SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(9) 2842–2858

DOI: 10.1177/0962280216688032
journals.sagepub.com/home/smm
SAGE

# Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: Meta-analysis with a joint model

**Takeshi Emura,[1] Masahiro Nakatochi,[2] Shigeyuki Matsui,[3] Hirofumi Michimae[4] and Virginie Rondeau[5]**

**Abstract**

Developing a personalized risk prediction model of death is fundamental for improving patient care and touches on the realm of personalized medicine. The increasing availability of genomic information and large-scale meta-analytic data sets for clinicians has motivated the extension of traditional survival prediction based on the Cox proportional hazards model. The aim of our paper is to develop a personalized risk prediction formula for death according to genetic factors and dynamic tumour progression status based on meta-analytic data. To this end, we extend the existing joint frailty-copula model to a model allowing for high-dimensional genetic factors. In addition, we propose a dynamic prediction formula to predict death given tumour progression events possibly occurring after treatment or surgery. For clinical use, we implement the computation software of the prediction formula in the *joint.Cox* R package. We also develop a tool to validate the performance of the prediction formula by assessing the prediction error. We illustrate the method with the meta-analysis of individual patient data on ovarian cancer patients.

**Keywords**

Compound covariate, copula, dependent censoring, risk prediction, semi-competing risk, surrogate endpoint

## 1 Introduction

In cancer studies, predicting risk of death is fundamental for improving patient care and optimizing treatment strategies. A common approach in survival analysis is to predict a dichotomous event status, death, or alive, within a given time window (e.g., five-years after treatment). The traditional prediction scheme is on the basis of conditional survival probability given clinical covariates collected at time $t = 0$, the time of treatment.[1,2] The conditional survival probability is easily estimated by fitting the Cox proportional hazards model and applying the Breslow estimator.[3]

Recent years have witnessed a rapid increase in the use of genomic factors to refine survival prediction models in medical research. Models incorporating genomic factors often lead to an improved prediction accuracy compared to models based solely only on traditional clinical covariates, as reported in breast cancer,[4,5] diffuse large-B-cell lymphoma,[6,7] lung cancer,[8–10] ovarian cancer,[11–13] and other cancers. In both medical and statistical contexts, evaluating predictive accuracy of survival models has been an active research area due to the high-dimensional nature of data.[14–21]

[1]Graduate Institute of Statistics, National Central University, Taoyuan City, Taiwan
[2]Statistical Analysis Section, Center for Advanced Medicine and Clinical Research, Nagoya University Hospital, Nagoya, Japan
[3]Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Japan
[4]Department of Clinical Medicine (Biostatistics), School of Pharmacy, Kitasato University, Tokyo, Japan
[5]INSERM CR1219 (Biostatistic), Université de Bordeaux, Bordeaux Cedex, France

**Corresponding author:**
Takeshi Emura, Graduate Institute of Statistics, National Central University, Zhongda Rd., Zhongli District, Taoyuan City 32001, Taiwan.
Email: takeshiemura@gmail.com

Survival prediction using genomic information can be further improved by a 'dynamic prediction' scheme that develops a risk prediction formula at a certain moment $t > 0$. To predict patient survival, dynamic prediction utilizes the record of intermediate events occurring to a patient. For example, tumour progression of a patient (e.g., relapse of cancer) may be strongly predictive of the patient's overall survival. However, such dynamic events are not available at time $t = 0$, as they evolve with time.

Different types of dynamic prediction have been proposed in the literature. The dynamic prediction scheme was initially developed under the landmark Cox model.[22,23] In recent years, there has been a noticeable trend using a joint model that accounts for the dependence between survival and other responses via frailty. Different frailty models have been developed to join different response types: survival and recurrent events,[24,25] survival and longitudinal covariates,[26–29] clustered multivariate survival responses,[30] survival, longitudinal covariates, and recurrent events.[31] However, these existing joint frailty models for dynamic prediction have not been adapted to handle high-dimensional genomic factors. In addition, dynamic prediction under the setting of individual patient data (IPD) meta-analysis has rarely been discussed in the literature.

In this context, this paper seeks to develop a personalized survival prediction formula by incorporating both genomic factors and dynamic history of tumour progression events under the meta-analytic setting. As for the statistical methods, we follow the meta-analytic approach with the joint frailty-copula model.[32] This copula-based approach results in a novel prediction formula compared to existing prediction schemes that focus solely on random effect joint models.[24–31]

The paper is organized as follows. Section 2 extends the joint frailty model to take into account for the high-dimensional genomic factors, where we utilize the compound covariate (CC) as a main technique for dimension reduction. Section 3 derives the proposed dynamic prediction formula. Section 4 considers a tool to validate the prediction formula by estimating prediction error. Section 5 illustrates the methods through the meta-analysis of ovarian cancer data. Section 6 concludes.

## 2 Joint model with meta-analysis

### 2.1 Motivating example: Meta-analysis of ovarian cancer patients

Association between the *CXCL12* gene expression and survival was reported by Popple et al.[11] in ovarian cancer patients. Their finding was confirmed by Ganzfried et al.[33] based on the meta-analysis of 14 independent studies. The multivariate Cox regression analyses performed by Popple et al.[11] and Ganzfried et al.[33] indicated that the *CXCL12* gene expression is predictive of patient survival, independently from other clinic-pathological covariates. A meta-analysis using a joint model further confirmed that the expression of *CXCL12* gene is predictive of both cancer relapse and death.[32] These studies focused solely on the *CXCL12* gene expression ignoring other gene expressions.

The extracted data from Ganzfried et al.[33] consist of 912 individual ovarian cancer patients (544 relapsed, 465 died, and 447 censored) from four independent studies (Table 1): the data extraction shall be detailed in Section 5. There are 11,756 gene expressions that are commonly available across the four studies, including the *CXCL12* gene expression.

**Table 1.** A meta-analytic data combining the four independent studies of ovarian cancer patients of Ganzfried et al.[33]

| Data set[a] | Median follow-up (days) | Sample size | The number of observed events (event rates) | | | The number of genes |
|---|---|---|---|---|---|---|
| | | | Relapse ($\delta_{ij} = 1$) | Death ($\delta^*_{ij} = 1$) | Censoring ($\delta^*_{ij} = 0$) | |
| GSE17260 | 1410 | $N_1 = 84$ | 59 (70%) | 38 (45%) | 46 (55%) | 18,548 |
| GSE30161 | 2513 | $N_2 = 58$ | 48 (83%) | 36 (62%) | 22 (38%) | 18,524 |
| GSE9891 | 1140 | $N_3 = 260$ | 185 (71%) | 113 (43%) | 147 (57%) | 18,524 |
| TCGA | 1721 | $N_4 = 510$ | 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 |

*Note:* The data are extracted from the *curatedOvarianData* R Bioconductor package of Ganzfried et al.;[33] see Section 5 for details.

[a]The data sets are signified as GEO accession number which can be used to search the public genomics data in the GEO (Gene Expression Omnibus) repository. Extracted studies are the subset having documented values of 'days-to-tumor-recurrence', 'days-to-death', 'recurrence status', and 'vital status' for all patients. The median follow-up time is calculated from the Kaplan–Meier survival curve for time-to-censoring for each study. The event rates are calculated for each study.

It is of our interest to examine how the high-dimensional genomic information can be incorporated into a joint model of relapse and death, instead of *CXCL12* alone. Waldron et al.[13] performed a meta-analysis on the data from Ganzfried et al.[33] to examine the predictive values of these available genes. They applied several existing genomic predictors for predicting overall survival and confirmed their ability to predict survival. However, they also noted the modest gain in prediction power, suggesting the need of further improvement to be of clinical value.[13]

Clinicians are often required to update their predictions of death in patients after intermediate events (e.g., relapse of cancer). It was observed that the increase in the risk of death linked to cancer relapse is much greater than the effect linked to the change in *CXCL12* expression.[32] To account for such dynamic disease mechanism and build a highly accurate prediction formula of death, we shall develop a joint model of relapse and death, involving both clinical and genomic information as covariates.

## 2.2 Data structure

Meta-analytic data consist of $G$ independent studies with the $i$th study containing $N_i$ subjects for $i = 1, 2, \ldots, G$. Let $X_{ij}$ be time-to-tumour progression (TTP), $D_{ij}$ be time-to-death, and $C_{ij}$ be independent censoring time for $i = 1, 2, \ldots, G$ and $j = 1, 2, \ldots, N_i$. We observe the first-occurring event time $T_{ij} = \min(X_{ij}, D_{ij}, C_{ij})$, the status of tumour progression $\delta_{ij} = \mathbf{I}(T_{ij} = X_{ij})$, the terminal event time $T^*_{ij} = \min(D_{ij}, C_{ij})$, and the status for death $\delta^*_{ij} = \mathbf{I}(T^*_{ij} = D_{ij})$, where $\mathbf{I}(\cdot)$ is the indicator function. The observed data are expressed as $(T_{ij}, T^*_{ij}, \delta_{ij}, \delta^*_{ij})$ for $i = 1, 2, \ldots, G$ and $j = 1, 2, \ldots, N_i$, as in literature.[32,34]

The aforementioned data structure follows the semi-competing risks setting:[35] the variable $X_{ij}$ may be censored by $D_{ij}$, but $D_{ij}$ is never censored by $X_{ij}$. If a patient has documented records of both TTP and time-to-death, they correspond to $T_{ij} = X_{ij}$, $T^*_{ij} = D_{ij}$, and $\delta_{ij} = \delta^*_{ij} = 1$. Since part of patients provide information on both $X_{ij}$ and $D_{ij}$, the parameter determining the level of dependence between $X_{ij}$ and $D_{ij}$ can be estimated from the data.[32,34,35]

## 2.3 Joint frailty-copula model

Rondeau et al.[34] proposed a joint model tailored for meta-analysis. Let $\mathbf{Z}_{1,ij}$ be a vector of clinical covariates associated with TTP, and $\mathbf{Z}_{2,ij}$ be a vector of clinical covariates associated with time-to-death. To capture the heterogeneity of the studies, Rondeau et al.[34] considered unobserved study-specific frailties $u_i$ following a Gamma distribution with a density

$$f_\eta(u) = \frac{1}{\Gamma(1/\eta)\eta^{1/\eta}} u^{\frac{1}{\eta}-1} \exp\left(-\frac{u}{\eta}\right)$$

where $\eta > 0$ is their variance. The two hazards are jointly specified as

$$\begin{cases} r_{ij}(t|u_i) = u_i r_0(t) \exp(\boldsymbol{\beta}_1' \mathbf{Z}_{1,ij}) & \text{(for tumour progression } X_{ij}) \\ \lambda_{ij}(t|u_i) = u_i^\alpha \lambda_0(t) \exp(\boldsymbol{\beta}_2' \mathbf{Z}_{2,ij}) & \text{(for death } D_{ij}) \end{cases} \quad (1)$$

The forms of the baseline hazards $r_0(\cdot)$ and $\lambda_0(\cdot)$ are flexibly modelled, for example, using cubic M-splines.[36] The parameters $\boldsymbol{\beta}_1$ (or $\boldsymbol{\beta}_2$) are interpreted as fixed effects of $\mathbf{Z}_{1,ij}$(or $\mathbf{Z}_{2,ij}$) across studies. The random effects $u_i$ ($i = 1, 2, \ldots, G$) act multiplicatively on the baseline hazard functions and reflect the intra-study dependence between $X_{ij}$ and $D_{ij}$. Positive dependence, independence, or negative dependence corresponds to $\alpha > 0$, $\alpha = 0$, or $\alpha < 0$, respectively. In the joint frailty model of Rondeau et al.,[34] the conditional independence between $X_{ij}$ and $D_{ij}$ is assumed (given $u_i$, $\mathbf{Z}_{1,ij}$, and $\mathbf{Z}_{2,ij}$).

Residual dependence arises if patient-level characteristics (clinical covariates or genes) affecting both $X_{ij}$ and $D_{ij}$ are ignored in the model.[32,37] Consider a setting where survival prediction is made using a few covariates (e.g., age, cancer grade, tumour size). Residual dependence implies that TTP gives additional predictive information for time-to-death beyond these covariates. Hence, residual dependence plays a role on survival prediction unless TTP is completely predictable by these covariates. In meta-analysis, researchers may access only a few covariates that are consistently obtained across studies.

Emura et al.[32] extended the joint frailty model of Rondeau et al.[34] by introducing intra-subject (residual) dependence with a copula model

$$\Pr(X_{ij} > x, D_{ij} > y|u_i) = C_\theta[S_{Xij}(x|u_i), S_{Dij}(y|u_i)] \quad (2)$$

where $C_\theta$ is a copula[38] with an unknown parameter $\theta$. Here, $S_{Xij}(x|u) = \exp\{-uR_0(x)e^{\boldsymbol{\beta}_1'\mathbf{Z}_{1,ij}}\}$ and $S_{Dij}(y|u) = \exp\{-u^\alpha \Lambda_0(y)e^{\boldsymbol{\beta}_2'\mathbf{Z}_{2,ij}}\}$ are conditional survival functions following the joint frailty model (1), and $R_0(x) = \int_0^x r_0(v)\mathrm{d}v$ and $\Lambda_0(y) = \int_0^y \lambda_0(v)\mathrm{d}v$ are the baseline cumulative hazard functions.

The model (2) is called the joint frailty-copula model.[32] The copula describes intra-subject dependence between $X_{ij}$ and $D_{ij}$ (given $u_i$, $\mathbf{Z}_{1,ij}$, and $\mathbf{Z}_{2,ij}$). A convenient example is the Clayton copula

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

where the copula parameter $\theta$ determines the amount of dependence and is related to Kendall's $\tau$ via $\tau = \theta/(\theta + 2)$ (pp. 116–117 of Nelsen[38]). If $\theta \to 0$, then $C_\theta(v, w) = vw$ and $\tau = 0$. Thus, the model reduces to the joint frailty model of Rondeau et al.[34]

## 2.4 Incorporating high-dimensional genomic factors

Theoretically, it might be possible to consider the joint frailty model (1) treating $\mathbf{Z}_{1,ij}$ and $\mathbf{Z}_{2,ij}$ as high-dimensional genetic factors. However, estimation becomes infeasible for such a model without penalization techniques. While many techniques penalizing the Cox partial likelihood for high-dimensional covariates are available,[18,19,21,39,40] they are not straightforwardly applied to the joint model requiring the complicated full likelihood computation.

In this context, rather than a complete multivariate technique, we adopt a simple approach based on Tukey's CC,[41] as commonly employed in medical studies with microarrays.[8,9,13,15,20,37,42–44] The CC is a weighted sum of gene expressions, where the weight attached to a gene is a regression coefficient from the univariate Cox regression for the gene. The competitive performance of the CC to more sophisticated multivariate techniques, such as ridge and Lasso, was previously reported.[42]

In the literature, the CC has been mainly utilized for purpose of making class prediction, especially classifying patients into either good or poor prognosis group.[8,9,13,20,37,42,43] For purpose of predicting survival probability, the CC was first employed in Matsui et al.[44] who proposed to build a patient-level survival prediction with a single event. In the sequel, we shall adopt the approach of Matsui et al.[44] to the joint model.

For each subject $(i, j)$, we consider two sets of genomic factors

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

where $q_1$ and $q_2$ can be large numbers. The set $\mathbf{V}_{ij}$ is determined as those genes that have low $P$-values of testing the null hypothesis $H_0 : b_k = 0$ in the univariate Cox model, $r_{ij}(t) = r_0(t)\exp(b_k V_{ij,k})$; the set $\mathbf{W}_{ij}$ is determined in a similar fashion. We generally recommend the $P$-value threshold of 0.001,[45] but it depends on many different factors, such as the total number of genes.

We first form two CCs

$$\mathrm{CC}_{1,ij} = \hat{b}_1 V_{ij,1} + \cdots + \hat{b}_{q_1} V_{ij,q_1} \quad \text{(associated with tumour progression } X_{ij})$$
$$\mathrm{CC}_{2,ij} = \hat{c}_1 W_{ij,1} + \cdots + \hat{c}_{q_2} W_{ij,q_2} \quad \text{(associated with death } D_{ij})$$

where the weights $\hat{b}_k$ and $\hat{c}_k$ are estimates of regression coefficients under univariate Cox models on $k$th gene, $r_{ij}(t) = r_0(t)\exp(b_k V_{ij,k})$, and $\lambda_{ij}(t) = \lambda_0(t)\exp(c_k W_{ij,k})$, respectively. Since the scale of $\mathrm{CC}_{1,ij}$ (or $\mathrm{CC}_{2,ij}$) depends on the number $q_1$ (or $q_2$), we suggest standardizing $\mathrm{CC}_{1,ij}$ (or $\mathrm{CC}_{2,ij}$) to have a mean of 0 and SD of 1.

We propose to treat the CCs as new covariates and construct the joint frailty-copula model

$$\begin{cases} r_{ij}(t|u_i) = u_i r_0(t)\exp(\boldsymbol{\beta}_1'\mathbf{Z}_{1,ij} + \gamma_1 \mathrm{CC}_{1,ij}) & \text{(for tumour progression } X_{ij}) \\ \lambda_{ij}(t|u_i) = u_i^\alpha \lambda_0(t)\exp(\boldsymbol{\beta}_2'\mathbf{Z}_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & \text{(for death } D_{ij}) \\ \Pr(X_{ij} > x, D_{ij} > y|u_i) = C_\theta[S_{Xij}(x|u_i), S_{Dij}(y|u_i)] & \text{(between } X_{ij} \text{ and } D_{ij}) \end{cases} \quad (3)$$

One can regard $\mathrm{CC}_{1,ij}$ (or $\mathrm{CC}_{2,ij}$) as a risk score: patients with larger $\mathrm{CC}_{1,ij}$ have higher risk for relapse if $\gamma_1 > 0$. The univariate Cox regression is used to determine the weight of each gene involved in the risk score.

Our experiences in applying the CCs suggest that the predictive accuracy is relatively insensitive to changing the $P$-value threshold around 0.001. We nevertheless recommend conducting a sensitivity analysis on the choice of the $P$-value threshold (see Section 5.4).

## 3 Dynamic prediction with the joint frailty-copula model

Our focus is on predicting the probability of death for a new patient from a new study at a given time point $t > 0$. This prediction is conditional on the patient's covariate information $\mathbf{Z} = (\mathbf{Z}_1,\ \mathbf{Z}_2,\ \mathrm{CC}_1,\ \mathrm{CC}_2)$ measured at time $t = 0$ and dynamic information measured at time $t > 0$.

Let $D$ be time-to-death and $X$ be TTP of the new patient. In dynamic prediction, prediction of death at time $t > 0$ is meaningful only when the patient is still alive (i.e., $D > t$). Given the covariate information, the conditional probability of death between $t$ and $t + w$ is predicted as

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

If we add the dynamic information on the presence or absence of tumour progression status, the above prediction is modified as follows. First, suppose that the patient does not have a history of tumour progression at time $t$ (i.e., $X > t$). Given that the patient is alive at time $t$, the probability of death between $t$ and $t + w$ is

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

Second, if the patient has a history of tumour progression before time $t$, the time of the tumour progression (i.e., $X = x$) is available at time $t$. Given that the patient is still alive at time $t$, the probability of death between $t$ and $t + w$ is

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

Here, TTP occurring before time $t$ is observable, but TTP occurring after $t$ is not.

Tumour progression occurring to a patient may be strongly associated with death occurring to the same patient. If so, the two probabilities, $F(t, t + w|\ X = x,\ \mathbf{Z})$ and $F(t, t + w|X > t,\ \mathbf{Z})$, show large discrepancy: the former is much larger than the latter. The individual-level (intra-subject) dependence between $X$ and $D$ is essential to discriminate the two probabilities. In the following, we consider a prediction formula accounting for the individual-level dependence by the joint frailty-copula model.

### 3.1 Proposed prediction formula

Let $S_X(x|u) = \exp\{-uR_0(x)e^{\boldsymbol{\beta}_1'\mathbf{Z}_1+\gamma_1\mathrm{CC}_1}\}$ and $S_D(y|u) = \exp\{-u^{\alpha}\Lambda_0(y)e^{\boldsymbol{\beta}_2'\mathbf{Z}_2+\gamma_2\mathrm{CC}_2}\}$ be the conditional survival functions for a new patient having covariates $\mathbf{Z} = (\mathbf{Z}_1,\ \mathbf{Z}_2,\ \mathrm{CC}_1,\ \mathrm{CC}_2)$ measured at time $t = 0$. We assume that the patient's TTP and time-to-death follow the joint frailty-copula model

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

Dynamic information available at time $t > 0$ is defined as

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

Under these settings, the dynamic prediction formula is

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

The formula is divided into the following exclusive cases (derivations given in Appendix 1):

- Given that the patient does not experience tumour progression before time $t$ (i.e., $X > t$)

$$\begin{aligned} F(t, t + w|X > t,\ \mathbf{Z}) &= \Pr(D \leq t + w|D > t,\ X > t,\ \mathbf{Z}) \\ &= \frac{\int_0^{\infty} (C_{\theta}[S_X(t|u),\ S_D(t|u)] - C_{\theta}[S_X(t|u),\ S_D(t + w|u)])f_{\eta}(u)\mathrm{d}u}{\int_0^{\infty} C_{\theta}[S_X(t|u),\ S_D(t|u)]f_{\eta}(u)\mathrm{d}u} \end{aligned}$$

- Given that the patient experiences tumour progression before time $t$ (i.e., $X = x,\ x \le t$)

$$F(t, t+w|X=x,\ \mathbf{Z}) = \Pr(D \le t+w|D>t,\ X=x,\ \mathbf{Z})$$
$$= \frac{\int_0^\infty \left(C_\theta^{[1,0]}[S_X(x|u),\ S_D(t|u)] - C_\theta^{[1,0]}[S_X(x|u),\ S_D(t+w|u)]\right) u S_X(x|u) f_\eta(u)\mathrm{d}u}{\int_0^\infty C_\theta^{[1,0]}[S_X(x|u),\ S_D(t|u)] u S_X(x|u) f_\eta(u)\mathrm{d}u}$$

where $C_\theta^{[1,0]}(v,\ w) = \partial C_\theta(v,\ w)/\partial v$. For instance, under the Clayton copula, the formula $C_\theta^{[1,0]}(v,\ w) = v^{-\theta-1}(v^{-\theta} + w^{-\theta} - 1)^{-1/\theta-1}$ is used to calculate the probability.

If $\theta = 0$, the Clayton copula reduces to the independence copula $C_{\theta=0}(v,\ w) = vw$ (pp. 116–117 of Nelsen[38]). Applying $C_{\theta=0}^{[1,0]}(v,\ w) = w$ to the prediction formulas, one has

$$F(t, t+w|X>t,\ \mathbf{Z}) = \frac{\int_0^\infty (S_D(t|u) - S_D(t+w|u)) S_X(t|u) f_\eta(u)\mathrm{d}u}{\int_0^\infty S_D(t|u) S_X(t|u) f_\eta(u)\mathrm{d}u}$$
$$F(t, t+w|X=x,\ \mathbf{Z}) = \frac{\int_0^\infty (S_D(t|u) - S_D(t+w|u)) u S_X(x|u) f_\eta(u)\mathrm{d}u}{\int_0^\infty S_D(t|u) u S_X(x|u) f_\eta(u)\mathrm{d}u}$$

This prediction formula corresponds to the joint frailty model of Rondeau et al.,[34] where progression events affect death only through the study-specific frailty.

The unknown parameters in the joint models are estimated using the penalized likelihood approach.[32,34] The estimates $(\hat\theta,\ \hat\eta,\ \hat{\boldsymbol\beta}_1,\ \hat{\boldsymbol\beta}_2,\ \hat r_0,\ \hat\lambda_0)$ under the Clayton copula or under the independence copula (given $\theta = 0$ in the Clayton copula) can be computed through the *joint.Cox* R package.[46] Estimation under the case of $\theta = 0$ (independence copula) can also be implemented through the *frailtypack* R package.[47] In *joint.Cox*,[46] the value of $\alpha$ must be given by users. In practice, the value of $\alpha$ is chosen based on the profile-likelihood approach.[32] The estimate $\hat F$ of $F$ is obtained once all the unknown parameters are estimated.

The calculation of the prediction probability $\hat F$ and its graphical representations are implemented in the *joint.Cox* R package.[46] Supplementary Material contains a simple example of calculating the prediction probability $\hat F$.

The variability of the estimate $\hat F$ is measured by the 95% confidence interval. We suggest the percentile confidence interval based on the Monte Carlo simulation method, as previously employed.[24,30,31] The detailed algorithms are provided in Appendix 2.

**Remark.** Mauguen et al.[24] considered a prediction formula of death given recurrent events in a single study (not in meta-analysis). In their setting, the number of previous recurrences (denoted as $J = 0,\ 1,\ 2,\ \ldots$ in Mauguen et al.[24]) influences the prediction probability of death. In our setting, we obtain two separate prediction probabilities, $F(t, t+w|\ X>t,\ \mathbf{Z})$ and $F(t, t+w|\ X=x,\ \mathbf{Z})$, corresponding to $J = 0$ and $J = 1$, respectively.

## 4 Evaluation of prediction error

In order to validate the performance of the proposed prediction formula, we introduce a traditional measure for predictive accuracy in survival analysis, known as the Brier score.[1,2]

### 4.1 Brier score

The true prediction error, defined as the Brier score, is

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

where the expectation is taken over the distribution of $(D, X, \mathbf{Z})$ given $\hat S(t,\ t+w|\ H(t,x),\ \mathbf{Z}) = 1 - \hat F(t,\ t+w|\ H(t,x),\ \mathbf{Z})$.

The idea of Graf et al.[1] and Gerds and Schumacher[2] can be applied to get an estimator of $Err(t, t+w)$. Let $Y(t) = \sum_{ij} \mathbf{I}(T_{ij}^* > t)$ be the number of subjects at risk at time $t$. Then, the estimated prediction error is

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

where

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

and where $\hat{G}(t)$ is the estimate of the censoring survival function $G(t) = P(C_{ij} > t)$. The inverse probability of censoring weight for $\hat{w}_{ij}(\cdot)$ corrects the bias due to censoring.[2] We use the Kaplan–Meier estimator for $\hat{G}$ by treating $T_{ij}^*$ as event time and $1 - \delta_{ij}^*$ as event indicator. The variability of $\hat{Err}(t, t+w)$ is measured in terms of the 95% point-wise confidence interval for each $(t, t+w)$. One can apply the bootstrap percentile interval, which is based on $B = 1000$ random samplings (with replacement) from the risk set $\{(i,j) : T_{ij}^* > t\}$.

The performance of the prediction formula $\hat{F}$ is evaluated by examining how the estimated error $\hat{Err}(t, t+w)$ reduces remarkably from a benchmark value. A common benchmark is the estimated prediction error by the Kaplan–Meier estimator

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

where $\hat{S}^{KM}(t, t+w) = \hat{S}^{KM}(t+w)/\hat{S}^{KM}(t)$ and $\hat{S}^{KM}$ is the Kaplan–Meier estimator by treating $T_{ij}^*$ as event time and $\delta_{ij}^*$ as event indicator for all $i$ and $j$. Prediction with the Kaplan–Meier estimator is interpreted as the 'average prediction', as it pools all the data and then calculates the overall prediction probability. Roughly speaking, $\hat{S}^{KM}$ is regarded as the best possible predictor when one cannot utilize any individual (covariate) information.

## 4.2 Optimism bias of estimated error $\hat{Err}(t, t+w)$

Typically, the estimator $\hat{Err}(t, t+w)$ yields some underestimation of the true value $Err(t, t+w)$ in finite sample. This is because we use the same data to estimate parameters in a given prediction model and to estimate the error, a typical phenomenon known as 'optimism bias' of the error estimate.[48] Prediction models involving high-dimensional factors require serious attention to the bias due to over-fitting.[16,20,43]

In our approach, the concern for the optimism bias becomes not severe due to the dimension reduction step of the CCs. In Supplemental Material, we evaluated the degree of the optimism bias by simulations, where the true prediction error $Err(t, t+w)$ is known by the simulation design. The simulations reveal that the estimated prediction error $\hat{Err}(t, t+w)$ shows modest downward bias but is still a good substitute for the true prediction error $Err(t, t+w)$. One can remove the optimism bias by applying a cross-validation, as previously employed under the joint models.[24] We detail the cross-validation scheme in Section 5.

# 5 Data analysis

In this data analysis, we aim to develop a dynamic prediction formula of death (i.e., overall survival) for ovarian cancer patients. The prediction formula was constructed on the basis of the joint frailty-copula model that incorporates both clinical and genomic covariates.

We used the subset of the ovarian cancer data of Ganzfried et al.[33] to estimate parameters in the model. We first applied the prescribed patient selection criterion given by the *curatedOvarianData* R package, and then extracted the studies that have documented values of 'days-to-tumor recurrence', 'days-to-death', 'recurrence status', and 'vital status' for all patients. The same process was carried out in our previous analysis of the same data.[32] However, our present analysis yielded a slightly reduced list of patients (Table 1). The reason may be due to the update of 'patientselection.config' file (from older version 1.0.3 to the latest version 1.8.0) in the *curatedOvarianData* R package to avoid some duplicate removal.[13]

Our extracted data consist of 912 ovarian cancer patients (544 relapsed, 465 died, and 447 censored) from four different studies (Table 1). Observed genomic factors are 11,756 gene expressions consistently available across the four studies. All the expression values are standardized (mean of 0 and SD of 1 in the 912 patients). Several clinical covariates in the data of Ganzfried et al.[33] contained missing values at study-level (e.g., age is completely missing in

two studies). Therefore, we focused on the two covariates considered in Ganzfried et al.[33]: the residual tumour size at surgery and the FIGO stage.

## 5.1 Compound covariate

Among 11,756 genes commonly available in the four studies, we chose a subset consisting of 6056 genes whose coefficient of variation in expression values in the 912 patients is greater than 3%. Then, the univariate Cox regression analyses give the CCs

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

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

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

involving 128 genes ($P$-value $< 0.001$ for time-to-death). In the above expressions, the gene symbols were ordered by their significance. For instance, *CXCL12* was the most strongly associated gene for time-to-relapse. Supplementary Material details known biological functions of selected genes (*TIMP2*, *PDPN*, *NCOA3*, *TEAD1*, *YWHAB*) involved in the expressions of $\mathrm{CC}_{1,ij}$ and $\mathrm{CC}_{2,ij}$. The full lists of genes with their coefficients are given in Supplementary Material. We standardized $\mathrm{CC}_{1,ij}$ and $\mathrm{CC}_{2,ij}$ to have mean of 0 and SD of 1.

## 5.2 Fitting joint frailty-copula model

The *joint.Cox* R package[46] was applied to fit the data to the joint frailty-copula model. We selected covariates in a stepwise fashion and arrived at a model

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

Here, the clinical covariate is the binary variable ($Z_{2,ij} = 0$ vs. $= 1$) on the residual tumour size at surgery ($<1$ cm vs. $\geq 1$ cm). The FIGO stage was not included in the model. The details of the covariate selection method are given in Supplemental Material.

All the regression coefficients were significant ($P$-value $< 0.05$). Their relative risks are $\exp(\beta_2) = 1.18$ (95%CI: 1.03–1.35), $\exp(\gamma_1) = 1.48$ (95%CI: 1.37–1.59), and $\exp(\gamma_2) = 1.56$ (95%CI: 1.44–1.70). The copula parameter was $\theta = 1.90$ (95%CI: 1.49–2.42), and the corresponding Kendall's tau was $\tau = 0.49$ (95%CI: 0.32–0.65), representing moderate positive dependence between time-to-relapse and time-to-death. The heterogeneity parameter was $Var(u_i) = \eta = 0.039$ (95%CI: 0.007–0.227).

We set the value $\alpha = 0$ which maximized the profile penalized likelihood[32] for the data. Hence, there is no heterogeneity of death rates among the four studies. Indeed, the heterogeneity of death rates among the four studies (Table 1) may be mostly explained by the heterogeneity of the median follow-up times.

## 5.3 Predicted risk of death for individual patients

To demonstrate the proposed prediction formula, we consider two hypothetical patients (named Patient 1 and Patient 2) having the following characteristics: Patient 1 has high-risk factors at $t = 0$ ($\mathrm{CC}_1 = 1$, $\mathrm{CC}_2 = 1$, the residual tumour size $\geq 1$ cm) but does not experience relapse during the follow-up. Patient 2 has low-risk factors at $t = 0$ ($\mathrm{CC}_1 = -1$, $\mathrm{CC}_2 = -1$, the residual tumour size $< 1$ cm) but experiences relapse at $x = 600$ days after surgery. The *joint.Cox* R package[46] was used to compute the dynamic prediction formula $F(t,\ t + w|\ H(t, x),\ \mathbf{Z})$ at prediction time $t = 500$ (early) or $t = 1000$ (late) over the prediction horizon $t + w \leq 3500$ (in days). The 95% confidence intervals are also given.

Figure 1 displays the predicted probabilities of death for Patient 1 and Patient 2. At the early prediction time ($t = 500$ days), Patient 1 has higher predicted probabilities of death due to the higher risk factors, compared to

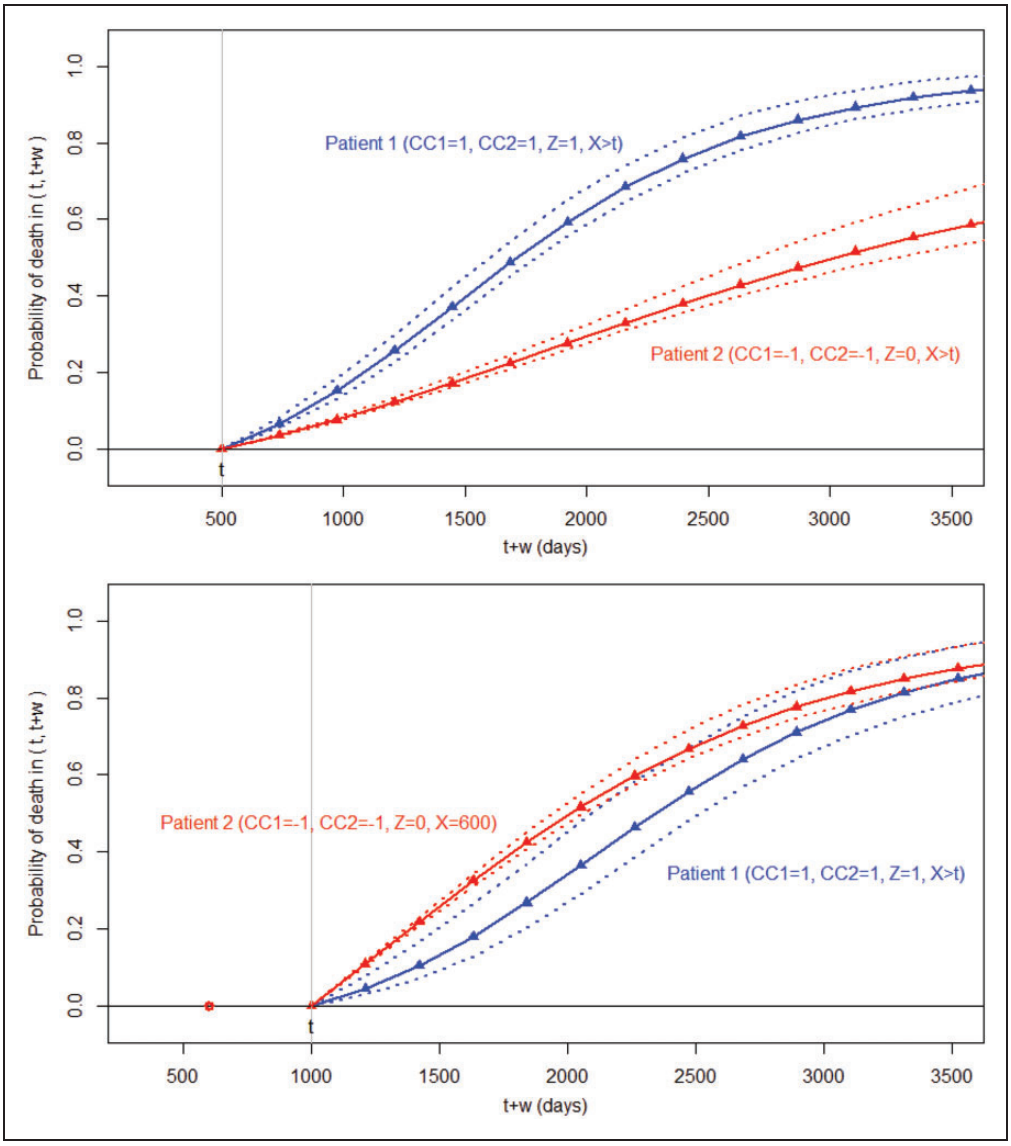


**Figure 1.** Probability of death between $t$ and $t + w$ for two patients: Patient 1 has high-risk factors at $t = 0$ ($CC_1 = 1$, $CC_2 = 1$, the residual tumour size $\geq 1$ cm) and does not experience relapse during the follow-up. Patient 2 has low-risk factors at $t = 0$ ($CC_1 = -1$, $CC_2 = -1$, the residual tumour size $< 1$ cm) and experiences relapse at 600 days. The vertical grey line corresponds to $t = 500$ or $t = 1000$ (days). Dotted lines represent the 95% confidence intervals.

Patient 2. However, at the late prediction time ($t = 1000$ days), the predicted probabilities of death for Patient 2 get consistently higher than those of Patient 1. Clearly, the increase in the risk of Patient 2 is due to the relapse occurred at 600 days. This explains that, for late prediction time, the relapse event is a stronger risk factor than all the three risk factors ($CC_1$, $CC_2$, and the residual tumour size).

## 5.4 Assessing prediction error

We assessed the prediction error of the proposed prediction formula that involves $CC_1$, $CC_2$, and the residual tumour size. We calculated the prediction error estimate $\hat{Err}(t, t + w)$ and the 95% confidence intervals. We then compared them with the benchmark value $\hat{Err}^{KM}(t, t + w)$, treating the latter as fixed and known.

Figure 2 displays the prediction error curve $\hat{Err}(t, t + w)$ of the dynamic prediction formula for a given prediction time $t = 500$ (early) or 1000 (late) with the range $t + w \leq 3000$ (in days). For the early prediction

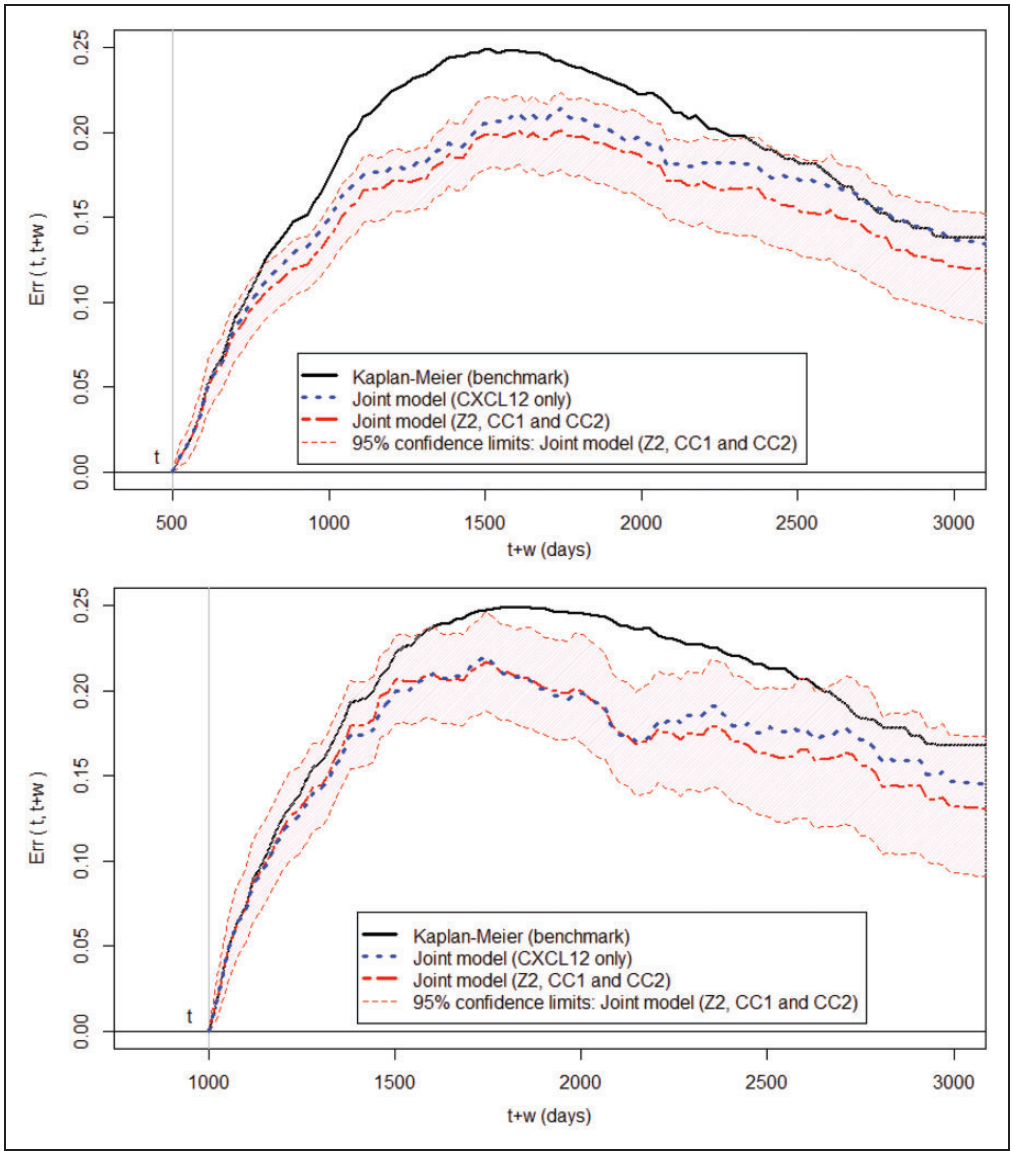


**Figure 2.** Comparison of prediction errors (Brier scores). (dashed line): joint model with Z2, CC1, and CC2, (shaded area): 95% confidence region with the joint model with Z2, CC1, and CC2, (dotted line): joint model with *CXCL12* alone, (vertical grey line): prediction time $t = 500$ or 1000 days.

time ($t = 500$), the prediction errors were smaller than the benchmark values in the whole range $500 \leq t + w \leq 3000$. The 95% confidence interval did not cover the benchmark value in the range $800 \leq t + w \leq 2300$, implying a significant reduction in prediction error. For the late prediction time ($t = 1000$), the prediction errors were also smaller than the benchmark values but had wider confidence intervals due to the smaller number of patients in the risk set.

We also compare the proposed dynamic prediction formula with the dynamic prediction formula incorporating *CXCL12* alone that was used in Emura et al.[32] For the early prediction time ($t = 500$), the *CXLC12*-alone formula had smaller prediction error than the benchmark model but had larger prediction error than the proposed prediction formula. However, in the late prediction time ($t = 1000$), the advantage of the proposed prediction formula over the *CXLC12*-alone formula becomes less clear. The reason may be that accumulated relapse events up to $t = 1000$ erase the impact of clinical and genomic covariates. A similar phenomenon was seen in different contexts of dynamic prediction.[24,32]

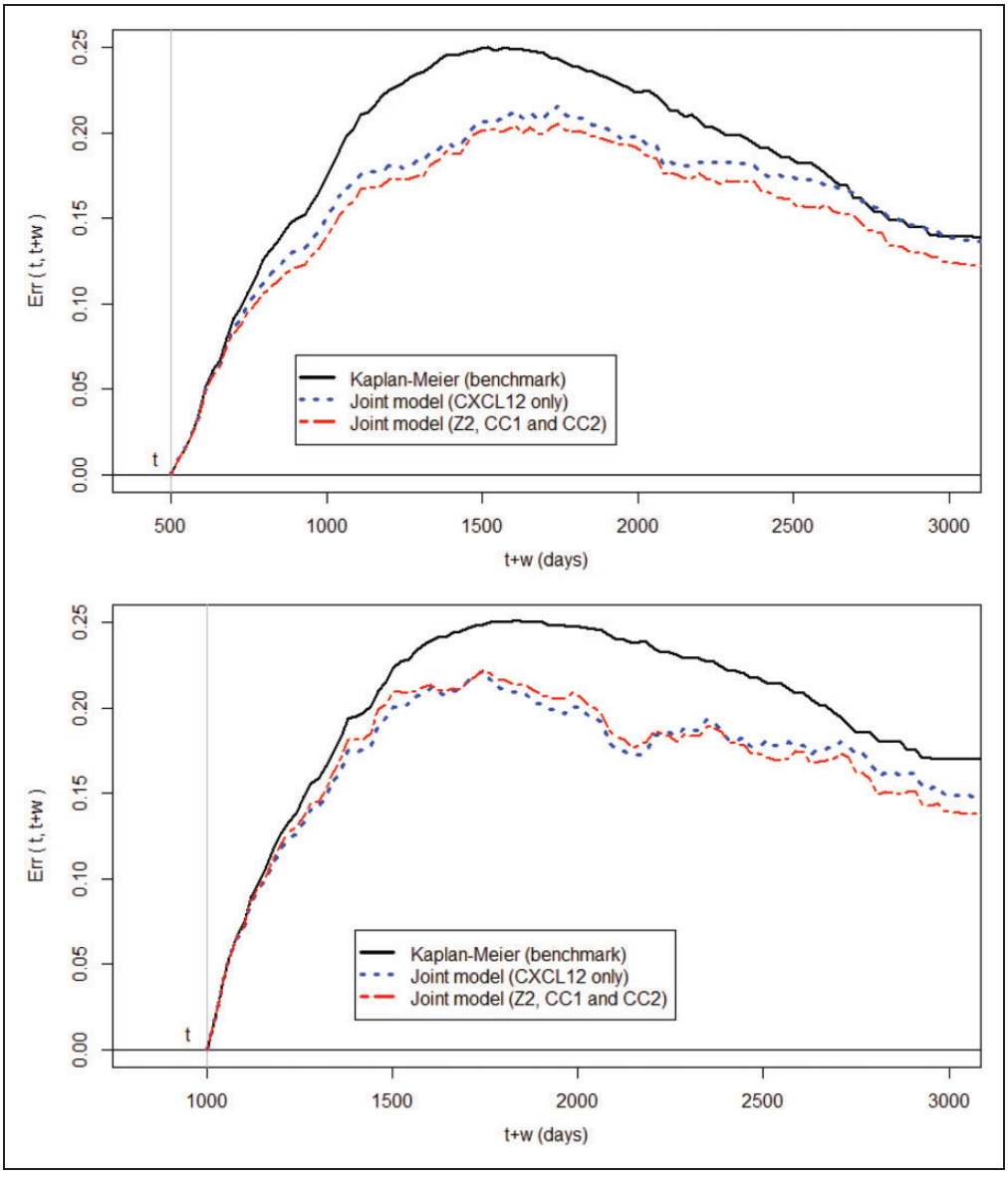


**Figure 3.** Comparison of cross-validated prediction errors (Brier scores). (dashed line): joint model with Z2, CC1, and CC2, (dotted line): joint model with *CXCL12* alone, (vertical grey line): prediction time $t$ =500 or 1000 days.

To assess the degree of the optimism bias, we modified $\hat{Err}(t, t+w)$ using leave-one-out cross-validation. We removed one patient $(i,j)$ from the risk set $\{(i,j) : T^*_{ij} > t\}$, and re-estimated the model parameters using the remaining subsamples (911 patients). All the model building steps, including selection of genes and estimation of regression coefficients, were based on the subsamples. Repeating this process for all patients in the risk set, the cross-validated prediction error was computed as

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

where $\hat{S}^{-(i,j)}(t, t+w|\cdot,\ \cdot)$ did not use any information for the left-out patient $(i,j)$.

Figure 3 reveals that cross-validated prediction error estimates do not differ too much from the prediction error estimates without cross-validation. For the proposed prediction formula involving the CCs, the variability due to selection of genes and estimation of coefficients in the CCs leads to minor increase in cross-validated

prediction errors over the prediction errors without cross-validation. This is a remarkable property of using the CC to handle high-dimensional genetic factors. The figure shows the clear advantage of the proposed dynamic prediction formula over the prediction by the Kaplan–Meier estimator. However, at the late prediction time ($t = 1000$), the benefit of the proposed dynamic prediction formula over the *CXCL12*-alone formula becomes questionable. Indeed, the *CXCL12*-alone formula can utilize the dynamic tumour progression information in the same manner as the proposed model. We have seen the same pattern in the prediction error estimates without cross-validation.

Finally, we conducted a sensitivity analysis by slightly increasing *P*-value threshold of 0.001 for the univariate gene selection (Section 5.1). Then, more genes were included into the CCs, but the prediction error curves did not show any visible change. This may be because the CCs were constructed in a stepwise manner: the CCs only updated the newly included genes after increasing the *P*-value threshold.

## 6 Conclusion and discussion

With the increasing availability of rich information, researchers may easily obtain meta-analytic data sets containing patients' multivariate time-to-events (death, tumour progression, PFS, etc.), high-dimensional genomic factors (e.g., gene expressions), and other clinical information. However, building a personalized survival prediction formula remains a challenging problem, as it requires the precise specifications of the dependency between multiple events (death and tumour progression) as well as their relationship with both genomic and clinical covariates. An additional challenge comes from the heterogeneity among studies ubiquitous in meta-analysis, which typically demands random effect or frailty modelling.

This paper offers a statistical approach using a joint model to implement a personalized dynamic prediction based on IPD meta-analysis with genomic factors. Our study touches on the realm of individualized or personalized medicine according to the definition of 'personalized' or 'individualized' medicine that aims to improve stratification and timing of health care by using biological and genomic information.[49]

We used the ovarian cancer data to demonstrate the individual-patient prediction of death according to his/her clinical and genomic information. Using the proposed prediction model, we observed that the information on both clinical and genomic factors allows clinicians to discriminate between good and poor prognosis patients at early prediction time (e.g., $t = 500$ days after surgery). On the other hand, at late prediction time (e.g., $t = 1000$ days after surgery), relapse information offers stronger predictive power than the clinical and genomic factors. While the high-relevance of relapse information on death is widely recognized, we confirmed it numerically through the proposed prediction formula.

If clinicians wish to draw the correct conclusion from the proposed dynamic prediction formula, the prediction time $t$ must be pre-specified (e.g., $t = 1000$ days after treatment). It is not a correct way to make prediction at the time of tumour progression for each patient. Hence, a sensible choice of $t$ is required that has to come from clinical perspective and not from statistical perspective.

The performance of the proposed prediction formula was assessed using the Brier score, the mean squared error of predicting dichotomous event (death or alive) in a time horizon. The estimated prediction error showed modest downward bias (optimism bias) relative to the true prediction error (simulations detailed in Supplementary Material). We have intentionally used the CCs to avoid the extreme overfitting by initially reducing the dimension of genomic factors with univariate selection. This is not a contradiction to a well-known phenomenon of formidable underestimation of prediction error for rich statistical models.[16] While one can always use cross-validation to adjust for the optimism bias in real data analysis, cross-validation has high computational demand for the joint model. Alternatively, the estimated prediction error without cross-validation can be suggested as a good substitute for the true prediction error. Otherwise, some analytical cross-validation method would be useful.[50]

To handle high-dimensional genetic factors, we adopted a simple approach based on Tukey's CC[41] followed by the univariate selection of genes with the *P*-value threshold of 0.001.[45] The univariate selection yielded 128 genes associated with time-to-relapse and 158 genes associated with time-to-death (*P*-value $< 0.001$). In our illustrative example of ovarian cancer patients, the joint model incorporating the CCs showed better predictive ability than the model incorporating *CXCL12* alone. For a new patient (not in the ovarian cancer data), we suggest using the same set of genes (128 genes for time-to-relapse and 158 genes time-to-death) and their coefficients to form CCs. As we described in Supplementary Material, some of these genes have known biological functions associated with patient survival, and the formulas of CCs appear to be consistent with these known functions. With these statistical and biological supports for the adopted approach, a personalized prediction of death for ovarian

cancer patients would be achieved. However, we recognize the need of testing our approach with an independent validation set of patients, before it is widely applied by clinicians.

Van Houwelingen et al.[40] applied a similar approach as Tukey's CC, where the coefficients attached to genes come from the ridge regression. This ridge approach could be used without the pre-selection of genes.[17,40] Instead, the ridge approach requires one to select a shrinkage parameter that plays a similar role as the *P*-value threshold of the CC.

We conducted simulation studies to compare the performance between the ridge-based approach and the CC approach as detailed in Supplementary Material. The results showed a remarkable contrast between the two approaches. The ridge provided smaller estimated prediction error (without cross-validation) than the CC. However, the ridge exhibited larger true prediction error over the CC. This over-fitting phenomenon of the ridge may be caused by a large number of genes in the model. Hence, with an appropriate *P*-value cut-off, the CC would be superior to the ridge in real prediction settings. One drawback of the CC is that the choice of the *P*-value cut-off is somewhat subjective. In this paper, we suggested the *P*-value threshold of 0.001,[45] though there is a more sophisticated approach that optimizes a cross-validated partial likelihood.[15]

Clearly, some dependence exists among time-to-death (overall survival), TTP, progression free survival (PFS), and their dependence pattern is essential to enhance the predictive value of death after progression events. In meta-analytic studies of cancer patients, the dependence is fundamental in validating the surrogacy of TTP or PFS for overall survival.[51–55] The joint frailty-copula model is a tailored model to analyse the dependence via copulas and the heterogeneity via frailty in meta-analytic settings. However, developing a formal validation process of surrogacy of TTP or PFS requires further extensions of the joint frailty-copula model, which would be our next topic for investigation.

Time-varying effects of clinical covariates and genetic factors are another important issue to be investigated. If clinical follow-up of patients is long, the prognostic effect of covariates may vary over time. For instance, time-varying effects of hormone receptors on recurrences (or metastases) were reported in breast cancer patients.[56–58]

One simplest way to introduce such time-varying effects of genes in the joint model is to add (gene $\times$ time) interaction terms, $\mathrm{CC}_{1,ij} \times f_1(t)$ and $\mathrm{CC}_{2\,ij} \times f_2(t)$, where $f_\ell(t)$, $\ell = 1, 2$, are flexibly chosen by users. One choice is $f_\ell(t) = \log(t+1)$[23,59] while a more elaborate choice is a B-spline approximation.[57]. In this way, the CCs account for 'common' time-varying effects of genes. This approach would be suitable if the majority of genes in the CCs share a similar time-varying effect on survival.

In reality, individual genes may have different time-varying effects on survival. For instance, one may categorize genes into two groups, namely a set of genes with short-term effect and a set of genes with long-term effect. In this way, time-varying effects can be homogeneous for genes in the same group. However, this strategy requires a way of grouping genes in order to reduce the heterogeneity of time-varying effects within a group.

Under time-varying effects, one can straightforwardly define the joint frailty-copula model but cannot exploit the computational advantage of the cubic spline models for the baseline hazards.[32] As a result, likelihood-based inference becomes computationally demanding. One possible alternative is to impose piecewise exponential models as in Mazroui et al.[57] It is fascinating that the Weibull or Pareto model for the baseline hazards may result in tractable forms under the time function $f_\ell(t) = \log(t+1)$. While we do not pursue these approaches here, a further development would be an interesting topic for investigation.

## Acknowledgements

The authors kindly thank the two anonymous referees for their valuable suggestions that improved the paper.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by Taiwan Ministry of Science and Technology (MOST 103-2118-M-008-MY2), CREST, JST, and a Grant-in-Aid for Scientific Research (16H06299) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## Supplemental material

Supplemental material is available for this article online.

## References


1. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999; **18**: 2529–2545.
2. Gerds TA and Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical J* 2006; **48**: 1029–1040.
3. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Ser B* 1972; **34**: 187–220.
4. Jenssen TK, Kuo WP, Stokke T, et al. Association between gene expressions in breast cancer and patient survival. *Hum Genet* 2002; **111**: 411–420.
5. Sabatier R, Finetti P, Adelaide J, et al. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. *PLoS One* 2011; **6**: e27656.
6. Lossos IS, Czerwinski DK, Alizadeh AA, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N Engl J Med* 2004; **350**: 1828–1837.
7. Alizadeh AA, Gentles AJ, Alencar AJ, 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**: 1350–1358.
8. Beer DG, Kardia SLR, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002; **8**: 816–824.
9. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; **356**: 11–20.
10. Shedden K, Taylor JMG, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; **14**: 822–827.
11. Popple A, Durrant LG, Spendlove I, et al. The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. *Br J Canc* 2012; **106**: 1306–1313.
12. Yoshihara K, Tsunoda T, Shigemizu D, et al. High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. *Clin Canc Res* 2012; **18**: 1374–1385.
13. Waldron L, Haibe-Kains B, Culhane AC, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J Natl Canc Inst* 2014; **106**: dju049.
14. Michiels S, Koscielny S and Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005; **365**: 488–492.
15. Matsui S. Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinformatics* 2006; **7**: 156.
16. Schumacher M, Binder H and Gerds T. Assessment of survival prediction models based on microarray data. *Bioinformatics* 2007; **23**: 1768–1774.
17. Bøvelstad HM, Nygård S, Storvold HL, et al. Predicting survival from microarray data – a comparative study. *Bioinformatics* 2007; **23**: 2080–2087.
18. Binder H and Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 2008; **9**: 14.
19. Bøvelstad HM, Nygård S and Borgan Ø. Survival prediction from clinico-genomic models – a comparative study. *BMC Bioinformatics* 2009; **10**: 1.
20. Simon RM, Subramanian J, Li MC, et al. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinformatics* 2011; **12**: 203–214.
21. Goeman J, Meijer R and 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.
22. Van Houwelingen H. Dynamic prediction by landmarking in event history analysis. *Scand J Stat* 2007; **34**: 70–85.
23. Van Houwelingen HC and Putter H. *Dynamic prediction in clinical survival analysis*. Boca Raton: CRC Press, 2011.
24. Mauguen A, Rachet B, Mathoulin-Pélissier S, et al. Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Stat Med* 2013; **32**: 5366–5380.
25. Mauguen A, Rachet B, Mathoulin-Pélissier S, et al. Validation of death prediction after breast cancer relapses using joint models. *BMC Med Res Methodol* 2015; **15**: 27.
26. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; **67**: 819–829.
27. Taylor JM, Park Y, Ankerst DP, et al. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 2013; **69**: 206–213.
28. Sène M, Taylor JM, Dignam JJ, et al. Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: development and validation. *Stat Methods in Med Res* 2016; **25**: 2972–2991.
29. Proust-Lima C, Séne M, Taylor JM, et al. Joint latent class models for longitudinal and time-to-event data: a review. *Stat Methods Med Res* 2014; **23**: 74–90.

30. Rondeau V, Mauguen A, Laurent A, et al. Dynamic prediction models for clustered and interval-censored outcomes: investigating the intra-couple correlation in the risk of dementia. *Stat Methods Med Res* 2015; **26**: 2168–2183.
31. Król A, Ferrer L, Pignon JP, et al. Joint model for left-censored longitudinal data, recurrent events and terminal event: predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000-05 trial. *Biometrics* 2016. DOI: 10.1111/biom.12490.
32. Emura T, Nakatochi M, Murotani K, et al. A joint frailty-copula model between tumour progression and death for meta-analysis. *Stat Methods Med Res* 2015; **26**: 2649–2666.
33. Ganzfried BF, Riester M, Haibe-Kains B, et al. Curated ovarian data: clinically annotated data for the ovarian cancer transcriptome. *Database* 2013. Article ID bat013. DOI:10.1093/database/bat013.
34. Rondeau V, Pignon JP and 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. *Stat Methods Med Res* 2015; **24**: 711–729.
35. Haneuse S and Lee KH. Semi-competing risks data analysis, accounting for death as a competing risk when the outcome of interest is nonterminal. *Circ Cardiovasc Qual Outcomes* 2016; **9**: 322–331.
36. Joly P, Commenges D and Letenneur L. A penalized likelihood approach for arbitrary censored and truncated data: application to age-specific incidence of dementia. *Biometrics* 1998; **54**: 185–194.
37. Emura T and Chen YH. Gene selection for survival data under dependent censoring, a copula-based approach. *Stat Methods Med Res* 2016; **25**: 2840–2857.
38. Nelsen RB. *An introduction to copulas*, 2nd ed. New York: Springer, 2006.
39. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997; **16**: 385–395.
40. Van Houwelingen HC, Bruinsma T, Hart AA, et al. Cross-validated Cox regression on microarray gene expression data. *Stat Med* 2006; **25**: 3201–3216.
41. Tukey JW. Tightening the clinical trial. *Contr Clin Trials* 1993; **14**: 266–285.
42. Emura T, Chen YH and Chen HY. Survival prediction based on compound covariate under Cox proportional hazard models. *PLoS One* 2012; **7**: e47627.
43. Radmacher MD, Mcshane LM and Simon RM. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002; **9**: 505–511.
44. Matsui S, Simon RM, Qu P, et al. Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clin Canc Res* 2012; **18**: 6065–6073.
45. Simon RM. *Design and analysis of DNA microarray investigations*. New York: Springer Science & Business Media, 2003.
46. 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.
47. Rondeau V, Gonzalez JR, Mazroui Y, et al. R frailtypack: general frailty models: shared, joint and nested frailty models with prediction. *CRAN*, version 2.8.3 2016-01-13.
48. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning*. New York: Springer, 2009.
49. Schleidgen S, Klingler C, Bertram T, et al. What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Med Ethic* 2013; **14**: 55.
50. Commenges D, Proust-Lima C, Samieri C, et al. A universal approximate cross-validation criterion for regular risk functions. *Int J Biostat* 2015; **11**: 51–67.
51. Burzykowski T, Molenberghs G, Buyse M, et al. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Appl Stat* 2001; **50**: 405–422.
52. Burzykowski T, Molenberghs G and Buyse M (eds) *The evaluation of surrogate endpoints*. New York: Springer, 2005.
53. Michiels S, Le Maître A, Buyse M, et al. Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data. *Lancet Oncol* 2009; **10**: 341–350.
54. Buyse M, Sargent DJ and Saad ED. Survival is not a good outcome for randomized trials with effective subsequent therapies. *J Clin Oncol* 2011; **29**: 4719–4720.
55. Oba K, Paoletti X, Alberts S, et al. Disease-free survival as a surrogate for overall survival in adjuvant trials of gastric cancer: a meta-analysis. *J Natl Canc Inst* 2013; **105**: 1600–1607.
56. Bellera CA, MacGrogan G, Debled M, et al. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med Res Methodol* 2010; **10**: 1.
57. Mazroui Y, Mauguen A, Mathoulin-Pélissier S, et al. Time-varying coefficients in a multivariate frailty model: application to breast cancer recurrences of several types and death. *Lifetime Data Anal* 2016; **22**: 191–215.
58. Baulies S, Belin L, Mallon P, et al. Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy. *Br J Canc* 2015; **113**: 30–36.
59. Putter H, Sasako M, Hartgrink H, et al. Long-term survival with non-proportional hazards: results from the Dutch Gastric Cancer Trial. *Stat Med* 2005; **24**: 2807–2821.

## Appendix 1: Derivation of prediction formula

- Given that the patient does not experience tumour progression before time $t$ (i.e., $X > t$)

$$\begin{aligned} F(t, t+w|X>t,\ \mathbf{Z}) &= \Pr(D \le t+w|D>t,\ X>t,\ \mathbf{Z}) \\ &= \frac{\Pr(D>t,\ X>t|\mathbf{Z}) - \Pr(D>t+w,\ X>t|\ \mathbf{Z})}{\Pr(D>t,\ X>t|\ \mathbf{Z})} \\ &= \frac{\int_0^\infty (\Pr(D>t,\ X>t|u,\ \mathbf{Z}) - \Pr(D>t+w,\ X>t|u,\ \mathbf{Z}))f_\eta(u)\mathrm{d}u}{\int_0^\infty \Pr(D>t,\ X>t|u,\ \mathbf{Z})f_\eta(u)\mathrm{d}u} \\ &= \frac{\int_0^\infty (C_\theta[S_X(t|u),\ S_D(t|u)] - C_\theta[S_X(t|u),\ S_D(t+w|u)])f_\eta(u)\mathrm{d}u}{\int_0^\infty C_\theta[S_X(t|u),\ S_D(t|u)]f_\eta(u)\mathrm{d}u}. \end{aligned}$$

- Given that the patient experiences tumour progression at time $x \le t$

$$\begin{aligned} &F(t, t+w|X=x,\ \mathbf{Z}) = \Pr(D \le t+w|D>t,\ X=x,\ \mathbf{Z}) \\ &= \frac{\Pr(D>t,\ X=x|\mathbf{Z}) - \Pr(D>t+w,\ X=x|\mathbf{Z})}{\Pr(D>t,\ X=x|\mathbf{Z})} \\ &= \frac{\int_0^\infty (\Pr(D>t,\ X=x|u,\ \mathbf{Z}) - \Pr(D>t+w,\ X=x|u,\ \mathbf{Z}))f_\eta(u)du}{\int_0^\infty \Pr(D>t,\ X=x|u,\ \mathbf{Z})f_\eta(u)\mathrm{d}u} \\ &= \frac{\int_0^\infty \left(-\frac{\partial}{\partial x}\Pr(D>t,\ X>x|u,\ \mathbf{Z}) - \left\{-\frac{\partial}{\partial x}\Pr(D>t+w,\ X>x|u,\ \mathbf{Z})\right\}\right)f_\eta(u)\mathrm{d}u}{\int_0^\infty -\frac{\partial}{\partial x}\Pr(D>t,\ X>x|u,\ \mathbf{Z})f_\eta(u)\mathrm{d}u} \\ &= \frac{\int_0^\infty \left(-\frac{\partial}{\partial x}C_\theta[S_X(x|u),\ S_D(t|u)] - \left\{-\frac{\partial}{\partial x}C_\theta[S_X(x|u),\ S_D(t+w|u)]\right\}\right)f_\eta(u)\mathrm{d}u}{\int_0^\infty -\frac{\partial}{\partial x}C_\theta[S_X(x|u),\ S_D(t|u)]f_\eta(u)\mathrm{d}u} \\ &= \frac{\int_0^\infty \left(C_\theta^{[1,0]}[S_X(x|u),\ S_D(t|u)] - C_\theta^{[1,0]}[S_X(x|u),\ S_D(t+w|u)]\right)uS_X(x|u)f_\eta(u)\mathrm{d}u}{\int_0^\infty C_\theta^{[1,0]}[S_X(x|u),\ S_D(t|u)]uS_X(x|u)f_\eta(u)\mathrm{d}u}, \end{aligned}$$

where the last expression follows from $-\partial S_X(x|u)/\partial x = ur_0(x)\exp(\boldsymbol{\beta}_1'\mathbf{Z} + \gamma_1 \mathrm{CC}_1)S_X(x|u)$.

## Appendix 2: Confidence interval for *F*

To enhance the accuracy of normal approximation, we consider log-transformed parameters $\tilde{\eta} = \log(\eta)$, $\tilde{\theta} = \log(\theta)$, $\tilde{\mathbf{g}} = \log(\mathbf{g})$ and $\tilde{\mathbf{h}} = \log(\mathbf{h})$, where $\mathbf{g} = (g_1, \ldots, g_{L_r})'$ and $\mathbf{h} = (h_1, \ldots, h_{L_\lambda})'$ are coefficients for $r_0(t) = \sum_{\ell=1}^{L_r} g_\ell M_\ell(t)$ and $\lambda_0(t) = \sum_{\ell=1}^{L_\lambda} h_\ell M_\ell(t)$, respectively. The log-likelihood function of Emura et al.[32] can be re-expressed as

$$\ell(\eta,\ \theta,\ \boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2,\ \mathbf{g},\ \mathbf{h}) = \ell(\exp(\tilde{\eta}),\ \exp(\tilde{\theta}),\ \boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2,\ \exp(\tilde{\mathbf{g}}),\ \exp(\tilde{\mathbf{h}})) = \tilde{\ell}(\tilde{\eta},\ \tilde{\theta},\ \boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2,\ \tilde{\mathbf{g}},\ \tilde{\mathbf{h}})$$

The penalized maximization in the *joint.Cox* R package is performed on $\tilde{\ell}$ rather than $\ell$ since the domains of $\tilde{\ell}$ are unrestricted. The penalized log-likelihood is

$$\tilde{\ell}(\tilde{\eta},\ \tilde{\theta},\ \boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2,\ \tilde{\mathbf{g}},\ \tilde{\mathbf{h}}) - \kappa_1 \int \ddot{r}_0(t)^2\mathrm{d}t - \kappa_2 \int \ddot{\lambda}_0(t)^2\mathrm{d}t \tag{4}$$

where $\ddot{f}(t) = d^2f(t)/dt^2$, and $(\kappa_1, \kappa_2)$ are positive smoothing parameters.

We generate 500 pairs of parameters from a multivariate normal distribution

$$\begin{aligned}&(\log(\hat{\eta}^*),\ \log(\hat{\theta}^*),\ \hat{\boldsymbol{\beta}}_1^*,\ \hat{\boldsymbol{\beta}}_2^*,\ \log(\hat{\mathbf{g}}^*),\ \log(\hat{\mathbf{h}}^*))\\ &\quad\sim N\Big(\text{Mean} = (\log(\hat{\eta}),\ \log(\hat{\theta}),\ \hat{\boldsymbol{\beta}}_1,\ \hat{\boldsymbol{\beta}}_2,\ \log(\hat{\mathbf{g}}),\ \log(\hat{\mathbf{h}})),\ \text{Covariance} = -\hat{\tilde{H}}_{PL}^{-1}(\kappa_1,\ \kappa_2)\Big)\end{aligned}$$

where $\hat{\tilde{H}}_{PL}(\kappa_1,\ \kappa_2)$ is the converged Hessian matrix for the penalized log-likelihood of equation (4). These parameters are used to compute 500 Monte Carlo values $\hat{F}^*(t,\ t+w|\ H(t,x),\ \mathbf{Z})$'s. The 95% confidence interval is obtained using the 2.5% and 97.5% points of $\hat{F}^*(t,\ t+w|\ H(t,x),\ \mathbf{Z})$'s.

# Supplementary Material for: Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model

**Takeshi Emura, Masahiro Nakatochi, Shigeyuki Matsui, Hirofumi Michimae and Virginie Rondeau**

**Corresponding to:** Takeshi Emura, Graduate Institute of Statistics, National Central University, Zhongda Rd., Zhongli District, Taoyuan City 32001, Taiwan
Email: takeshiemura@gmail.com

Supplementary Material contains the following items to supply the main article.

**S1: An example of calculating the prediction probability** $\hat{F}$

**S2: Simulations to assess optimism bias**

**S3: Lists of genes associated with survival (P-value<0.001)**

**S4. Biological functions of genes associated with death and relapse**

**S5. Variable selection**

## S1: An example of calculating the prediction probability $\hat{F}$

A simple example allows one to see how the proposed prediction formula works. We consider parameter estimates $(\hat{\theta}, \hat{\eta}, \hat{\alpha}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{r}_0(\cdot), \hat{\lambda}_0(\cdot))$ defined as $\hat{\theta}=6$, $\hat{\eta}=0.5$, $\hat{\alpha}=1$, $\hat{\boldsymbol{\beta}}_1=\hat{\boldsymbol{\beta}}_2=1$, and $\hat{r}_0(t)=\sum_{\ell=1}^{5}\hat{g}_\ell M_\ell(t)$ and $\hat{\lambda}_0(t)=\sum_{\ell=1}^{5}\hat{h}_\ell M_\ell(t)$, where $\hat{h}_\ell=\hat{g}_\ell=1$ for all $\ell$. Here, $M_\ell(t)$, $\ell=1,\ldots,5$, are the five cubic M-spline bases (Emura et al., 2015) defined on the range $\xi_1\le t\le\xi_3$, where $\xi_1=0$ and $\xi_3=3$. The value $\hat{\theta}=6$ implies strong intra-subject dependence (Kendall's tau = 0.75) between time-to-tumour progression (TTP) and time-to-death, and the value $\hat{\eta}=0.5$ implies moderate amount of heterogeneity between studies. We wish to predict the probability of death by $\hat{F}(t,t+w|H(t,x),\mathbf{Z})$ for a patient with a covariate $\mathbf{Z}=(\mathbf{Z}_1,\mathbf{Z}_2)=(1,1)$ and given his/her progression status at a prediction time $t=1$. We consider two cases for TTP occurring before $t=1$: $X=0.2$ and $X=0.8$. After installing the *joint.Cox* R package (Emura 2016), we run the following commands:

```
> w=c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0,1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,2.0)
> F.windows(time=1,X=0.2,widths=w,Z1=1,Z2=1,beta1=1,beta2=1,eta=0.5,theta=6,
+           alpha=1,g=rep(1,5),h=rep(1,5),xi1=0,xi3=3)
      t   w   X F_event_at_X  F_noevent
 [1,] 1 0.0 0.2    0.0000000 0.00000000
 [2,] 1 0.1 0.2    0.2242580 0.06203191
 [3,] 1 0.2 0.2    0.3842101 0.13066666
 [4,] 1 0.3 0.2    0.5021955 0.19724611
 [5,] 1 0.4 0.2    0.5915647 0.25886806
 [6,] 1 0.5 0.2    0.6606891 0.31492470
 [7,] 1 0.6 0.2    0.7150498 0.36557577
 [8,] 1 0.7 0.2    0.7584282 0.41127527
 [9,] 1 0.8 0.2    0.7934999 0.45261118
[10,] 1 0.9 0.2    0.8226970 0.49019403
[11,] 1 1.0 0.2    0.8480790 0.52460305
[12,] 1 1.1 0.2    0.8696267 0.55636284
[13,] 1 1.2 0.2    0.8880923 0.58593418
[14,] 1 1.3 0.2    0.9040385 0.61371132
[15,] 1 1.4 0.2    0.9178853 0.64002216
[16,] 1 1.5 0.2    0.9299454 0.66512968
[17,] 1 1.6 0.2    0.9404502 0.68923412
[18,] 1 1.7 0.2    0.9495726 0.71247607
[19,] 1 1.8 0.2    0.9574454 0.73494047
[20,] 1 1.9 0.2    0.9641782 0.75666223
[21,] 1 2.0 0.2    0.9698715 0.77763348
> F.windows(time=1,X=0.8,widths=w,Z1=1,Z2=1,beta1=1,beta2=1,eta=0.5,theta=6,
+           alpha=1,g=rep(1,5),h=rep(1,5),xi1=0,xi3=3)
      t   w   X F_event_at_X  F_noevent
 [1,] 1 0.0 0.8    0.0000000 0.00000000
 [2,] 1 0.1 0.8    0.4238257 0.06203191
 [3,] 1 0.2 0.8    0.6378359 0.13066666
 [4,] 1 0.3 0.8    0.7573313 0.19724611
 [5,] 1 0.4 0.8    0.8293982 0.25886806
 [6,] 1 0.5 0.8    0.8754908 0.31492470
 [7,] 1 0.6 0.8    0.9063526 0.36557577
 [8,] 1 0.7 0.8    0.9278059 0.41127527
 [9,] 1 0.8 0.8    0.9432126 0.45261118
[10,] 1 0.9 0.8    0.9545995 0.49019403
[11,] 1 1.0 0.8    0.9632329 0.52460305
[12,] 1 1.1 0.8    0.9699296 0.55636284
[13,] 1 1.2 0.8    0.9752285 0.58593418
[14,] 1 1.3 0.8    0.9794958 0.61371132
[15,] 1 1.4 0.8    0.9826720 0.64002216
[16,] 1 1.5 0.8    0.9856990 0.66512968
[17,] 1 1.6 0.8    0.9882681 0.68923412
[18,] 1 1.7 0.8    0.9904528 0.71247607
[19,] 1 1.8 0.8    0.9923035 0.73494047
[20,] 1 1.9 0.8    0.9938563 0.75666223
[21,] 1 2.0 0.8    0.9951401 0.77763348
```

Above outputs are summarized as follows:

*Predictive probability of death* $\hat{F}(t, t+w \mid H(t,x), (Z_1, Z_2) = (1,1))$

(i) Given that a patient had no tumour progression at time $t = 1$ (i.e., $X > 1$), the probability of death between $t = 1$ and $t + w = 2$ is **0.525**.

(ii) Given that a patient had tumour progression at time $x = 0.2$ (i.e., $X = 0.2$), the probability of death between $t = 1$ and $t + w = 2$ is **0.848**.

(iii) Given that a patient had tumour progression at time $x = 0.8$ (i.e., $X = 0.8$), the probability of death between $t = 1$ and $t + w = 2$ is **0.963**.

The outputs are also presented in Figure S1. The figure shows that the prior occurrence of tumour progression remarkably increases the predictive probability of death. The TTP of $X = 0.8$ yields higher predictive probability of death than the TTP of $X = 0.2$. This is the consequence of the strong intra-subject dependence between TTP and death (Kendall's tau = 0.75).

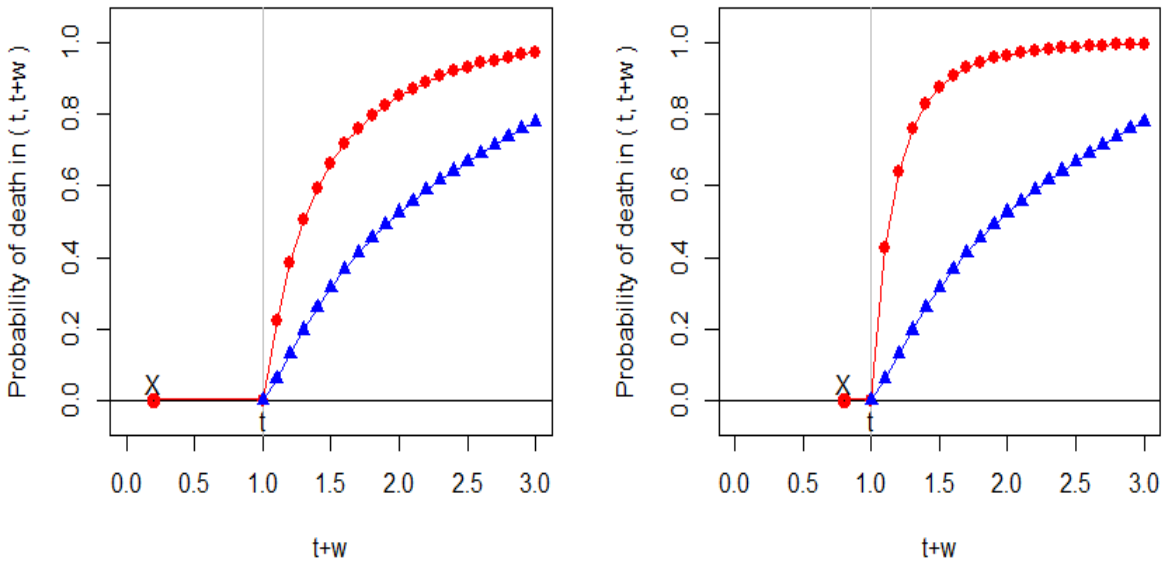


**Figure S1.** Predictive probabilities of death between $t = 1$ and $t + w \in [1, 3]$. The blue symbol (▲) signifies a patient who did not experience tumour progression before $t = 1$. The red symbol (●) signifies a patient who experienced tumour progression at $X = 0.2$ (left panel) and a patient who experienced tumour progression at time $X = 0.8$ (right panel).

**Ranges of prediction**: The two baseline hazards $r_0(t)$ and $\lambda_0(t)$ are identifiable in a range $\xi_1 \le t \le \xi_3$, where $\xi_1 = \min_{i,j}(T_{ij})$ is the smallest even time and $\xi_3 = \max_{i,j}(T_{ij}^*)$ is the maximal follow-up time, as computed by the *joint.Cox* R package. Accordingly, the probability $\hat{F}(t, t+w \mid X = x, \mathbf{Z})$ can be defined only within the range $\xi_1 \le x \le t < t + w \le \xi_3$. Some warning messages will be produced if the inputted values are beyond the range.

**S2: Simulations to assess optimism bias**

Simulations were conducted to evaluate the degree of the optimism bias for estimating prediction error of the proposed prediction formula. We also compare two different methods of summarizing high-dimensional genetic factors: 1) compound covariate predictor, and 2) ridge-based predictor.

Let $G=5$ be the number of studies and $N_i=200$ be the number of subjects in each study for $i=1,2,...,5$. The frailty $u_i$ followed a gamma distribution with variance $\eta$, and a clinical covariate $Z_{ij}$ followed the standard normal distribution $N(0,1)$, truncated between -3 and 3. We generated a vector of genomic covariates $\mathbf{U}_{ij}=(U_{ij,1},\ldots,U_{ij,q})$ from a multivariate uniform distribution with all the margins having mean$=0$ and SD$=1$, where $q=200$ is the number of genes. The corresponding coefficients are

$$\boldsymbol{\xi}'=(\underbrace{0.03,...,0.03}_{\times 15},\underbrace{-0.03,...,-0.03}_{\times 15},\underbrace{0,...,0}_{\times 170}).$$

Here, we have assumed the existence of the two blocks of correlated genes (i.e., pathways): the first corresponds to the 15 positive coefficients, and the second corresponds to the 15 negative coefficients. Specifically, $Corr(U_{ij,k},U_{ij,\ell})=0.5$ for $1\le k<l\le 15$ or $16\le k<l\le 30$; $Corr(U_{ij,k},U_{ij,\ell})=0$ otherwise. We generated such gene expressions by *X.pathway* routine in the R *compound.Cox* package (Emura et al., 2016). This correlation structure mimics the setting of gene pathways, where the two sets of genes informative for survival are correlated (see Binder et al., 2009; Emura et al., 2012).

Given $u_i$, $Z_{ij}$, and $\mathbf{U}_{ij}$, the distribution of $X_{ij}$ and $D_{ij}$ followed the joint frailty-copula model (Emura et al., 2015)

$$\begin{cases} r_{ij}(t\,|\,u_i)=u_i r_0(t)\exp(\beta_1 Z_{ij}+\boldsymbol{\xi}'\mathbf{U}_{ij}) & (\text{ for } X_{ij}) \\ \lambda_{ij}(t\,|\,u_i)=u_i\lambda_0(t)\exp(\beta_2 Z_{ij}+\boldsymbol{\xi}'\mathbf{U}_{ij}) & (\text{ for } D_{ij}) \\ \Pr(X_{ij}>x, D_{ij}>y\,|\,u_i)=[\exp\{\theta R_{ij}(x\,|\,u_i)\}+\exp\{\theta\Lambda_{ij}(y\,|\,u_i)\}-1]^{-1/\theta} \end{cases}$$

where $\lambda_0(t)=r_0(t)=1$, $R_{ij}(x\,|\,u_i)=\int_0^x r_{ij}(t\,|\,u_i)dt$ and $\Lambda_{ij}(y\,|\,u_i)=\int_0^y \lambda_{ij}(t\,|\,u_i)dt$. The amount of intra-subject dependence (between $X_{ij}$ and $D_{ij}$) is determined by the association parameter $\theta$ while the amount of intra-study dependence is determined by the frailty variance $\eta$. We set

$\theta = 6$ (Kendall's tau = 0.75) and $\eta = 0.5$ throughout the simulations. Censoring variable $C_{ij}$ followed a uniform distribution on $(0, 5)$ that yielded about 30% censored subjects.

For a dataset generated from the aforementioned models, we summarize the high-dimensional genetic factors in two different ways:

**Method (1) Compound covariate**

We form two compound covariate (CC) predictors

$$\mathrm{CC}_{1,ij} = \hat{b}_1 V_{ij,1} + \cdots + \hat{b}_{q_1} V_{ij,q_1} \qquad (\text{associated with tumour progression } X_{ij})$$
$$\mathrm{CC}_{2,ij} = \hat{c}_1 W_{ij,1} + \cdots + \hat{c}_{q_2} W_{ij,q_2} \qquad (\text{associated with death } D_{ij})$$

where the weights $\hat{b}_k$ and $\hat{c}_k$ are estimates of regression coefficients under univariate Cox models on *k*-th gene, $r_{ij}(t) = r_0(t)\exp(b_k V_{ij,k})$, and $\lambda_{ij}(t) = \lambda_0(t)\exp(c_k W_{ij,k})$, respectively. We determined the number of genes $q_1$ in $\mathrm{CC}_{1,ij}$ by thresholding P-value<0.2 of testing the null hypothesis $H_0 : b_k = 0$ in the univariate Cox model. The number $q_2$ in $\mathrm{CC}_{2,ij}$ is determined similarly. This implies that two subsets $\mathbf{V}_{ij} = (V_{ij,1}, \ldots, V_{ij,q_1})$ and $\mathbf{W}_{ij} = (W_{ij,1}, \ldots, W_{ij,q_2})$ of $\mathbf{U}_{ij} = (U_{ij,1}, \ldots, U_{ij,q})$ are used for prediction.

**Method (2) Multivariate Ridge regression ($L_2$-penalized Cox regression)**

We form two ridge-based predictors

$$\mathrm{Ridge}_{1,ij} = \hat{\xi}_1 U_{ij,1} + \cdots + \hat{\xi}_q U_{ij,q} = \hat{\boldsymbol{\xi}}'\mathbf{U}_{\mathrm{ij}} \qquad (\text{associated with tumour progression } X_{ij})$$
$$\mathrm{Ridge}_{2,ij} = \hat{\varsigma}_1 U_{ij,1} + \cdots + \hat{\varsigma}_q U_{ij,q} = \hat{\boldsymbol{\varsigma}}'\mathbf{U}_{\mathrm{ij}} \qquad (\text{associated with death } D_{ij})$$

where the weights $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\varsigma}}$ are the ridge estimates (Verweij and van Houwelingen, 1994) of regression coefficients under multivariate Cox models $r_{ij}(t) = r_0(t)\exp(\boldsymbol{\xi}'\mathbf{U}_{ij})$, and $\lambda_{ij}(t) = \lambda_0(t)\exp(\boldsymbol{\varsigma}'\mathbf{U}_{ij})$, respectively. We apply the R command *optL2(,fold=5)* in the package *penalized* (Goeman et al. 2016) to obtain the ridge estimates, where the shrinkage parameter is optimized by 5-fold cross-validation.

Method (1) uses a pre-selection of genes while Method (2) uses all available genes of $p = 200$. In Method (1), every simulation run results in a different subset of genes. For instance, on average, the number $q_1 \approx 60$ is included in the formula of $\mathrm{CC}_{1,ij}$. In Method (2), every simulation run results in a different amount of shrinkage parameter.

The CC predictors [or ridge-based predictors] are treated as new covariates and incorporated into the joint frailty-copula model as

$$\begin{cases} r_{ij}(t \mid u_i) = u_i r_0(t) \exp(\beta_1 Z_{ij} + \gamma_1 \mathrm{CC}_{1,ij}) & (\text{for } X_{ij}) \\ \lambda_{ij}(t \mid u_i) = u_i \lambda_0(t) \exp(\beta_2 Z_{ij} + \gamma_2 \mathrm{CC}_{2,ij}) & (\text{for } D_{ij}) \\ \Pr(X_{ij} > x, D_{ij} > y \mid u_i) = [\exp\{\theta R_{ij}(x \mid u_i)\} + \exp\{\theta \Lambda_{ij}(y \mid u_i)\} - 1]^{-1/\theta} \end{cases}$$

We obtained parameter estimates and dynamic prediction formula by using the *joint.Cox* R package. The ridge-based predictors are incorporated in a similar way.

Based on the fitted results of the joint models, we calculated estimates of prediction error, denoted as $\hat{E}rr(t, t+w)$ that is defined in the main article. We also fitted the joint frailty model under an assumed value of $\theta = 0$, and calculated estimates of prediction error. These estimated prediction errors are compared with the benchmark value $\hat{E}rr^{KM}(t, t+w)$ defined in the main article. We did not use cross-validation since our objective was to see the degree of the optimism bias.

To evaluate the optimism bias, we compared $\hat{E}rr(t, t+w)$ with the true prediction error $Err(t, t+w)$. To this end, we generated independent test data $(X_{ij}, D_{ij}, Z_{ij}, \mathbf{U}_{ij})$ in the same algorithms described before. The true prediction error is then approximated as

$$Err(t, t+w) \approx \frac{\sum_{ij} \mathbf{I}(D_{ij} > t)\{\mathbf{I}(D_{ij} > t+w) - \hat{S}(t, t+w \mid H(t, X_{ij}), \mathbf{Z}_{ij}, \mathbf{U}_{ij})\}^2}{\sum_{ij} \mathbf{I}(D_{ij} > t)},$$

where the summation is taken over all subjects in the test data $i = 1, 2, ..., 5$ and $j = 1, 2, ..., 200$. Estimates in the formula of $\hat{S}$ do not use the test data. The true prediction error under the null model, $Err^{KM}(t, t+w)$, is approximated similarly.

The estimated (and true) prediction errors are reported based on the average of 50 Monte Carlo replications at time $t = 0.5$ or 1.5 and range $0 \le w \le 3$. Results are summarized in Figure S2.

Left panels of Figure S2 display the *true* prediction errors between different prediction models. All types of the joint models had smaller prediction error than the null model (the Kaplan-Meier estimator), indicating the advantage of using the joint models in prediction. Among the joint models, the smallest prediction error was achieved by the model that utilizes both clinical covariates and CCs. If the CCs are replaced by the ridge-based predictors, the prediction error shows a small increase. If the model ignores the CCs (utilizes only clinical covariates), the prediction error clearly inflates. Also, if the model ignores the intra-subject dependence, the prediction error inflates. The model using only clinical covariates (ignoring both the CCs and intra-subject dependence) performs worse among all the joint models, but it still shows a clear advantage over the null model.

Right panels of Figure S2 display the *estimated* prediction errors between different prediction models. The estimated prediction errors showed very similar patterns with the true prediction errors. That is, if the model ignores either the genetic factors or the intra-subject dependence, the prediction error increases. Unlike the case of the true prediction error, the best prediction scheme was now achieved by the joint model that utilizes both clinical covariates and ridge-based predictors. This phenomenon may be explained as the over-fitting of the ridge-based predictors that use all the 200 genes for prediction; the ridge-based predictors may work better on the training data, but not as much on the testing data.

Comparing between the true and estimated prediction errors, it is clear that the estimated prediction errors showed downward biases relative to the true prediction errors. Nevertheless, the estimated prediction errors contained sufficient information to determine the relative performance between different prediction models. One exception is the joint model that uses the ridge-based predictor that yielded the over-fitting phenomenon.

In conclusion, we have confirmed that the estimated prediction error $\hat{E}rr(t, t+w)$ exhibits modest downward bias, but is still a good substitute for the true prediction error $Err(t, t+w)$. This is an important advantage of using the CC to avoid extreme overfitting.

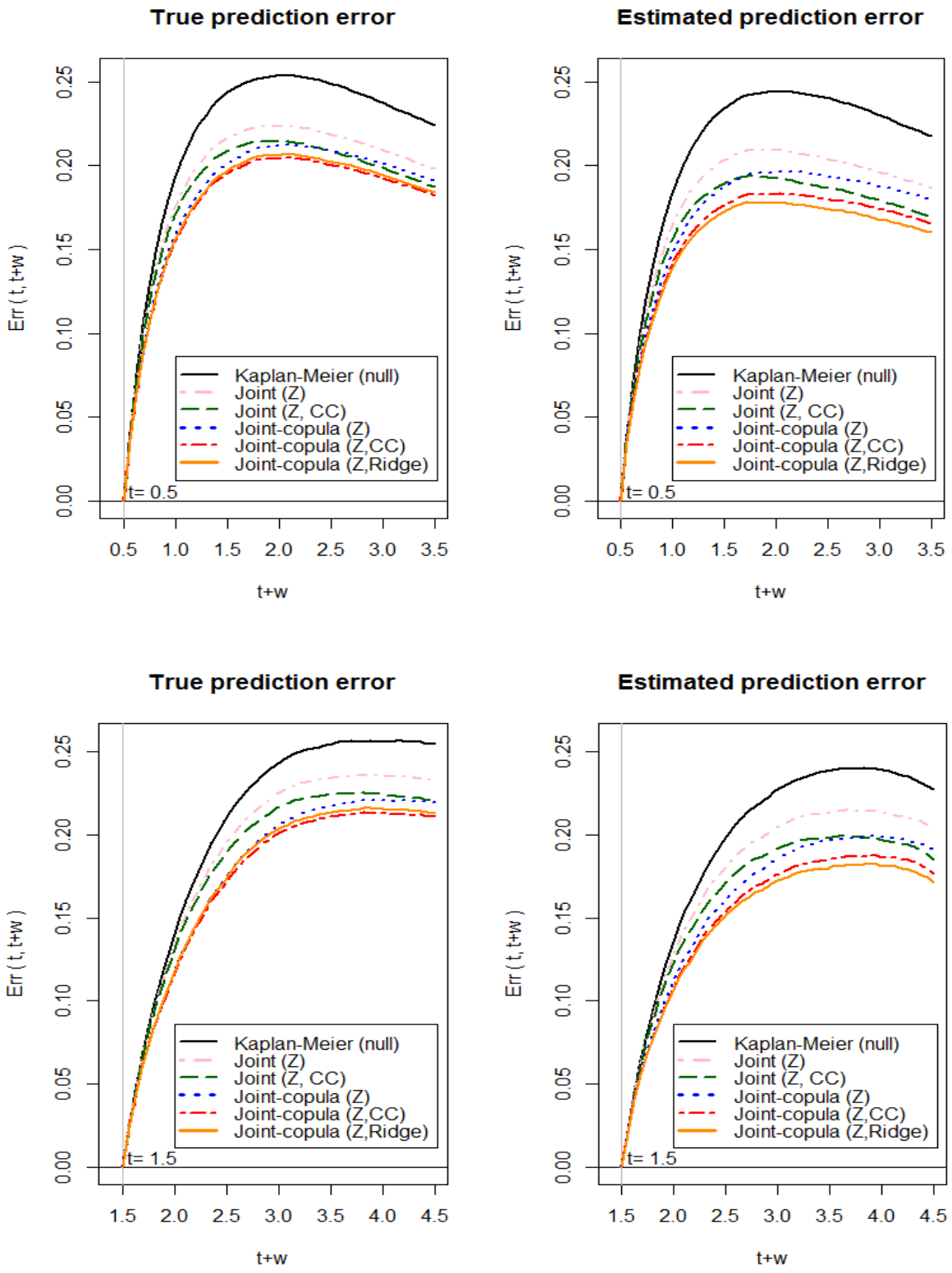


**Figure S2.** The true and estimated prediction errors under the joint models with clinical covariate (Z) and genetic factors (CC or Ridge). The Clayton copula (Joint-copula) or independence copula (Joint) is used for intra-subject dependence.

**S3. Lists of genes associated with survival (P-value<0.001)**

We present the results of performing gene selection based on univariate Cox regression analyses. In the univariate Cox model, we treat time-to-relapse (or time-to-death) as a response and a single gene expression as a covariate. The results of testing the null hypothesis of no covariate effect are used to select and order genes.

**158 genes associated with time-to-relapse (P-value<0.001)**

| Gene | P-value | coefficient |
|---|---|---|
| CXCL12 | 1.36E-08 | 0.249 |
| TIMP2 | 1.03E-07 | 0.235 |
| PDPN | 1.68E-07 | 0.222 |
| TUBB6 | 3.28E-07 | 0.228 |
| ANKRD27 | 1.91E-06 | 0.198 |
| COL3A1 | 2.52E-06 | 0.21 |
| CRYAB | 2.80E-06 | 0.204 |
| FOSL2 | 2.84E-06 | 0.206 |
| DNAJC8 | 2.98E-06 | 0.2 |
| TEAD1 | 3.13E-06 | 0.195 |
| SPARC | 6.69E-06 | 0.197 |
| RARRES1 | 6.77E-06 | 0.197 |
| CLIC4 | 6.87E-06 | 0.207 |
| TIMP3 | 6.97E-06 | 0.191 |
| PCYT1A | 7.77E-06 | 0.183 |
| ITGB1 | 8.02E-06 | 0.223 |
| NCOA3 | 8.18E-06 | 0.194 |
| FAP | 8.26E-06 | 0.191 |
| ASAP3 | 8.41E-06 | 0.182 |
| FGF1 | 8.47E-06 | 0.181 |
| THEMIS2 | 9.58E-06 | 0.192 |
| COL11A1 | 1.13E-05 | 0.187 |
| LOX | 1.35E-05 | 0.188 |
| PDGFD | 1.48E-05 | 0.183 |
| KLHL25 | 1.59E-05 | -0.191 |
| CRISPLD2 | 1.74E-05 | 0.183 |
| NUAK1 | 1.96E-05 | 0.185 |
| MEOX2 | 2.24E-05 | 0.166 |

| | | |
|---|---|---|
| HP1BP3 | 2.52E-05 | 0.172 |
| GPATCH1 | 2.54E-05 | 0.176 |
| INHBA | 2.69E-05 | 0.179 |
| SERPINE1 | 2.93E-05 | 0.173 |
| LGALS1 | 3.33E-05 | 0.189 |
| FERMT2 | 3.79E-05 | 0.188 |
| HSD17B6 | 4.08E-05 | 0.172 |
| KIAA1598 | 4.47E-05 | 0.185 |
| COL5A1 | 4.60E-05 | 0.175 |
| FABP4 | 4.62E-05 | 0.169 |
| VSIG4 | 4.91E-05 | 0.179 |
| ZFP36 | 4.99E-05 | 0.17 |
| ZFP36L2 | 5.27E-05 | 0.179 |
| COMP | 5.39E-05 | 0.161 |
| POSTN | 5.69E-05 | 0.175 |
| CYR61 | 5.73E-05 | 0.176 |
| C1QTNF3 | 6.09E-05 | 0.167 |
| CDV3 | 6.10E-05 | 0.174 |
| GAS1 | 6.16E-05 | 0.183 |
| NPY | 6.61E-05 | 0.154 |
| PDE1A | 7.01E-05 | 0.156 |
| N4BP2L2 | 7.31E-05 | 0.17 |
| COL10A1 | 7.35E-05 | 0.168 |
| B4GALT5 | 7.39E-05 | 0.17 |
| DDX27 | 7.86E-05 | 0.174 |
| CTSK | 8.55E-05 | 0.165 |
| VCAN | 8.68E-05 | 0.173 |
| COL5A2 | 8.73E-05 | 0.171 |
| DVL3 | 8.90E-05 | 0.168 |
| TAGLN | 9.03E-05 | 0.175 |
| CCNL1 | 9.64E-05 | 0.164 |
| DPYSL3 | 9.96E-05 | 0.18 |
| RPS16 | 1.06E-04 | 0.203 |
| ADAM12 | 1.09E-04 | 0.16 |
| SLC12A8 | 1.13E-04 | 0.168 |
| SH3PXD2A | 1.14E-04 | 0.157 |
| GJC1 | 1.15E-04 | 0.16 |
| ZNF148 | 1.18E-04 | 0.163 |

| | | |
|---|---|---|
| TPM4 | 1.19E-04 | 0.173 |
| USP48 | 1.27E-04 | 0.161 |
| FN1 | 1.29E-04 | 0.174 |
| JUN | 1.35E-04 | 0.165 |
| CEBPB | 1.44E-04 | 0.177 |
| DNAJC13 | 1.45E-04 | 0.16 |
| LUZP1 | 1.57E-04 | 0.166 |
| PLSCR4 | 1.59E-04 | 0.155 |
| TGM5 | 1.82E-04 | -0.169 |
| HLTF | 1.83E-04 | 0.165 |
| ADORA3 | 1.87E-04 | 0.16 |
| EIF3K | 1.87E-04 | 0.168 |
| DNAJB4 | 1.91E-04 | 0.161 |
| SULF1 | 1.98E-04 | 0.163 |
| TESK1 | 1.98E-04 | -0.163 |
| TUBB2A | 2.06E-04 | 0.157 |
| LUM | 2.10E-04 | 0.165 |
| KIN | 2.14E-04 | 0.166 |
| CALD1 | 2.20E-04 | 0.168 |
| STAU1 | 2.23E-04 | 0.17 |
| FAM69A | 2.28E-04 | 0.155 |
| EPYC | 2.32E-04 | 0.151 |
| PPIC | 2.35E-04 | 0.161 |
| COL16A1 | 2.47E-04 | 0.161 |
| NOTCH2 | 2.59E-04 | 0.16 |
| PSMC4 | 2.64E-04 | 0.157 |
| ENPP1 | 2.79E-04 | 0.163 |
| TPM2 | 2.87E-04 | 0.156 |
| ARHGAP28 | 2.95E-04 | 0.175 |
| SGK1 | 3.00E-04 | 0.158 |
| CSE1L | 3.04E-04 | 0.17 |
| OAT | 3.16E-04 | 0.162 |
| MXD1 | 3.32E-04 | 0.159 |
| L2HGDH | 3.33E-04 | -0.158 |
| ARHGAP29 | 3.35E-04 | 0.167 |
| DCUN1D1 | 3.39E-04 | 0.156 |
| KRT7 | 3.40E-04 | 0.166 |
| PLAU | 3.59E-04 | 0.159 |

| | | |
|---|---|---|
| AP3S1 | 3.61E-04 | 0.158 |
| RAB32 | 3.74E-04 | 0.151 |
| KPNA6 | 3.97E-04 | 0.146 |
| MFN1 | 3.98E-04 | 0.16 |
| MCL1 | 4.00E-04 | 0.152 |
| GFRA1 | 4.11E-04 | 0.109 |
| KIAA0355 | 4.18E-04 | 0.155 |
| PGRMC1 | 4.18E-04 | -0.155 |
| KIAA0226 | 4.39E-04 | 0.147 |
| SPHK1 | 4.40E-04 | 0.149 |
| ELK1 | 4.64E-04 | -0.161 |
| METTL9 | 4.64E-04 | -0.148 |
| MAPRE1 | 4.72E-04 | 0.156 |
| MRPS22 | 4.77E-04 | 0.156 |
| MICAL2 | 4.83E-04 | 0.159 |
| OLFML2B | 4.87E-04 | 0.15 |
| PRDM2 | 5.07E-04 | 0.152 |
| RAB31 | 5.26E-04 | 0.155 |
| ARTN | 5.30E-04 | -0.157 |
| NNMT | 5.39E-04 | 0.156 |
| GFRA3 | 5.78E-04 | -0.194 |
| CDC42 | 5.84E-04 | 0.154 |
| ABI3BP | 6.07E-04 | 0.126 |
| DIAPH3 | 6.08E-04 | 0.144 |
| SUPT5H | 6.10E-04 | 0.143 |
| RAB22A | 6.49E-04 | 0.142 |
| PLOD2 | 6.61E-04 | 0.151 |
| GLIPR1 | 6.61E-04 | 0.153 |
| URI1 | 6.72E-04 | 0.139 |
| TP73-AS1 | 6.78E-04 | 0.144 |
| GABRG3 | 6.83E-04 | -0.156 |
| TJP1 | 6.88E-04 | 0.156 |
| LPP | 6.90E-04 | 0.147 |
| KRTAP5-8 | 7.05E-04 | -0.154 |
| YWHAB | 7.05E-04 | 0.169 |
| MXRA8 | 7.36E-04 | 0.145 |
| EFNB2 | 7.66E-04 | 0.145 |
| NDRG3 | 7.82E-04 | 0.138 |

| | | |
|---|---|---|
| NINJ1 | 7.82E-04 | -0.146 |
| TSC22D2 | 7.88E-04 | 0.134 |
| TUFT1 | 7.88E-04 | 0.147 |
| FSTL1 | 8.10E-04 | 0.148 |
| AP2M1 | 8.25E-04 | 0.145 |
| BCAP31 | 8.31E-04 | -0.147 |
| SKIL | 8.34E-04 | 0.144 |
| ZMYM1 | 8.47E-04 | 0.14 |
| NTM | 8.69E-04 | 0.138 |
| CCNE1 | 8.85E-04 | 0.145 |
| MAP7D1 | 8.91E-04 | 0.144 |
| TBCB | 8.91E-04 | 0.143 |
| ZNF79 | 9.31E-04 | -0.153 |
| PARD3 | 9.50E-04 | 0.14 |
| BRD4 | 9.61E-04 | 0.14 |
| MMP12 | 9.77E-04 | -0.152 |

**128 genes associated with time-to-death (P-value<0.001)**

| Gene | P-value | coefficient |
|---|---|---|
| **NCOA3** | **6.88E-07** | **0.237** |
| **TEAD1** | **1.32E-06** | **0.223** |
| **YWHAB** | **1.38E-06** | **0.263** |
| PSMC4 | 1.77E-06 | 0.214 |
| PDP1 | 3.26E-06 | 0.226 |
| TUBB6 | 3.37E-06 | 0.228 |
| STAU1 | 4.67E-06 | 0.234 |
| GPATCH1 | 4.72E-06 | 0.202 |
| RPS16 | 4.91E-06 | 0.258 |
| B4GALT5 | 6.62E-06 | 0.215 |
| ASAP3 | 1.02E-05 | 0.199 |
| HP1BP3 | 1.47E-05 | 0.197 |
| DDX27 | 1.61E-05 | 0.211 |
| NUAK1 | 1.69E-05 | 0.199 |
| ENPP1 | 2.00E-05 | 0.197 |
| KIAA0355 | 2.53E-05 | 0.194 |
| COL16A1 | 2.56E-05 | 0.197 |
| SH3PXD2A | 2.61E-05 | 0.181 |

| | | |
|---|---|---|
| CRYAB | 2.79E-05 | 0.195 |
| RECQL | 3.15E-05 | 0.189 |
| DLGAP4 | 4.46E-05 | 0.183 |
| CPNE1 | 4.84E-05 | 0.193 |
| FABP4 | 5.02E-05 | 0.176 |
| N4BP2L2 | 5.50E-05 | 0.191 |
| NCOA6 | 5.53E-05 | 0.196 |
| LSM14A | 6.54E-05 | 0.193 |
| TBCB | 6.66E-05 | 0.193 |
| COL5A1 | 8.44E-05 | 0.184 |
| ARHGAP28 | 8.59E-05 | 0.201 |
| LEP | 8.80E-05 | 0.168 |
| LPL | 9.22E-05 | 0.177 |
| INHBA | 1.02E-04 | 0.177 |
| PCDH9 | 1.02E-04 | 0.158 |
| FSTL1 | 1.05E-04 | 0.192 |
| BYSL | 1.05E-04 | -0.192 |
| RND3 | 1.10E-04 | 0.189 |
| MAPRE1 | 1.12E-04 | 0.186 |
| AP3S1 | 1.17E-04 | 0.191 |
| PLXNA1 | 1.20E-04 | 0.188 |
| ZFP36 | 1.23E-04 | 0.173 |
| HLA-DOB | 1.27E-04 | -0.193 |
| URI1 | 1.28E-04 | 0.17 |
| COMP | 1.31E-04 | 0.162 |
| OAT | 1.33E-04 | 0.189 |
| APMAP | 1.33E-04 | 0.186 |
| GAS1 | 1.33E-04 | 0.194 |
| IL2RG | 1.39E-04 | -0.189 |
| NOTCH2NL | 1.47E-04 | 0.183 |
| CXCL12 | 1.48E-04 | 0.183 |
| LUZP1 | 1.50E-04 | 0.18 |
| PSMD8 | 1.60E-04 | 0.172 |
| ZNF148 | 1.64E-04 | 0.177 |
| ANKRD27 | 1.66E-04 | 0.166 |
| TIMP3 | 1.72E-04 | 0.174 |
| CLIC4 | 1.79E-04 | 0.19 |
| PAK4 | 1.85E-04 | 0.163 |

| | | |
|---|---|---|
| RAI14 | 1.86E-04 | 0.177 |
| COL3A1 | 1.96E-04 | 0.183 |
| CYTH3 | 2.02E-04 | 0.169 |
| COL11A1 | 2.33E-04 | 0.167 |
| FAP | 2.37E-04 | 0.174 |
| TJP1 | 2.69E-04 | 0.185 |
| RAB13 | 2.69E-04 | 0.174 |
| KDELC1 | 2.79E-04 | 0.165 |
| JUN | 2.84E-04 | 0.172 |
| CTNNBL1 | 2.89E-04 | 0.164 |
| TSPAN9 | 2.91E-04 | 0.178 |
| EIF3K | 2.96E-04 | 0.176 |
| RARRES1 | 2.98E-04 | 0.174 |
| SLAMF7 | 2.98E-04 | -0.187 |
| SACS | 2.99E-04 | 0.16 |
| ZFP36L2 | 3.04E-04 | 0.175 |
| LOX | 3.18E-04 | 0.173 |
| ITGB1 | 3.21E-04 | 0.197 |
| PHF20 | 3.37E-04 | 0.168 |
| CASP8 | 3.43E-04 | -0.166 |
| CRISPLD2 | 3.43E-04 | 0.165 |
| KIN | 3.48E-04 | 0.175 |
| MMP12 | 3.64E-04 | -0.187 |
| RIN2 | 3.79E-04 | 0.178 |
| EMP1 | 3.88E-04 | 0.174 |
| TUBB2A | 3.92E-04 | 0.164 |
| PDPN | 3.93E-04 | 0.161 |
| CD79A | 4.05E-04 | -0.18 |
| FGF1 | 4.32E-04 | 0.157 |
| C1QTNF3 | 4.38E-04 | 0.156 |
| SUPT5H | 4.39E-04 | 0.155 |
| MEOX2 | 4.52E-04 | 0.149 |
| EFNB2 | 4.56E-04 | 0.169 |
| JAM2 | 4.71E-04 | 0.16 |
| SPARC | 4.73E-04 | 0.166 |
| SMG5 | 4.85E-04 | 0.166 |
| COL5A2 | 4.90E-04 | 0.166 |
| TTI1 | 4.91E-04 | 0.16 |

| | | |
|---|---|---|
| SLC37A4 | 4.98E-04 | -0.153 |
| CYBRD1 | 5.31E-04 | 0.164 |
| GABRG3 | 5.36E-04 | -0.182 |
| SOCS5 | 5.38E-04 | 0.167 |
| TP53BP2 | 5.49E-04 | 0.162 |
| GFRA1 | 5.50E-04 | 0.126 |
| HSD17B6 | 5.70E-04 | 0.159 |
| USP48 | 5.71E-04 | 0.16 |
| ITPKC | 5.87E-04 | 0.149 |
| RBM39 | 6.03E-04 | 0.171 |
| HOXA5 | 6.09E-04 | 0.158 |
| TBCC | 6.43E-04 | -0.165 |
| CYR61 | 6.65E-04 | 0.161 |
| OMD | 6.66E-04 | 0.146 |
| MCL1 | 6.85E-04 | 0.161 |
| CXCL9 | 6.94E-04 | -0.163 |
| SSR4 | 7.00E-04 | -0.159 |
| GJC1 | 7.27E-04 | 0.156 |
| LUM | 7.45E-04 | 0.166 |
| COX7A2P2 | 7.78E-04 | -0.163 |
| DYNLRB1 | 7.90E-04 | 0.164 |
| NR1H3 | 8.13E-04 | -0.158 |
| SKI | 8.15E-04 | 0.148 |
| ASAP1 | 8.25E-04 | 0.153 |
| DNAJC13 | 8.59E-04 | 0.157 |
| TESK1 | 8.73E-04 | -0.161 |
| ASB7 | 8.81E-04 | -0.159 |
| CCL18 | 9.10E-04 | -0.181 |
| FBL | 9.21E-04 | 0.161 |
| CDK19 | 9.23E-04 | 0.149 |
| GZMB | 9.35E-04 | -0.166 |
| FOXN3 | 9.47E-04 | 0.156 |
| ELN | 9.69E-04 | 0.141 |
| KCNH4 | 9.83E-04 | -0.157 |

**S4. Biological functions of genes associated with death and relapse**

The lists of genes associated with survival in Section S3 give the compound covariates

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

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

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

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

Below, we detail known biological functions of selected genes (*TIMP2*, *PDPN*, *NCOA3*, *TEAD1*, *YWHAB*) involved in the expressions of $CC_{1,ij}$ and $CC_{2,ij}$.

- ***TIMP2***

*TIMP2* is a member of the TIMP gene family. The proteins encoded by this gene family are natural inhibitors of the matrix metalloproteinases (MMPs). MMPs and their inhibitors (TIMP gene family) play an important regulatory role in the homeostasis of the extracellular matrix (Halon et al., 2012). In addition to inhibitors of MMPs, *TIMP2* has additional functions that are associated with cell proliferation and survival (Bourboulia et al., 2011). In our study, the overexpression of the gene was highly associated with time-to-relapse (Coefficient=0.235, P-value=$1.03\times10^{-7}$).

- ***PDPN***

The *PDPN* gene encodes the podoplanin protein. It is reported that cancer cells with higher *PDPN* expression have higher malignant potential due to enhanced platelet aggregation, which promotes alteration of metastasis, cell motility, and epithelial-mesenchymal transition (Shindo et al., 2013). Zhang et al. (2011) reported that overexpression of *PDPN* in fibroblasts is significantly associated with a poor prognosis in ovarian carcinoma. In our study, the overexpression of the gene was highly associated with time-to-relapse (Coefficient=0.222, P-value=$1.68\times10^{-7}$) and time-to-death (Coefficient=0.161, P-value=$3.93\times10^{-4}$).

- ***NCOA3***

The *NCOA3* gene encodes a nuclear receptor coactivator, and amplification of the gene occurs in breast and ovarian cancers (Anzick et al., 1997). The overexpression of *NCOA3* is associated with tumor size (Spears et al., 2012) and tamoxifen resistance (Osborne et al., 2003), which are involved in the progression. Yoshida et al. (2005) reported that *NCOA3* could contribute to ovarian cancer progression by promoting cell migration. In our study, the overexpression of the gene was highly associated with time-to-relapse (Coefficient=0.194, P-value= $8.18\times10^{-6}$) and time-to-death (Coefficient=0.237, P-value= $6.88\times10^{-7}$). This result is consistent with the function of these reports.

- ***TEAD1***

*TEAD1* encodes a ubiquitous transcriptional enhancer factor that is a member of the TEA/ATTS domain family. It is reported that the protein level of *TEAD1* was associated with poor prognosis in prostate cancer patients (Knight et al., 2008). In our study, the overexpression of the gene was highly associated with time-to-relapse (Coefficient=0.195, P-value= $3.13\times10^{-6}$) and time-to-death (Coefficient=0.223, P-value= $1.32\times10^{-6}$).

- ***YWHAB***

*YWHAB* encodes a protein belonging to the 14-3-3 family of proteins, members of which mediate signal transduction by binding to phosphoserine-containing proteins. It is reported that the protein of *YWHAB* can regulate cell survival, proliferation, and motility (Tzivion, 2006). Actually, it is reported that overexpression of this gene promotes tumor progression and was associated with extrahepatic metastasis and worse survival in hepatocellular carcinoma (Liu et al., 2011). In our study, the overexpression of the gene was highly associated with time-to-relapse (Coefficient=0.169, P-value= $7.05\times10^{-4}$) and time-to-death (Coefficient=0.263, P-value= $1.38\times10^{-6}$).

**S5. Variable selection**

Table S1 presents the results of forward variable selection. Model selection is guided with the likelihood cross-validation (LCV) criterion. The LCV accounts for the number of parameters in the model with penalized likelihood approaches via equation LCV=logL-DF, where logL is the log-likelihood value and DF is the (effective) degree of freedom. More details can be seen from the source codes of the *joint.Cox* R package (Emura 2016). The larger LCV value corresponds to the better model.

We start from the simplest model including only *CXCL12* expression and ignoring clinical covariates (Model 1, Table S1), which was used in Emura et al. (2015). The estimates for the relative risks of *CXCL12* are comparable to previously reported results (Ganzfried et al. 2013; Emura et al. 2015).

By replacing *CXCL12* with the CCs, the LCV criterion improved remarkably (Model 2, Table S2). The estimates for the relative risks are greater than those based on *CXCL12* alone. Hence, we keep the CCs to the model rather than *CXCL12* alone.

The addition of the clinical covariates $Z_1 = Z_2$ (=0 vs. =1) on the residual tumour size at surgery (<1cm vs. ≥ 1cm) further improved the LCV criterion (Model 4, Table 1). Furthermore, their addition almost did not alter the relative risks for the genomic factors (Model 2 vs. Model 4). This implies that the residual tumour size and CCs are independent predictors of survival. However, we observe that the lower confidence bound of the relative risk for $Z_1$ reached the null value of 1.

Therefore, we considered the model dropping $Z_1$ for relapse but still keeping $Z_2$ for death (Model 3 in Table 1). With this model, the LCV criterion was maximized among all the models (Table 1).

Inclusion of FIGO stage to all the models (Models 1-4) did not improve the LCV criterion and the corresponding regression coefficients were nonsignificant (not shown). Hence, FIGO stage is not included in the model.

Thus, we conclude that the most satisfactory model is Model 3 in Table S1, namely

$$\begin{cases} r_{ij}(t\,|\,u_i) = u_i r_0(t)\exp(\gamma_1 \mathrm{CC}_{1,ij}) & (\text{for } X_{ij}) \\ \lambda_{ij}(t\,|\,u_i) = u_i^{\alpha}\lambda_0(t)\exp(\beta_2 Z_{2,ij} + \gamma_2 \mathrm{CC}_{2,ij}) & (\text{for } D_{ij}) \\ \Pr(X_{ij} > x, D_{ij} > y\,|\,u_i) = [\exp\{\theta R_{ij}(x\,|\,u_i)\} + \exp\{\theta \Lambda_{ij}(y\,|\,u_i)\} - 1]^{-1/\theta} \end{cases}$$

**Table S1.** The joint analysis of time-to-relapse and time-to-death for ovarian cancer patients based on the meta-analytic data (four studies, 912 patients) of Ganzfried et al. (2013).

| | **Model 1**: | **Model 2**: | **Model 3 (chosen)**: | **Model 4**: |
|---|---|---|---|---|
| | $CC_1 = CXCL12$<br>$CC_2 = CXCL12$ | $CC_1$=158 genes<br>$CC_2$=128 genes | $Z_2$=Res. tumour<br>$CC_1$=158 genes<br>$CC_2$=128 genes | $Z_1$=Res. tumour<br>$Z_2$=Res. tumour<br>$CC_1$=158 genes<br>$CC_2$=128 genes |
| RR for time-to-relapse (95% CI) | | | | |
| $Z_1$ | - | - | - | 1.17 (1.00-1.37) |
| $CC_1$ | 1.24 (1.14-1.33) | 1.47 (1.37-1.59) | 1.48 (1.37-1.59) | 1.45 (1.35-1.57) |
| RR for time-to-death (95% CI) | | | | |
| $Z_2$ | - | - | 1.18 (1.03-1.35) | 1.30 (1.10-1.53) |
| $CC_2$ | 1.19 (1.09-1.30) | 1.56 (1.43-1.70) | 1.56 (1.44-1.70) | 1.55 (1.42-1.69) |
| Parameter estimate (95% CI) | | | | |
| $\eta$ | 0.028 | 0.035 | 0.039 | 0.040 |
| | (0.005-0.155) | (0.006-0.200) | (0.007-0.227) | (0.007-0.233) |
| $\theta$ | 2.20 | 1.87 | 1.90 | 1.94 |
| | (1.78-2.72) | (1.47-2.39) | (1.49-2.42) | (1.52-2.46) |
| $\tau$ | 0.52 | 0.48 | 0.49 | 0.49 |
| | (0.37-0.68) | (0.32-0.65) | (0.32-0.65) | (0.33-0.66) |
| $\alpha$ | 0 | 0 | 0 | 0 |
| Likelihood cross-validation (LCV) | | | | |
| LCV | -8150.89 | -8088.77 | -8086.36 | -8087.77 |
| DF | 12.16 | 12.01 | 12.07 | 13.99 |

RR, relative risk; Res. tumour (Residual tumour), Z=1 (>1cm) vs Z=0 (<=1cm); CI, confidence interval; LCV=logL-DF, the likelihood cross-validation criterion which accounts for the effective number of parameters (DF, degree of freedom) in the model. The larger LCV value corresponds to the better model.

## References


Anzick SL, Kononen J, Walker RL, et al. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science* 1997; **277**: 965-968.

Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 2009; **25**: 890-896.

Bourboulia D, Jensen-Taubman S, Rittler MR, et al. Endogenous angiogenesis inhibitor blocks tumor growth via direct and indirect effects on tumor microenvironment. *Am J Pathol* 2011; **179**: 2589-2600.

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.

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.

Emura T, Chen HY, Chen YH. compound.Cox: estimation and gene selection based on the compound covariate method under the Cox proportional hazard model, *CRAN* 2016; version 3**.0**: 2016-11-28.

Emura T. 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.

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.

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**: 2016-05-27.

Halon A, Nowak-Markwitz E, Donizy P, et al. Enhanced immunoreactivity of TIMP-2 in the stromal compartment of tumor as a marker of favorable prognosis in ovarian cancer patients. *J Histochem Cytochem* 2012; **60**: 491-501.

Knight JF, Shepherd CJ, Rizzo S, et al. TEAD1 and c-Cbl are novel prostate basal cell markers that correlate with poor clinical outcome in prostate cancer. *Br J Cancer* 2008; **99**: 1849-1858.

Liu TA, Jan YJ, Ko BS, et al. Increased expression of 14-3-3beta promotes tumor progression and predicts extrahepatic metastasis and worse survival in hepatocellular carcinoma. *Am J Pathol* 2011 **179**: 2698-2708.

Osborne CK, Bardou V, Hopp TA, et al. Role of the estrogen receptor coactivator AIB1 (SRC-3) and HER-2/neu in tamoxifen resistance in breast cancer. *J Natl Cancer Inst* 2003; **95**: 353-361.

Shindo K, Aishima S, Ohuchida K, et al. Podoplanin expression in cancer-associated fibroblasts enhances tumor progression of invasive ductal carcinoma of the pancreas. *Mol Cancer* 2013; **12**: 168

Spears M, Oesterreich S, Migliaccio I, et al. The p160 ER co-regulators predict outcome in ER negative breast cancer. *Breast Cancer Res Treat* 2012; **131**: 463-472.

Tzivion G, Gupta VS, Kaplun L, Balan V. 14-3-3 proteins as potential oncogenes. *Semin Cancer Biol* 2006; **16**: 203-213.

Yoshida H, Liu J, Samuel S, Cheng W, Rosen D, Naora H. Steroid receptor coactivator-3, a homolog of Taiman that controls cell migration in the Drosophila ovary, regulates migration of human ovarian cancer cells. *Mol Cell Endocrinol* 2005; **245**: 77-85.

Verveij PJM, van Houwelingen HC, Penalized likelihood in Cox regression. *Stat. in Med*. 1994; **13**: 2427-2436.

Zhang Y, Tang H, Cai J, et al. Ovarian cancer-associated fibroblasts contribute to epithelial ovarian carcinoma metastasis by promoting angiogenesis, lymphangiogenesis and tumor cell invasion. *Cancer Lett* 2011; **303**: 47-55.