$$\begin{cases} r(t \mid u) = u\ r_0(t)\exp(\boldsymbol{\beta}_1'\mathbf{Z}_1) & \text{for TTP} \\ \lambda(t \mid u) = u^{\alpha}\lambda_0(t)\exp(\boldsymbol{\beta}_2'\mathbf{Z}_2) & \text{for OS} \end{cases}$$

Clinical + Genetic covariates

High-dimensional Gene expressions:

Breast cancer:

van't Veer et al. (2002); van de Vijver et al. (2002) → *MammaPrint* (**70 genes**)
Sotiriou et al. (2006); Haibe-Kains et al. (2006) → *GGI* (**93 genes**)

Ovarian cancer:

Yoshihara et al. (2010) Yoshihara et al. (2012) → *Ridge PI* (**88 genes**, or **126 genes**)
Emura et al. (2018) → *Compound covariate* (**128 genes**)

Lymphoma:

Rosenwald et al. (2002) → *Outcome-predictor score* (**17 genes**)
Matsui S (2006) → *Compound covariate* (**75 genes** or **85 genes**)

Clayton copula model

$$\Pr(X > x, D > y \mid u) = [\, S_X(x \mid u)^{-\theta} + S_D(y \mid u)^{-\theta} - 1 \,]^{-1/\theta}$$

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

$\theta > 0$: Positive dependence

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

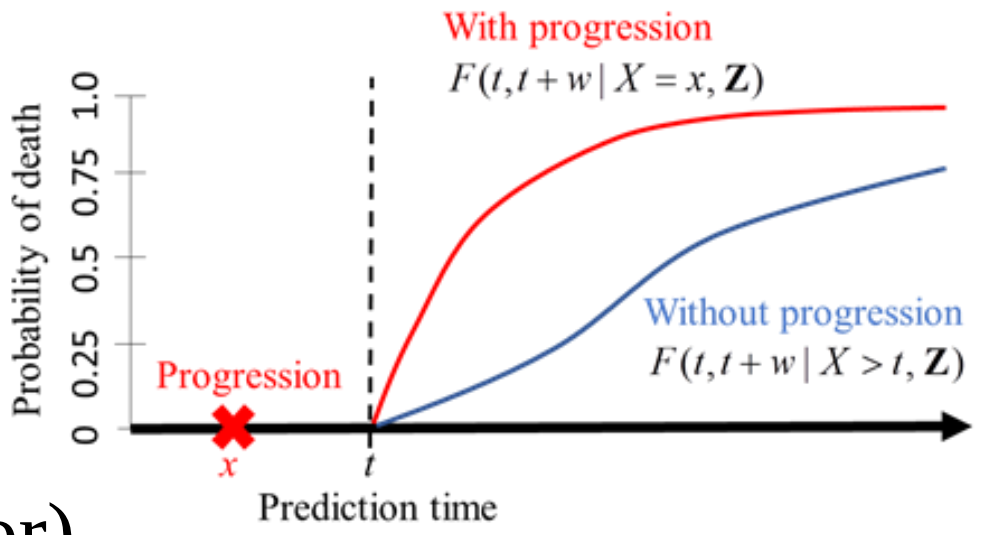
- Kendall's tau $= \dfrac{\theta}{\theta + 2}$

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

Schematic algorithm

1. Set parameters: $\boldsymbol{\beta}$. , $(r_0(\cdot), \lambda_0(\cdot))$, η, α, and θ
2. Set patient information 
3. Set prediction time
4. Draw the plot
5. Validate the results

   (assess prediction error)



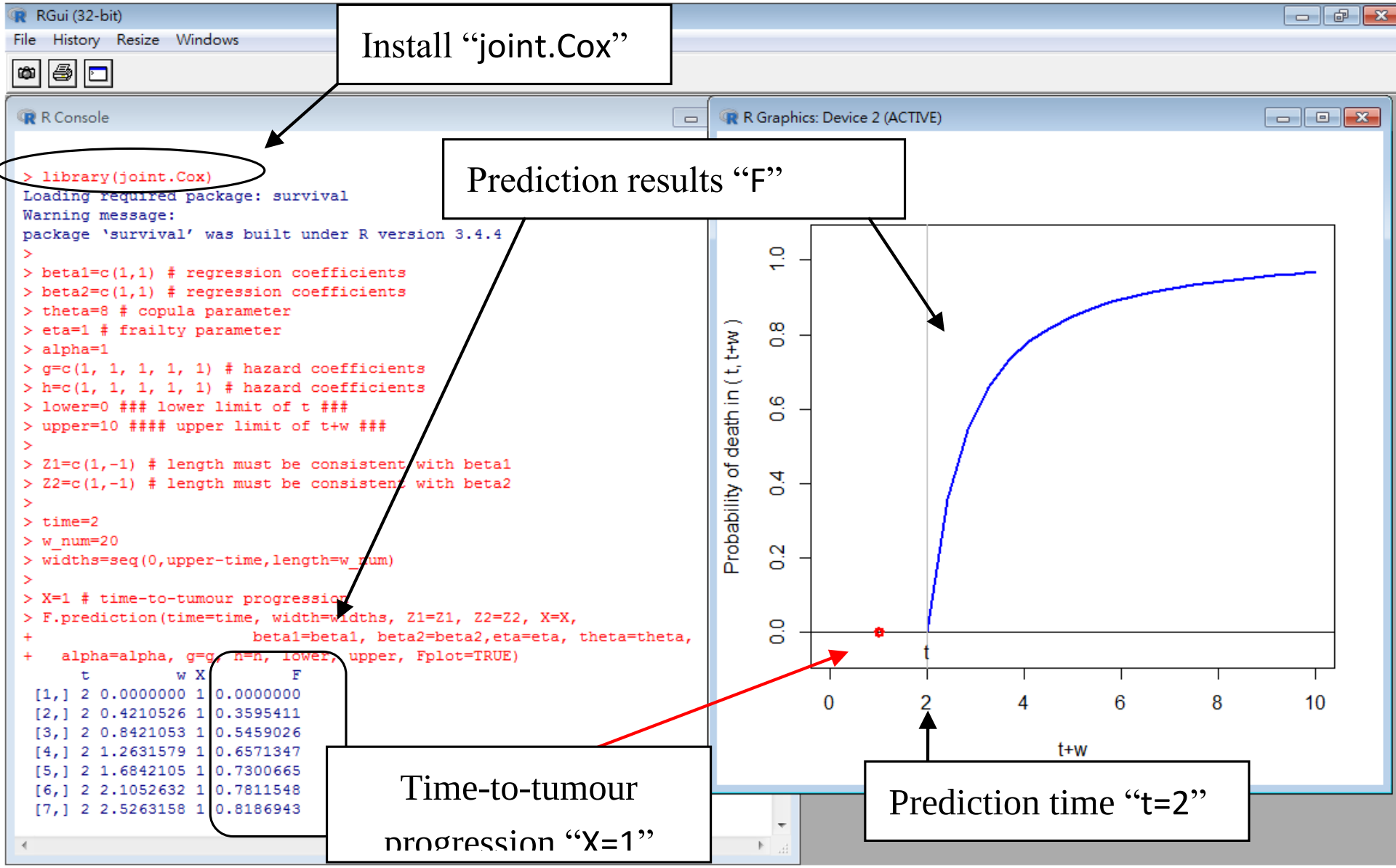


Figure 2: The screenshot of the R console after running the codes.

Example: Breast cancer data (Haibe-Kains et al. 2006)

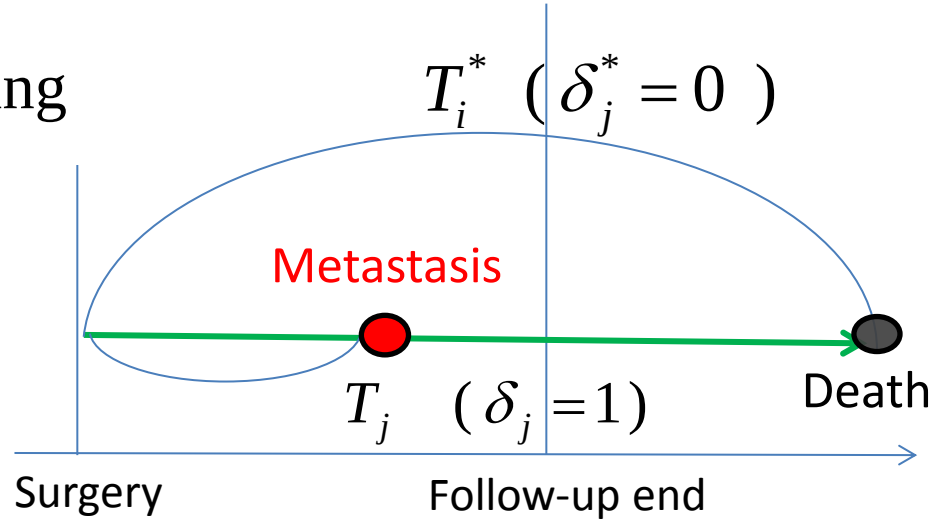
T_i : time-to-metastasis or censoring

δ_i : metastatic status (0 or 1)

T_i^* : time- to- death or censoring

δ_i^* : vital status (0 or 1)

$\mathbf{Z}_i$: Covariates



⇓

- Estrogen receptor status (**ER**=1 for positive; =0 for negative)
- Tumor size (**Size**=1 for > 2 cm; =0 for ≤ 2 cm)
- Lymph nodal status (**Node**=1 for present; =0 for absent)
- Age at diagnosis (**Age**=1 for age ≤40; =2 for 40<age≤50; =3 for age>50)
- The 70-gene signature developed by van't Veer et al. [1, 2]
  (**MammaPrint**=1 for high; =-1 for low)
- The gene expression grade index (GGI) developed by Sotiriou et al. [3]
  (**GGI**=1 for high; =-1 for low)

Breast cancer data (Haibe-Kains et al. 2006)

Dataset[a]	Maximum (median) follow-up days	N	No. of events (event rates)		
			Metastasis	Death	Censoring
CAL	5,165 (4,219)	109	24 (22%)	75 (69%)	34 (31%)
NIK	6,694 (3,232)	295	101 (34%)	79 (27%)	216 (73%)
TRANSBIG	9,108 (5,101)	196	62 (32%)	56 (29%)	140 (71%)
UCSF	8,267 (2,799)	120	19 (16%)	39 (32%)	81 (68%)
Total	**9,108 (3,769)**	**720**	**206 (29%)**	**249 (35%)**	**471 (65%)**

CAL = U of California, San Francisco, California Pacific Medical Center (United States)
NKI = National Kanker Instituut (the Netherlands)
TRANSBIG = dataset collected by the TransBIG consortium (Europe)
UCSF = U of California, San Francisco (United States).

Fit the joint frailty-copula model (Emura et al. 2017)

$$
\begin{cases}
r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\boldsymbol{\beta}_1' \mathbf{Z}_{1,ij}) & \Leftarrow \text{hazard for metastatis} \\
\lambda_{ij}(t \mid u_i) = u_i^{\alpha} \lambda_0(t) \exp(\boldsymbol{\beta}_2' \mathbf{Z}_{2,ij}) & \Leftarrow \text{hazard for death} \\
\Pr(X_{ij} > x, D_{ij} > y \mid u_i) = [S_X(x \mid u)^{-\theta} + S_D(y \mid u)^{-\theta} - 1]^{-1/\theta} & \Leftarrow \text{Clayton copula}
\end{cases}
$$

↓Maximum Penalized Likelihood Estimator (R package *joint.Cox*)

$$\hat{\boldsymbol{\beta}}_1' \mathbf{Z}_1 = (-0.15 \times \text{Age}) + (-0.23 \times \text{ER}) + (0.27 \times \text{Size}) + (0.20 \times \text{MammaPrint}) + (0.19 \times \text{GGI})$$

$$\hat{\boldsymbol{\beta}}_2' \mathbf{Z}_2 = (-0.36 \times \text{ER}) + (0.14 \times \text{Node}) + (0.27 \times \text{Size}) + (0.17 \times \text{MammaPrint}) + (0.25 \times \text{GGI})$$

$$\hat{r}_0(t) = 0.20 \times M_1(t) + 0.39 \times M_2(t) + 0.19 \times M_3(t) + 0.43 \times M_4(t) + 0.25 \times M_5(t)$$

$$\hat{\lambda}_0(t) = 0.05 \times M_1(t) + 0.37 \times M_2(t) + 0.38 \times M_3(t) + 0.09 \times M_4(t) + 0.00 \times M_5(t)$$

$$\hat{\theta} = 10.7 \text{ (95\%CI: 8.6-13.4)}$$

```
beta1=c(-0.15, -0.23, 0.27, 0.20, 0.19)
beta2=c(-0.36, 0.14, 0.27, 0.17, 0.25)
g=c(0.20, 0.39, 0.19, 0.43, 0.25) # baseline hazard coefficients
h=c(0.05, 0.37, 0.38, 0.09, 0.00) # baseline hazard coefficients
theta=10.7
eta=0.067
```

Set parameters

Step 2: Set patient information

Patient 1:

Age at diagnosis = 45; Estrogen receptor = positive; Tumor size > 2cm;

Lymph nodal status = present; MammaPrint = High; GGI= High

The patient-level information for covariates are set as

$$\mathbf{Z}_1 = (\text{Age, ER, Size, MammaPrint, GGI}),$$

$$\mathbf{Z}_2 = (\text{ER, Node, Size, MammaPrint, GGI}).$$

Hence, we set the following values for our proposed algorithm:

```
Z1=c("age"=2,"er"=1,"size"=1,"MAMMAPRINT"=1,"GGI"=1)
Z2=c("er"=1,"node"=1,"size"=1,"MAMMAPRINT"=1,"GGI"=1)
```

Step 3: Set prediction time

Patient 1:

Age at diagnosis = 45; Estrogen receptor = positive; Tumor size > 2cm;

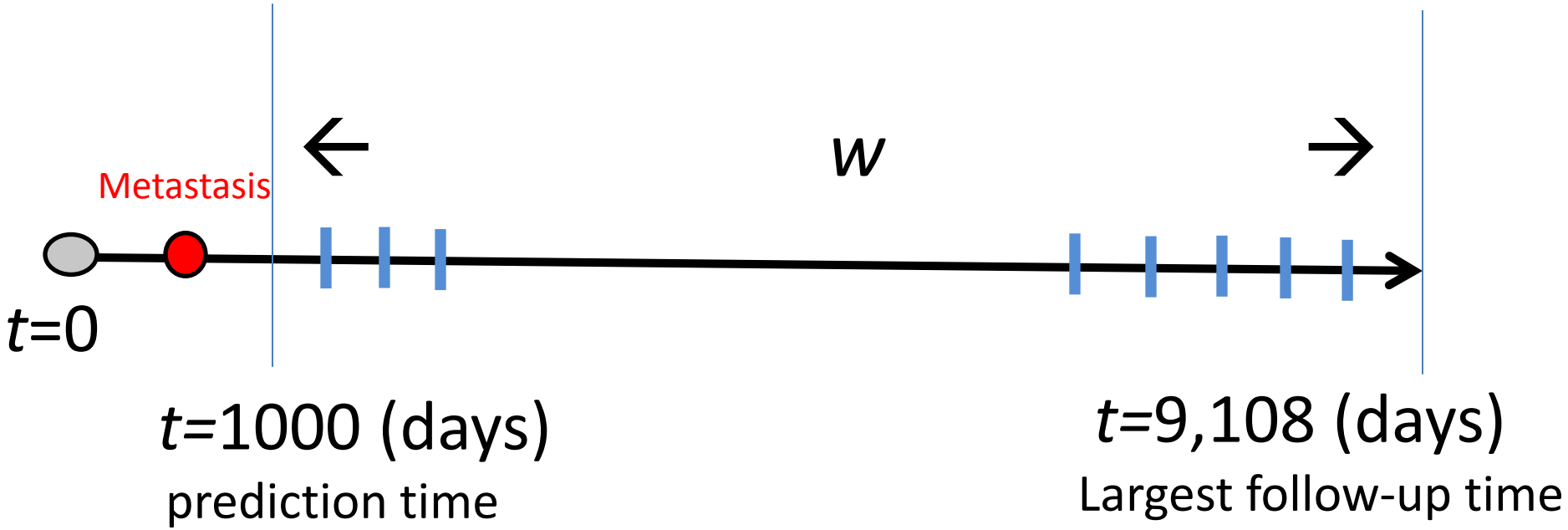
Lymph nodal status = present; MammaPrint = High; GGI= High

5 patients with the same status as Patient1 (in the dataset)

- 2 patients developed metastasis <1000 days
- 3 patients developed metastasis >1000 days

➔ Set our prediction time at t=1000 days

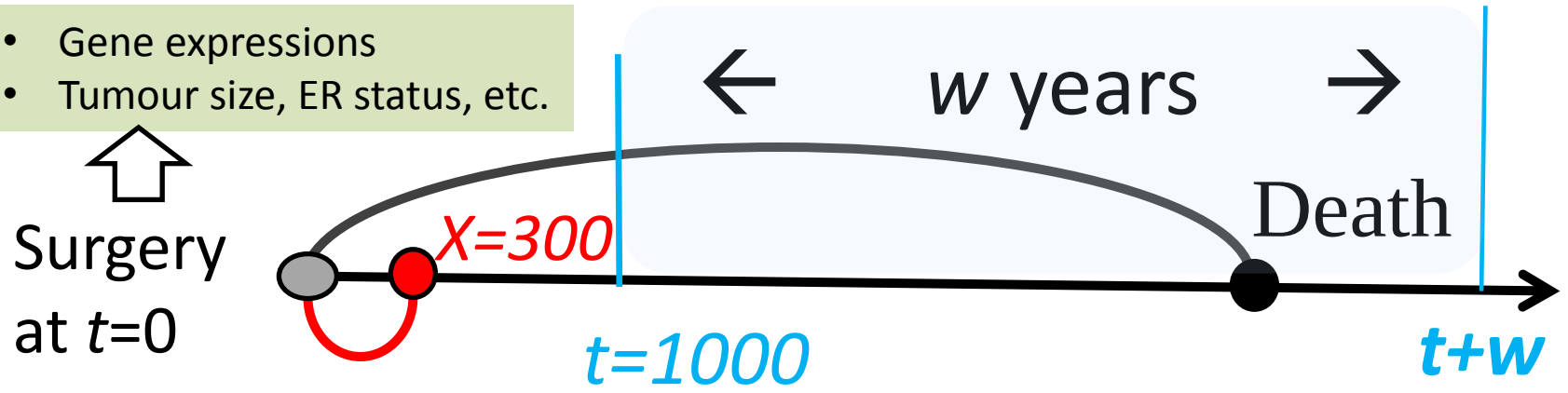
Step 3: Set prediction horizens




```
time=1000
w_num=20
widths=seq(0,upper-time,length=w_num)
> round(widths,0)
[1]     0   427   853  1280  1707  2134  2560  2987  3414  3841  4267
[12]  4694  5121  5548  5974  6401  6828  7255  7681  8108
```

Step 4: Draw the plot

- Gene expressions
- Tumour size, ER status, etc.



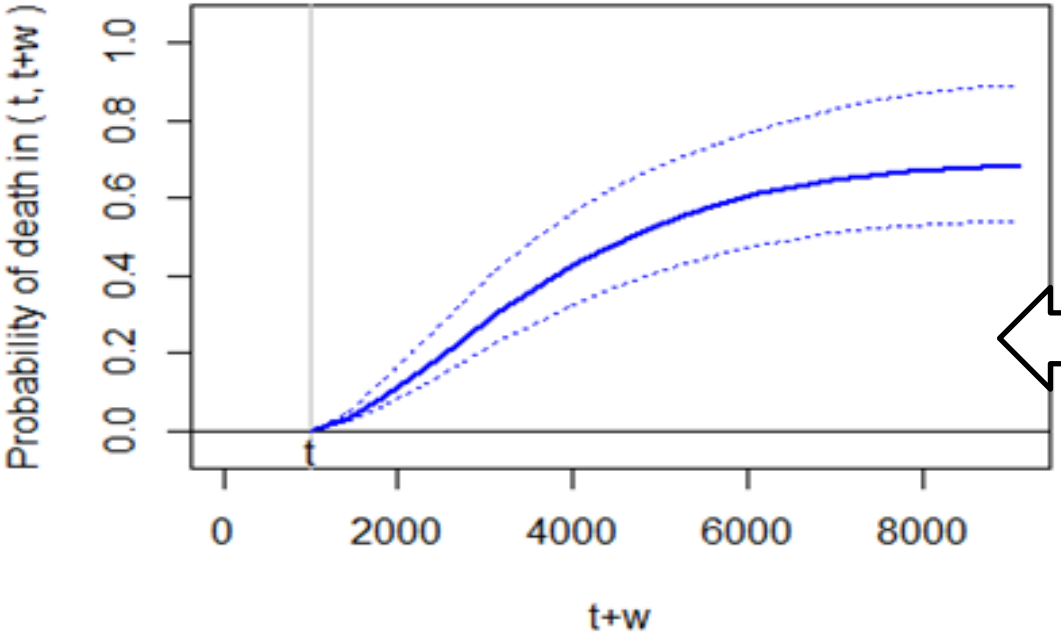
Probability of death (t, t+w)

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

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

↓ compute by an R package *joint.Cox*

```
F.prediction(time=time,  width=widths,  Z1=Z1,  Z2=Z2,  X=300,
                    beta1=beta1,  beta2=beta2,eta=eta,  theta=theta,
                    alpha=alpha,  g=g,  h=h,  lower,  upper,  Fplot=TRUE)
```

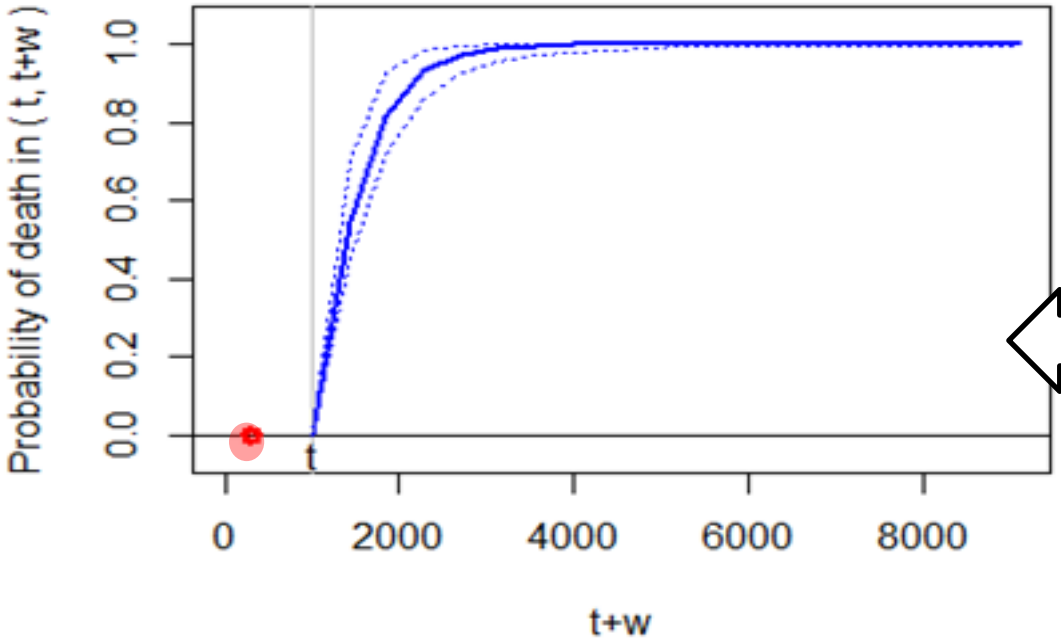


No metastasis
at t=1000 days

$$RR = (\hat{\theta} + 1)$$

$$= 11.7 \text{ (95\%CI: 9.6-14.4)}$$

11 times higher risk



Metastasis
at *x*=300 days

Step 5: Validate the results

Three criteria to be met:

(i) The 95%CI not too wide

(ii) The prediction error sufficiently small

(iii) The model not over-fitting

Brier score (prediction error)

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

where $\hat{S}(t,t+w|\cdot) = 1 - \hat{F}(t,t+w|\cdot)$

Ref: Gerds and Schumacher (2006, Biometrical J)

Estimation of Brier score

- Under the joint model:

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

↑ Compute a bootstrap 95%CI

Re-sample from the risk set of size $Y(t)=\sum_{ij}\mathbf{I}(T_{ij}^{*}>t)$

- Under the null model:

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

Validation criterion:

[95%CI of $Err(t,t+w)$] << $\hat{Err}^{KM}(t,t+w)$

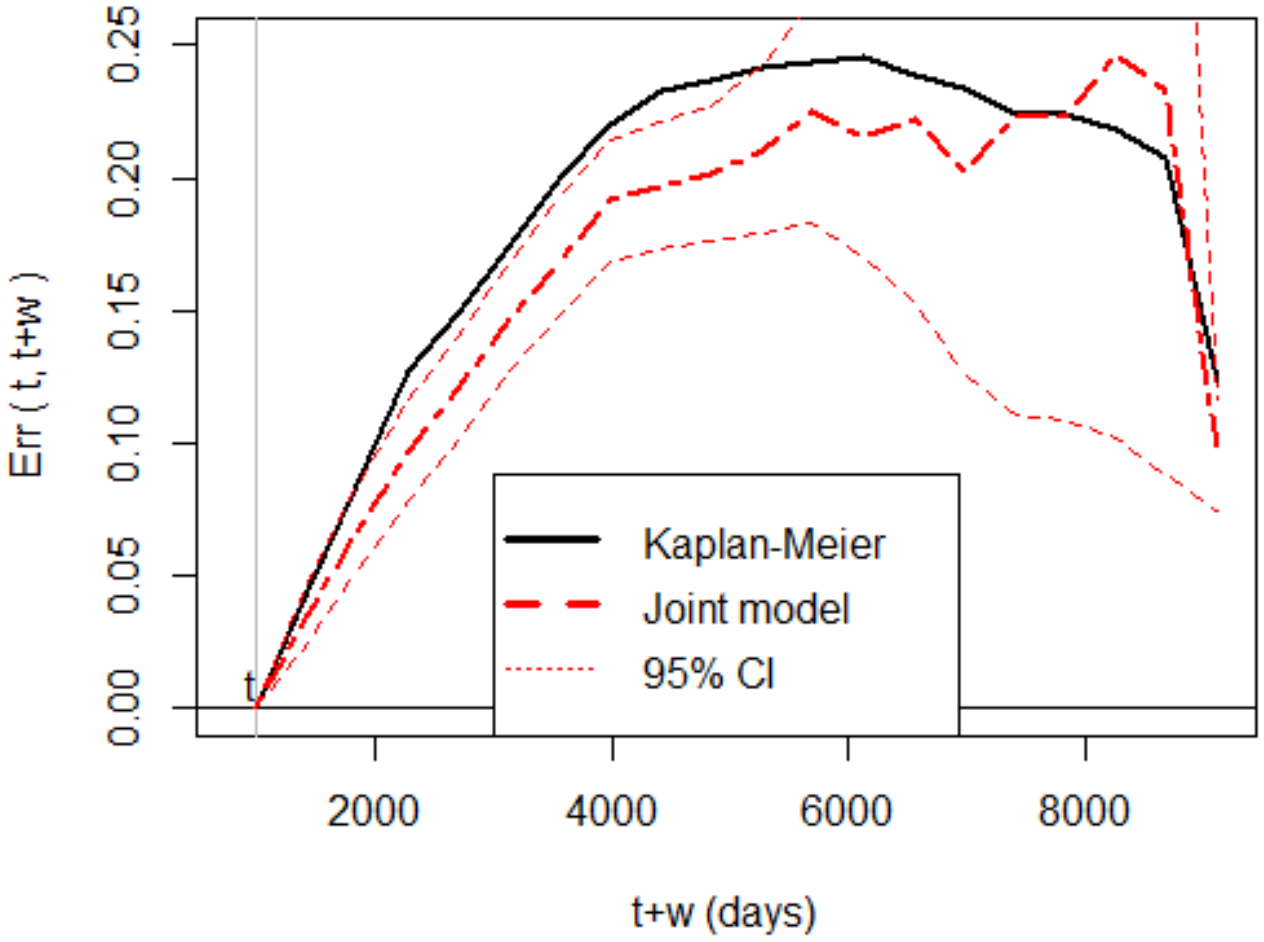


Figure 5: Estimated prediction errors (Brier scores) using the breast cancer data. The prediction tims is set at $t = 1000$ days.

The joint model over-fitting?

Estimator of Brier score

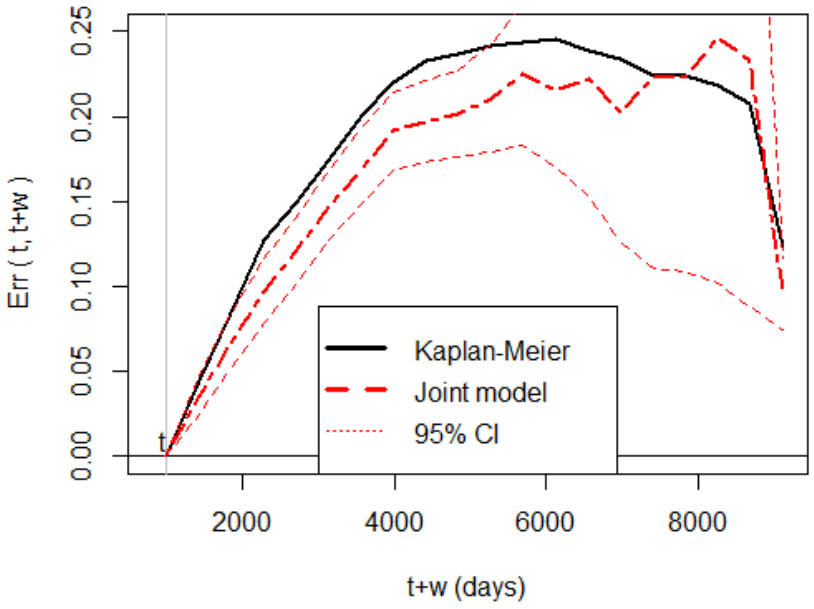
$$\hat{Err}(t,t+w)=\frac{1}{Y(t)}\sum_{ij}\mathbf{I}(T_{ij}^{*}>t)\hat{w}_{ij}(t,t+w)\{\mathbf{I}(T_{ij}^{*}>t+w)-\hat{S}(t,t+w|H(t,T_{ij}),\mathbf{Z}_{ij})\}^{2}$$

Cross-validated estimator

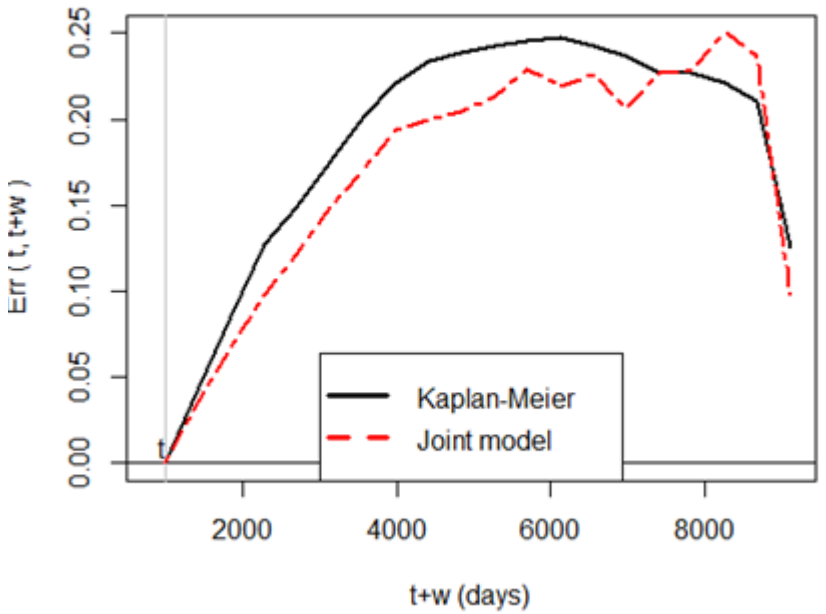
Leave-one-out estimator

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

Not cross-validated

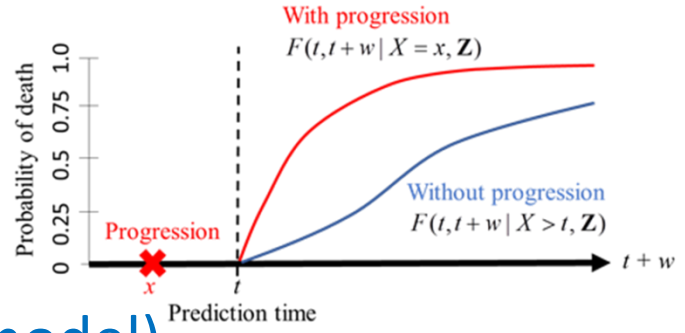


Cross-validated



Summary & Discussions

- A guide for clinicians to apply *joint.Cox*
  - How to fit a joint model
  - How to set prediction time
  - How to draw the plot
  - How to validate the results.



- Bivariate joint survival model (copula model)
  - Intermediate event (TTP) and overall survival (OS)
- TTP is outcome, not covariate
  - TTP can be a primary endpoint (at time $t = 0$)
  - But TTP can be a predictor (at time $t > 0$)
- Optimism bias of prediction error
  - Mainly come from high-dimensional gene expressions ($p>>n$)

  (Sol 1) Use existing scores such as *MammaPrint* (70 genes) and *GGI* (93 genes)

  (Sol 2) Use compound covariate (univariate feature selection)
  Little bias even if selection & predictor development is performed within each cross-validation fold (Emura et al. 2018)

  (Sol 3) Use R packages for feature selection & predictor development:
  SGL (Simon et al. 2013), *penalized* (Goeman et al. 2017), *SIS*, *compound.Cox*, etc..