Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2017, Vol. 26(6) 2649–2666
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280215604510
journals.sagepub.com/home/smm
SAGE

A joint frailty-copula model between tumour progression and death for meta-analysis

Takeshi Emura,[1] Masahiro Nakatochi,[2] Kenta Murotani[2] and Virginie Rondeau[3]

Abstract

Dependent censoring often arises in biomedical studies when time to tumour progression (e.g., relapse of cancer) is censored by an informative terminal event (e.g., death). For meta-analysis combining existing studies, a joint survival model between tumour progression and death has been considered under semicompeting risks, which induces dependence through the study-specific frailty. Our paper here utilizes copulas to generalize the joint frailty model by introducing additional source of dependence arising from intra-subject association between tumour progression and death. The practical value of the new model is particularly evident for meta-analyses in which only a few covariates are consistently measured across studies and hence there exist residual dependence. The covariate effects are formulated through the Cox proportional hazards model, and the baseline hazards are nonparametrically modeled on a basis of splines. The estimator is then obtained by maximizing a penalized log-likelihood function. We also show that the present methodologies are easily modified for the competing risks or recurrent event data, and are generalized to accommodate left-truncation. Simulations are performed to examine the performance of the proposed estimator. The method is applied to a meta-analysis for assessing a recently suggested biomarker *CXCL12* for survival in ovarian cancer patients. We implement our proposed methods in R *joint.Cox* package.

Keywords

Dependent censoring, dynamic prediction, semicompeting risk, spline, surrogate endpoint, survival analysis

1 Introduction

Several endpoints have been adopted to appropriately demonstrate a clinically convincing effect of treatments, covariates, or biomarkers on survival outcomes. A common endpoint is death, which would be a clinically most significant event for a patient. The overall survival (OS) refers to the time from the patient entry to death due to any cause. Owning to the unambiguity of the definition, OS has been the gold standard endpoint in many cancer studies.[1–3] Another commonly used endpoint is the time to tumour progression (TTP) defined as the time from the entry to the first evidence of disease progression (e.g. relapse, loco-regional progression or distant metastasis). If TTP is used as a primary endpoint, death is one possible cause for censoring (i.e. competing risk). The first occurring event between TTP and OS is progression-free survival (PFS), formally defined as PFS = min{OS, TTP}. Many researchers adopt PFS, rather than TTP alone, since PFS is more associated with OS or PFS itself is a more effective endpoint than OS.[4] The appropriate choice of OS, TTP and PFS as an endpoint and their association has been discussed by many authors.[1,3–7]

In most medical studies, the Cox proportional hazards model[8] is applied to OS, TTP, or PFS separately. The primary interest of such studies is to understand the marginal effect of some covariates on a selected endpoint. Clearly, some dependence exists among OS, TTP, and PFS, and its dependence pattern may be essential to

[1] Graduate Institute of Statistics, National Central University, Jhongli City, Taoyuan, Taiwan
[2] Center for Advanced Medicine and Clinical Research, Nagoya University Hospital, Japan
[3] INSERM CR897 (Biostatistic), Université Bordeaux Segalen, Bordeaux Cedex, France

Corresponding author:
Takeshi Emura, Graduate Institute of Statistics, National Central University, Jhongda Road, Jhongli City, Taoyuan 32001, Taiwan.
Email: takeshiemura@gmail.com; emura@stat.ncu.edu.tw

understand the disease progression mechanisms.[3,5,6] In meta-analytic studies, the dependence is examined to validate the surrogacy of TTP or PFS for OS.[2,3,9] If TTP is the primary endpoint, the Cox regression analysis requires the assumption that TTP and OS are independent (given covariates). This assumption, however, appears to be rarely true owning to the strong link between TTP and OS; death may occur soon after progression.

While the separate (marginal) Cox regression analyses are simple, the observed effects of biomarkers can be biased. Emura and Chen[10] give a systematic reason that the Cox model encounters the bias owning to residual dependence in genetic biomarker search. In a similar reason, the problem of potential bias is evident in meta-analysis for which only a few covariates are consistently measured across studies and hence there is residual dependence.

This paper is concerned about meta-analysis for the joint assessment of TTP and OS using a joint (bivariate) statistical model. While many bivariate survival models useful for fitting TTP and OS have been already proposed in the literature of multivariate survival analysis,[11–13] they are not tailored for meta-analysis. An appropriate model for meta-analysis may include a study-specific (random) effect to explain the variation due to heterogeneity among studies and allow an adequate parameterization of the target population effect.[14] In this respect, Burzykowski et al.[15] propose a parametric Weibull regression and a two-stage semiparametric procedure useful for fitting TTP and OS in meta-analysis. See also Chapter 11 of Burzykowski et al.[9] However, these inference methods are designed for the standard bivariate survival data in which TTP and OS are subject to independent right-censoring only. More realistic setting is the so-called semicompeting risks setting[16] in which OS can dependently censor TTP (i.e. progression is never observed after death). Under the semicompeting risks data, Rondeau et al.[6] developed a joint-frailty model to take into account for the study specific random effect, which in turn induces the dependency between TTP and OS. For estimation, they utilized penalized likelihood techniques under nonparametric models for the two baseline hazards. This approach deals with the dependency between TTP and OS at the study-level; however, there may exist a residual dependency at the patient-level in these meta-analyses.

So this paper aims to develop a copula-based approach for jointly performing the Cox regressions for TTP and OS in meta-analysis. For this purpose, we follow the semicompeting risks framework of Fine et al.,[16] and then generalize the joint-frailty model of Rondeau et al.[6] to incorporate additional source of dependence arising from intra-subject dependence based on copulas. We show that statistical inference methods follow similarly to the penalized likelihood approach of Rondeau et al.[6] even under our broader class of copula models. We supplement the development of the efficient computational schemes by building our original R package "*joint.Cox*".[17] In addition, we demonstrate the proposed methods with the meta-analysis of assessing a recently suggested biomarker *CXCL12* for survival in ovarian cancer patients.[18,19] Finally, we show that the present methodologies are easily modified for the competing risks or recurrent event data, and are generalized to accommodate left-truncation. This implies that our proposal offers a unified framework accommodating a variety of data types in survival analysis.

This paper is organized as follows. Section 2 describes the background. Section 3 introduces our proposed methods. Section 4 conducts simulations and Section 5 performs a meta-analysis of real data. Section 6 discusses extensions of our methods. Section 7 concludes.

2 Background

2.1 Data structure

Meta-analytic data consists of G independent studies with the i-th study containing N_i subjects. Let X_{ij} be time to tumour progression (TTP), D_{ij} be overall survival (OS; i.e., time to death), and C_{ij} be independent and uninformative 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 indicator of progression $\delta_{ij} = \mathbf{I}(T_{ij} = X_{ij})$, where $\mathbf{I}(\cdot)$ is the indicator function, the terminal event time $T^*_{ij} = \min(D_{ij}, C_{ij})$ and the indicator for death $\delta^*_{ij} = \mathbf{I}(T^*_{ij} = D_{ij})$. The data consist of $(T_{ij},\ T^*_{ij},\ \delta_{ij},\ \delta^*_{ij})$ for $i = 1, 2, \ldots, G$ and $j = 1, 2, \ldots, N_i$ (Table 1).

The observation scheme mentioned above is termed "semicompeting risks".[16] This is different from the usual competing-risks data in which two events can censor each other and only the first occurring event time is observable. The semicompeting risks data offers more information about the population than the competing risks data that encounters the un-identifiability about the model of (X_{ij}, D_{ij}).[20]

2.2 Motivating example: meta-analysis of ovarian cancer patients

We consider a recently reported *CXCL12* gene expression as a predictive biomarker of survival in ovarian cancer.[18,19] It has been known that *CXCL12* promotes tumour growth, participates in tumour metastasis, and

Table 1. Four mutually exclusive cases under semi-competing risks.

First occurring event	Terminal event	T_{ij}	T_{ij}^*	δ_{ij}	δ_{ij}^*	Likelihood contribution
Tumour progression	Death	X_{ij}	D_{ij}	1	1	$\Pr(X_{ij} = T_{ij}, D_{ij} = T_{ij}^*)$
Tumour progression	Censoring	X_{ij}	C_{ij}	1	0	$\Pr(X_{ij} = T_{ij}, D_{ij} > T_{ij}^*)$
Death	Death	D_{ij}	D_{ij}	0	1	$\Pr(X_{ij} > T_{ij}, D_{ij} = T_{ij}^*)$
Censoring	Censoring	C_{ij}	C_{ij}	0	0	$\Pr(X_{ij} > T_{ij}, D_{ij} > T_{ij}^*)$

Table 2. A meta-analytic data combining the four independent studies of ovarian cancer patients of Ganzfried et al.[19]

			The number of observed events (event rates %)		
Dataset[a]	Median follow-up (months)	Sample size	Relapse ($\delta_{ij} = 1$)	Death ($\delta_{ij}^* = 1$)	Censoring ($\delta_{ij}^* = 0$)
GSE17260	47	$N_1 = 110$	76 (69%)	46 (42%)	64 (58%)
GSE30161	83	$N_2 = 58$	48 (83%)	36 (62%)	22 (38%)
GSE9891	36	$N_3 = 278$	185 (67%)	113 (41%)	165 (59%)
TCGA	52	$N_4 = 557$	266 (48%)	290 (52%)	267 (48%)
Total		$\sum_{i=1}^{4} N_i = 1003$	575 (57%)	485 (48%)	518 (52%)

Notes: The data are loaded from R Bioconductor curatedOvarianData package of Ganzfried et al.[19]
[a]Dataset is signified as GEO accession number which can be used to search the public genomics data in the GEO (Gene Expression Omnibus) repository. Event rates (%) are the percentage of experiencing a particular event (Relapse, Death or Censoring) within a study; the sums of death and censoring percentages are 100% since each subject can experience only one terminal event (Death or Censoring).

suppresses tumour immunity.[21] The statistical significance of the *CXCL12* expression on survival is first examined by Popple et al.,[18] and is further confirmed by Ganzfried et al.[19] based on the meta-analysis of 14 independent studies. These results are based on the standard Cox regression analysis treating OS as the endpoint. While their analysis focuses solely on OS, it is of our interest to study the significance of *CXCL12* expression on TTP as well as the association between TTP and OS. Therefore, we wish to develop a model to assess the effect of *CXCL12* expression jointly on relapse (TTP) and death (OS). Since the availability of TTP information is limited to $G = 4$ studies (the remaining 10 studies provide OS information only), we concentrate our analyses on them (Table 2). The study sizes considerably vary as $N_1 = 110$, $N_2 = 58$, $N_3 = 278$, and $N_4 = 557$ (total 1003 patients). Table 2 shows the number of observed events for the three events types (relapse, death, and censoring). First, one can recognize that the GSE17260 and GSE9891 studies show quite similar event rates. Compared to the two studies, the GSE30161 study exhibits a higher rate of relapse while the TCGA study exhibits a lower rate of relapse. This suggests a presence of heterogeneity of hazard rates among the four studies, which may not be fully explained by the different follow-up lengths. Our investigation on original papers suggests the heterogeneity of the serous subtype and ethnicity among the four studies. However, such covariates may not be available as the individual-patient data. A typical strategy of meta-analysis is to take into account of the heterogeneity with unobserved random effects.[14] A joint analysis of TTP and OS under meta-analysis with random effects is suggested by Rondeau et al.[6] and implemented by R *frailtypack* package.[22]

2.3 The joint frailty model of Rondeau et al.[6]

This section introduces the joint frailty model of Rondeau et al.,[6] a model tailored for meta-analysis. Let $\mathbf{Z}_{ij}$ be a p-vector of covariates. Consider an unobserved frailty u_i following a density $f_\eta(u_i)$ with $E_\eta(u_i) = 1$ and $Var_\eta(u_i) = \eta$. The hazards for TTP and OS in the joint frailty model[6] are specified as

$$\begin{cases} r_{ij}(t|u_i) = u_i r_0(t) \exp(\boldsymbol{\beta}'_1 \mathbf{Z}_{ij}) & \text{(time-to-progression } X_{ij}) \\ \lambda_{ij}(t|u_i) = u_i^\alpha \lambda_0(t) \exp(\boldsymbol{\beta}'_2 \mathbf{Z}_{ij}) & \text{(overall survival } D_{ij}) \end{cases} \tag{1}$$

The parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are interpreted as fixed (constant) effects of $\mathbf{Z}_{ij}$ on TTP and OS across studies. The forms of the baseline hazards r_0 and λ_0 are not specified. The heterogeneity of the hazards among studies are quantified by a study-specific frailty value u_i. Hence, the model is tailored for meta-analyses that estimate the fixed effect. The intra-cluster dependence between TTP and OS arises by the common u_i in the two hazard functions. The parameter α allows a difference in the amount of heterogeneity between TTP and OS.

If the focus is only on a single event, one can apply the shared frailty model

$$\lambda_{ij}(t|u_i) = u_i\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_{ij})$$

for clustered survival data. Meta-analysis can be performed by treating the study unit as a cluster. The shared frailty model is extensively studied in biostatistics.[12,23]

Clearly, the joint fraily model (1) is a more elaborate model than the shared frailty model. The advantage of the joint model over the usual frailty model is that one can form a bivariate model of TTP and OS, which allows us to investigate the association and dynamics between TTP and OS. See also Rondeau et al.[24] who propose a joint model between recurrent event process and OS, and discuss the advantage of the joint model.

3 Proposed methods

We generalize the joint frailty model of Rondeau et al.[6] by accounting for the intra-subject dependence between TTP and OS in addition to the intra-cluster dependence. We also develop inference procedures by generalizing the penalized likelihood approach of Rondeau et al.[6]

3.1 Joint frailty-copula model

In the joint frailty model (1), the time to tumour progression (TTP, X_{ij}) and overall survival (OS, D_{ij}) are conditionally independent given u_i and $\mathbf{Z}_{ij}$. It is natural to think that there is residual dependence between X_{ij} and D_{ij} within the subject level. One of common reasons to yield residual dependence is insufficiently collected covariates $\mathbf{Z}_{ij}$. In meta-analysis, such residual dependence is a legitimate concern since researchers may access only a few covariates that are consistently obtained across studies. In addition, a strong link between TTP and OS is clear as physicians may encounter death soon after tumour progression.

We relax the conditional independence assumed in Rondeau et al.[6] by introducing the intra-subject dependence with a copula model

$$\Pr(X_{ij} > x\ ,\ D_{ij} > y|u_i) = C_\theta[\ \exp\{\ -R_{ij}(x\ |u_i)\ \},\ \ \exp\{\ -\Lambda_{ij}(\ y\ |u_i)\ \}\ ], \tag{2}$$

where C_θ is a copula[25] with an unknown parameter θ, and

$$R_{ij}(x\ |u_i) = \int_0^x r_{ij}(v\ |u_i)\mathrm{d}v, \quad \Lambda_{ij}(\ y\ |u_i) = \int_0^y \lambda_{ij}(v\ |u_i)\mathrm{d}v,$$

are the cumulative hazards, where r_{ij} and λ_{ij} follow the joint frailty model (1). We call the set of models (1) and (2) "joint frailty-copula model".

The copula describes the dependency between X_{ij} and D_{ij} given u_i. For instance, a mathematically convenient example is the Clayton copula

$$C_\theta(v,\ w) = (v^{-\theta} + w^{-\theta} - 1)^{-1/\theta}, \qquad \theta \geq 0. \tag{3}$$

The copula parameter θ determines the amount of dependence and is related to Kendall's τ via $\tau(X_{ij},\ D_{ij}|u_i) = \theta/(\theta + 2)$. If $\theta \to 0$, then $C_\theta(v, w) = vw$ with $\tau(X_{ij}, D_{ij}|u_i) = 0$ and our model reduces to the joint frailty model of Rondeau et al.[6]

3.2 Likelihood under the joint frailty-copula model

Let $R_{ij}(t) = R_0(t)\exp(\boldsymbol{\beta}'_1\mathbf{Z}_{ij})$ and $\Lambda_{ij}(t) = \Lambda_0(t)\exp(\boldsymbol{\beta}'_2\mathbf{Z}_{ij})$ be cumulative hazard functions for TTP and OS, respectively. The corresponding hazard functions are $r_{ij}(t) = dR_{ij}(t)/dt$ and $\lambda_{ij}(t) =\ d\Lambda_{ij}(t)/dt$. Also, let

$m_i = \sum_{j=1}^{N_i} \delta_{ij}$ (or $m_i^* = \sum_{j=1}^{N_i} \delta_{ij}^*$) be the number of occurrences of TTP (or OS) within the i-th study. Then, the log-likelihood under the models (1) and (2) is

$$\begin{aligned} \ell\,(\alpha, \eta, \theta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r_0, \lambda_0) = \sum_{i=1}^{G} \Bigg[& \sum_{j=1}^{N_i} \{\delta_{ij} \log r_{ij}(T_{ij}) + \delta_{ij}^* \log \lambda_{ij}(T_{ij}^*)\} \\ & + \log \int_0^\infty \Bigg\{ u_i^{m_i + \alpha m_i^*} \prod_{j=1}^{N_i} \psi_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)]^{\delta_{ij}} \psi_\theta^*[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)]^{\delta_{ij}^*} \\ & \times \Theta_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)]^{\delta_{ij}\delta_{ij}^*} D_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)] \Bigg\} f_\eta(u_i) du_i \Bigg], \end{aligned} \tag{4}$$

where $D_\theta[s, t] = C_\theta[\exp(-s), \ \exp(-t)]$, $\psi_\theta = D_\theta^{[1,\,0]}/D_\theta$, $D_\theta^{[1,\,0]} = -\partial D_\theta/\partial s$, $\psi_\theta^* = D_\theta^{[0,\,1]}/D_\theta$, $D_\theta^{[0,\,1]} = -\partial D_\theta/\partial t$, $\Theta_\theta = D_\theta^{[1,\,1]} D_\theta / D_\theta^{[1,\,0]} D_\theta^{[0,\,1]}$ and $D_\theta^{[1,\,1]} = \partial^2 D_\theta/\partial s \partial t$. The derivation is given in Appendix A (Supplementary material online [available at: http://smm.sagepub.com/]). For the frailty distribution, one typically chooses the gamma density

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

where $\eta > 0$ is the variance.

Under the Clayton copula in equation (3), the log-likelihood function has a particularly simple form. Letting $A_\theta(s, t) = \exp(\theta s) + \exp(\theta t) - 1$, one obtains $D_\theta(s, t) = A_\theta(s, t)^{-1/\theta}$, $\psi_\theta(s, t) = \exp(\theta s)/A_\theta(s, t)$, $\psi_\theta^*(s, t) = \exp(\theta t)/A_\theta(s, t)$, and $\Theta_\theta(s, t) = 1 + \theta$. By substituting these forms into equation (4), the likelihood function is readily calculated.

The case of independence copula $C_\theta(v, \ w) = vw$ yields to $D_\theta(s, t) = \exp(-s - t)$ and $\psi_\theta(s, t) = \psi_\theta^*(s, t) = \Theta_\theta(s, t) = 1$. They are also derived as the limit $\theta \to 0$under the Clayton copula. Then, equation (4) reduces to the log-likelihood of Rondeau et al.[6]

$$\begin{aligned} \ell\,(\alpha, \eta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r_0, \lambda_0) = \sum_{i=1}^{G} \Bigg[& \sum_{j=1}^{N_i} \{\delta_{ij} \log r_{ij}(T_{ij}) + \delta_{ij}^* \log \lambda_{ij}(T_{ij}^*)\} \\ & + \log \int_0^\infty \Bigg\{ u_i^{m_i + \alpha m_i^*} \exp\Bigg(-u_i \sum_{j=1}^{N_i} R_{ij}(T_{ij}) - u_i^\alpha \sum_{j=1}^{N_i} \Lambda_{ij}(T_{ij}^*) \Bigg) \Bigg\} f_\eta(u_i) du_i \Bigg]. \end{aligned}$$

This implies that the proposed method covers the original one as a special case.

3.3 Approximation by splines

We suggest approximating the baseline hazards r_0 and λ_0 using splines, which are well-established tools for nonparametric hazard estimation.[26–28] The splines yield a smooth estimate of the hazard function that is not achieved by the discrete approximation of the nonparametric maximum likelihood estimation (NPMLE). This advantage is appealing under the joint model that tries to capture the dynamic behavior of two hazards for TTP and OS. In addition, since the splines are calculated efficiently, they are attractive to work on the complicated likelihood for the joint model.

We set $r_0(t) = \sum_{\ell=1}^{L_r} g_\ell M_\ell(t)$, where $M_\ell(t)$, $\ell = 1, 2, \ldots, L_r$, are the cubic M-spline (a variant of B-spline) bases and $g_\ell \geq 0$, $\ell = 1, 2, \ldots, L_r$, are unknown parameters. We choose $L_r = 5$ that gives good flexibility in curve estimation.[29] The cumulative hazard is $R_0(t) = \sum_{\ell=1}^{L_r} g_\ell I_\ell(t)$, where $I_\ell\,(t)$ is the integration of $M_\ell\,(t)$, called the I-spline basis. The approximation $\lambda_0(t) = \sum_{\ell=1}^{L_\lambda} h_\ell M_\ell(t)$, $h_\ell \geq 0$, can be done similarly. Appendix B (Supplementary material online [available at: http://smm.sagepub.com/]) provides the explicit formulas for $M_\ell\,(t)$ and $I_\ell\,(t)$ with $\ell = 1, \ldots, 5 = L_r = L_\lambda$.

3.4 Penalized likelihood inference

Inference under the cubic spline approximation is implemented with the aid of the penalized maximum likelihood (ML) estimator which maximizes

$$\ell\,(\alpha,\eta,\theta,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,r_0,\lambda_0)-\kappa_1\int\ddot{r}_0(t)^2\mathrm{d}t-\kappa_2\int\ddot{\lambda}_0(t)^2\mathrm{d}t \tag{5}$$

where $\ddot{f}(t)=d^2f(t)/dt^2$, and (κ_1,κ_2) are positive smoothing parameters. The penalty terms represent the roughness of the hazard functions and are rewritten as

$$\int\ddot{r}_0(t)^2dt=\sum_{k=1}^{L_r}\sum_{\ell=1}^{L_r}g_kg_\ell\int\ddot{M}_k(t)\ddot{M}_\ell(t)\mathrm{d}t,\qquad\int\ddot{\lambda}_0(t)^2\mathrm{d}t=\sum_{k=1}^{L_\lambda}\sum_{\ell=1}^{L_\lambda}h_kh_\ell\int\ddot{M}_k(t)\ddot{M}_\ell(t)\mathrm{d}t.$$

After some calculations, the preceding formulas become simple quadratic forms in $\mathbf{g}=(g_1,\ldots,g_{L_r})'$ and $\mathbf{h}=(h_1,\ldots,h_{L_\lambda})'$ (see an example for $L_r=L_\lambda=5$ in Appendix B, Supplementary material online [available at: http://smm.sagepub.com/]). The penalization incorporates the prior knowledge that the hazard does not change very quickly over time and is typically smooth in real applications.

The standard error (SE) is obtained from the inverse of the converged Hessian of the penalized log-likelihood,[26,27] and the confidence interval is formed by the normal approximation. For instance, the 95% confidence interval (CI) for β_1 is

$$\hat{\beta}_1\pm1.96\times\mathrm{SE}(\beta_1)=\hat{\beta}_1\pm1.96\times\sqrt{-\{\,\hat{H}_{PL}^{-1}(\kappa_1,\ \kappa_2)\,\}_{\beta_1}}$$

where $\{\,\hat{H}_{PL}^{-1}(\kappa_1,\ \kappa_2)\,\}_{\beta_1}$ is the relevant component of the converged Hessian matrix for the penalized log-likelihood. Similarly, the 95% CI for the baseline hazard $r_0(x)$ is

$$\hat{r}_0(x)\pm1.96\times\mathrm{SE}\{\,\hat{r}_0(x)\,\}=\mathbf{M}'(x)\hat{\mathbf{g}}\pm1.96\times\sqrt{-\mathbf{M}'(x)\{\,\hat{H}_{PL}^{-1}(\kappa_1,\ \kappa_2)\,\}_{\mathbf{g}}\mathbf{M}(x)},$$

where $\mathbf{M}(x)=(M_1(x),\ldots,M_{L_r}(x))'$.

3.5 Software and computation

We implement automatic computing routines in R *joint.Cox* package (version 2.0).[17] All the numerical results in this paper are produced by the package.

For a given pair of (κ_1,κ_2), the routine maximizes the penalized likelihood in equation (5) by a subroutine R *nlm* with the initial values $\eta=1$, $\theta=1$, $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2=0$, $r_0(t)=\sum_{\ell=1}^{L_r}M_\ell(t)$ and $\lambda_0(t)=\sum_{\ell=1}^{L_\lambda}M_\ell(t)$. The integration in equation (4) is done by a subroutine R *integrate*.

An estimate of α can be highly unstable, especially for small study sizes, e.g. $G=5$. We suggest fixing α (e.g. $\alpha=0$ or $\alpha=1$). The case of $\alpha=0$ refers to the situation that the frailty influences the hazard for TTP, but not for OS. This might be suitable for the case where the observed event rates for death are relatively homogeneous across studies, as in our motivating example (Table 2). The case of $\alpha=1$ refers to the situation that the frailty has the same influence for TTP and OS. In our routine, the user can specify any value of α for which the maximization of the penalized likelihood is performed.

In practice, the following forward selection scheme is recommended for choosing α. The first step is to fit the model with $\alpha^{(0)}=0$. If the penalized log-likelihood value is reduced by fitting with $\alpha^{(1)}=0.25$, then $\alpha^{(0)}=0$ is the choice. Otherwise, we try the update $\alpha^{(2)}=0.5$. These steps are continued by trying $\alpha^{(k+1)}=\alpha^{(k)}+0.25$, $k=1,2,\ldots,$ until the update does not improve the penalized likelihood values from the previous steps. We used this procedure in the real data analyses.

An approximate likelihood cross-validation (LCV) has been used to choose the best smoothing parameter for penalized likelihood inference.[26,27] Since the LCV is developed in the case of the standard Cox model and has not been extended to joint models for choosing $(\kappa_1,\ \kappa_2)$, we only automatically choose κ_1 and κ_2 in two separate standard Cox models for TTP and OS, respectively, and then use these values into the penalized likelihood. The details can be obtained from our source codes. Our routine allows the user to specify a grid for κ_1 and κ_2, and then shows the two LCV curves and their maximum values. One may need to try several plausible grids by visually examining the LCV curves as the adequate choice of the grid considerably depends on many factors such as the number of observed events.

4 Simulations

We conducted simulations to evaluate the performance of the proposed method and to compare our proposal with the method of Rondeau et al.[6]

4.1 Simulation designs

We set two different scenarios:

Scenario (I): $G = 5$ and $N_i = 100$ (or 200) for $i = 1, 2, \ldots, 5$,
Scenario (II): $G = 30$ and $N_i = 10$ for $i = 1, 2, \ldots, 30$.

The number of studies $G = 5$ is common in meta-analyses. For instance, Sabatier et al.[30] examined the effect of *ECRG4* expression on survival by combining $G = 6$ independent studies. Our ovarian cancer meta-analysis (Section 5) is only $G = 4$. The case of $G = 30$ corresponds to a larger pool of studies with the smaller number of subjects.

For each study $i = 1, 2, \ldots, G$, a frailty u_i followed a gamma distribution with variance $\eta = 0.5$. For each subject $j = 1, 2, \ldots, N_i$, a covariate Z_{ij} followed a uniform distribution on (0, 1). Given u_i and Z_{ij}, the distribution of X_{ij} (TTP) and D_{ij} (OS) followed the joint frailty-copula model.

$$\Pr(X_{ij} > x\ ,\ D_{ij} > y | u_i) = [\ \exp\{\ \theta R_{ij}(x\ | u_i)\ \} + \ \exp\{\ \theta \Lambda_{ij}(\ y\ | u_i)\ \} - 1\ ]^{-1/\theta}$$

where $R_{ij}(x\ | u_i) = u_i R_0(x) \exp(\beta_1 Z_{ij})$ and $\Lambda_{ij}(\ y\ | u_i) = u_i^{\alpha} \Lambda_0(\ y) \exp(\beta_2 Z_{ij})$, where $dR_0(x)/dx = r_0(x) = 1$ and $d\Lambda_0(y)/dy = \lambda_0(y) = 1$ were set to be constants. We assumed that $\alpha = 1$ is known and not estimated (see Section 3.5). To introduce the intra-subject dependence between TTP and OS, the association parameter was set at $\theta = 2$ or 6 that corresponds to Kendall's τ equal to 0.5 or 0.75, respectively. Independent censoring variable C_{ij} followed a uniform distribution on (0, 5) that yielded about 16–37% censored subjects. After generating data, we fitted the joint model by using our R package *joint.Cox* (Section 3.5) to estimate β_1, β_2, η, $r_0(\cdot)$, $\lambda_0(\cdot)$ and θ. We based our simulations on 500 replications. All the R codes for simulations are available upon request to the corresponding author.

4.2 Simulation results

Table 3 shows the simulation results under Scenario (I). The parameter estimates appear to be nearly unbiased for regression parameters β_1 and β_2. The standard deviation (SD) of the estimates decreases as the number of subjects increases from $N_i \equiv 100$ to 200 (Table 3). Also, the standard error (SE) accurately captures the SD. Accordingly, the resulting 95% confidence intervals give correct coverage probabilities.

Table 3 also reveals that the estimates for the true copula parameter $\theta = 2$ or $\theta = 6$ are fairly accurate with correct SEs and coverage probabilities. Hence, the estimates provide reliable inference on the degree of intra-subject dependence between TTP and OS.

The estimates exhibit bias for the frailty parameter η, and the bias does not vanish even when the number of subjects increases from $N_i \equiv 100$ to 200 (Table 3). This problem is caused by the small number of studies, $G = 5$. Owning to this reason, the SE is lower than the SD, and the confidence interval has systematic under-coverage. However, the problem becomes less serious by increasing G up to 30 (see the case of $G = 30$, Table 4).

Figure 1 shows the plot of the estimated baseline hazard functions for the first 50 replications. Overall, the estimated hazard functions are good approximation to the true hazards, and their variation reduces as the number of subjects increase from $N_i \equiv 100$ to 200.

Table 4 compares the proposed method with the method of Rondeau et al.[6] under Scenario (II). The proposed estimators are nearly unbiased for β_1, β_2, η, and θ. On the other hand, the method of Rondeau et al.[6] yields some modest biases, especially for β_1. The reason for the biases is attributed to the violation of the intra-subject independence assumption made in the model of Rondeau et al.[6]

In summary, our simulation results have confirmed that accurate inference for regression coefficients and hazard functions is possible even when the study size G is small. This property is important since many existing meta-analyses have small study size. In addition, the proposed method offers better performance than the method of Rondeau et al.[6] when intra-subject dependence exists in the underlying model.

Table 3. Simulation results for the proposed method under Scenario (I) ($G = 5$ studies) based on 500 replications.

		$N_i = 100$				$N_i = 200$			
	Parameter	Mean	SD	SE	CP%	Mean	SD	SE	CP%
CEN=16%	$\beta_1 = 1$	1.003	0.189	0.194	0.96	1.004	0.135	0.135	0.95
	$\beta_2 = 1$	1.010	0.154	0.163	0.96	1.004	0.114	0.114	0.95
	$\eta = 0.5$	0.408	0.264	0.248	0.88	0.399	0.289	0.238	0.82
	$\theta = 2$	2.023	0.247	0.242	0.95	2.015	0.178	0.169	0.94
	κ_1	58.8	176.1	–	–	26.9	100.8	–	–
	κ_2	268.5	418.0	–	–	191.9	363.4	–	–
CEN=32%	$\beta_1 = -1$	−1.001	0.236	0.230	0.95	−1.001	0.157	0.160	0.95
	$\beta_2 = -1$	−1.000	0.194	0.192	0.95	−1.001	0.136	0.134	0.95
	$\eta = 0.5$	0.404	0.263	0.246	0.88	0.395	0.281	0.237	0.82
	$\theta = 2$	2.038	0.296	0.294	0.96	2.019	0.209	0.203	0.94
	κ_1	256.2	389.9	–	–	124.4	276.4	–	–
	κ_2	555.4	470.3	–	–	521.7	469.9	–	–
CEN=18%	$\beta_1 = 1$	1.006	0.154	0.161	0.95	1.004	0.114	0.112	0.95
	$\beta_2 = 1$	1.011	0.143	0.151	0.95	1.004	0.107	0.105	0.95
	$\eta = 0.5$	0.411	0.268	0.249	0.87	0.397	0.279	0.237	0.82
	$\theta = 6$	6.089	0.567	0.561	0.94	6.036	0.396	0.390	0.94
	κ_1	114.1	273.9	–	–	56.7	181.6	–	–
	κ_2	279.9	423.4	–	–	213.5	380.4	–	–
CEN=37%	$\beta_1 = -1$	−1.002	0.197	0.194	0.94	−1.000	0.134	0.135	0.95
	$\beta_2 = -1$	−1.001	0.177	0.179	0.95	−1.001	0.124	0.124	0.96
	$\eta = 0.5$	0.407	0.268	0.248	0.88	0.394	0.274	0.236	0.83
	$\theta = 6$	6.129	0.690	0.672	0.95	6.056	0.462	0.463	0.95
	κ_1	301.5	414.4	–	–	123.5	275.6	–	–
	κ_2	551.8	468.6	–	–	517.8	464.7	–	–

CEN = the percentage that both death and progression are censored; $100 \times \Pr(X_{ij} > C_{ij}, D_{ij} > C_{ij})$. SD = the sample standard deviation of the estimates. SE = the average of the standard errors. CP% = the coverage ratio for the 95% confidence intervals.

Remark: In spite of the superior performance of our proposal over the method of Rondeau et al.,[6] the magnitude of the biases in the method of Rondeau et al.[6] is modest even under the strong intra-subject dependence ($\theta = 6$, $\tau = 0.75$). The real advantage of modeling intra-subject dependence appears when the parameter for the intra-subject dependence is utilized for prediction (Section 5).

5 Meta-analysis of ovarian cancer patients

We performed meta-analysis for ovarian cancer patients of Ganzfried et al.[19] by using the proposed joint model. Following the original paper, we defined TTP as time to recurrence and OS as time to death. Our implementation was based on the individual-patients data available from R Bioconductor package, *curatedOvarianData*. The details of the data are described in Section 2.2 and summarized in Table 2.

The meta-analysis of Ganzfried et al.[19] focused on the significance of *CXCL12* as a univariate predictor of OS, where other clinical characteristics were removed. Such a univariate assessment has been an important step in the stage of gene selection or gene filtering[10,31–33] before building more elaborate predictors. Since the availability of clinical characteristics varied substantially across the studies, we also focused on the univariate assessment of *CXCL12* on survival. As we discussed in Section 2.2, our joint analysis was based on the four ovarian cancer studies for which both TTP and OS are available. We reported the effect of *CXCL12* in terms of the relative risk (RR), say RR = $\exp(\beta_1)$ for recurrence (TTP), and RR = $\exp(\beta_2)$ for death (OS), and their 95% confidence interval (CI). We used the standardized expression values of *CXCL12* as directly available in the data.[19] The expression values had mean 3×10^{-17} and standard deviation 0.999 within the four studies. Hence, RR refers to the increase of the risk for one standard deviation change in *CXCL12* expression.

The results of fitting the proposed method under the Clayton copula are summarized in Table 5. The RR of *CXCL12* on death is significantly greater (RR = 1.18, 95% CI: 1.08–1.29) than the null value (RR = 1). Note that

Table 4. Simulation results comparing the proposed method with the method of Rondeau et al.[6] under Scenario (II) ($G = 30$ studies; $N_i = 10$ subjects) based on 500 replications.

		Proposed method				Method of Rondeau et al.[6]			
	Parameter	Mean	SD	SE	CP%	Mean	SD	SE	CP%
CEN = 16%	$\beta_1 = 1$	1.007	0.263	0.261	0.94	0.929	0.351	0.325	0.92
	$\beta_2 = 1$	1.010	0.231	0.226	0.95	1.007	0.259	0.243	0.93
	$\eta = 0.5$	0.490	0.150	0.146	0.94	0.505	0.149	0.144	0.94
	$\theta = 2$	2.054	0.323	0.330	0.95	0 (fixed)	–	–	–
	κ_1	69.4	169.9	–	–	69.4	169.9	–	–
	κ_2	178.1	350.1	–	–	178.1	350.1	–	–
CEN = 32%	$\beta_1 = -1$	−1.002	0.330	0.303	0.92	−0.897	0.408	0.361	0.91
	$\beta_2 = -1$	−1.007	0.281	0.258	0.93	−1.028	0.304	0.273	0.92
	$\eta = 0.5$	0.495	0.162	0.157	0.95	0.505	0.158	0.151	0.94
	$\theta = 2$	2.061	0.386	0.395	0.95	0 (fixed)	–	–	–
	κ_1	259.5	391.3	–	–	259.5	391.3	–	–
	κ_2	563.3	471.1	–	–	563.3	471.1	–	–
CEN = 18%	$\beta_1 = 1$	1.010	0.229	0.220	0.94	0.961	0.357	0.329	0.93
	$\beta_2 = 1$	1.011	0.220	0.209	0.93	1.006	0.265	0.242	0.93
	$\eta = 0.5$	0.492	0.147	0.145	0.94	0.530	0.157	0.149	0.93
	$\theta = 6$	6.172	0.705	0.760	0.95	0 (fixed)	–	–	–
	κ_1	131.4	275.8	–	–	131.4	275.8	–	–
	κ_2	195.5	362.2	–	–	195.5	362.2	–	–
CEN = 37%	$\beta_1 = -1$	−1.010	0.286	0.259	0.92	−0.928	0.429	0.374	0.90
	$\beta_2 = -1$	−1.014	0.268	0.242	0.92	−1.054	0.316	0.274	0.89
	$\eta = 0.5$	0.496	0.160	0.157	0.96	0.536	0.165	0.159	0.94
	$\theta = 6$	6.218	0.824	0.904	0.96	0 (fixed)	–	–	–
	κ_1	334.7	428.0	–	–	334.7	428.0	–	–
	κ_2	562.2	472.4	–	–	562.2	472.4	–	–

CEN = the percentage that both death and progression are censored; $100 \times \Pr(X_{ij} > C_{ij}, D_{ij} > C_{ij})$. SD = the sample standard deviation of the estimates. SE = the average of the standard errors. CP% = the coverage ratio for the 95% confidence intervals.

this effect is very close to the previously reported effect (RR = 1.15, 95% CI: 1.09–1.23)[19] obtained under the separate Cox regression for death. In our joint analysis, the effect of *CXCL12* on recurrence is even higher (RR = 1.22, 95% CI: 1.13–1.32) than that on death. Our result suggests that the expression of *CXCL12* is a potential biomarker predictive of tumour recurrence in ovarian cancer patients.

Figure 2 shows the two estimated baseline hazards for recurrence and death. The baseline hazard for recurrence is high on early stage and gradually decreases as time passes. On the other hand, the hazard for death is initially low and reaches a peak at around 2000 days. Hereafter, the hazard of death is consistently higher than that of recurrence. This result agrees with the descriptive statistics of Table 2 that the majority of the first occurring events are recurrence. Some practical suggestions to physicians are as follows: we suggest carefully monitoring patients for cancer recurrence before 2000 days, and after 2000 days, shifting more attention to other life-threatening symptoms. These joint assessments of the two event risks may not be straightforward by fitting two separate Cox models to recurrence and death.

The estimate of the copula parameter was $\theta = 2.35$ (95% CI = 1.90–2.90), confirming the presence of positive dependence between TTP and OS at patient level ($\tau = 0.54$, 95% CI = 0.49–0.59). This implies that relapse occurring before death increases the risk of death by 3.35 times through the hazard ratio

$$\frac{\lambda_{ij}(y|X_{ij} = x,\ Z_{ij},\ u_i)}{\lambda_{ij}(y|X_{ij} > x,\ Z_{ij},\ u_i)} = \theta + 1 = 3.35, \qquad y \geq x$$

where $\lambda_{ij}(y|A)$ is the predictive hazard of death at a prediction time y given that event A occurs[34] (see Appendix C, Supplementary material online [available at: http://smm.sagepub.com/], for the details).

From our analysis on baseline hazards, physicians are recommended to change a way to monitor patients at time $x = 2000$ (days). Figure 3 highlights the effect. The predictive hazard of death with relapse

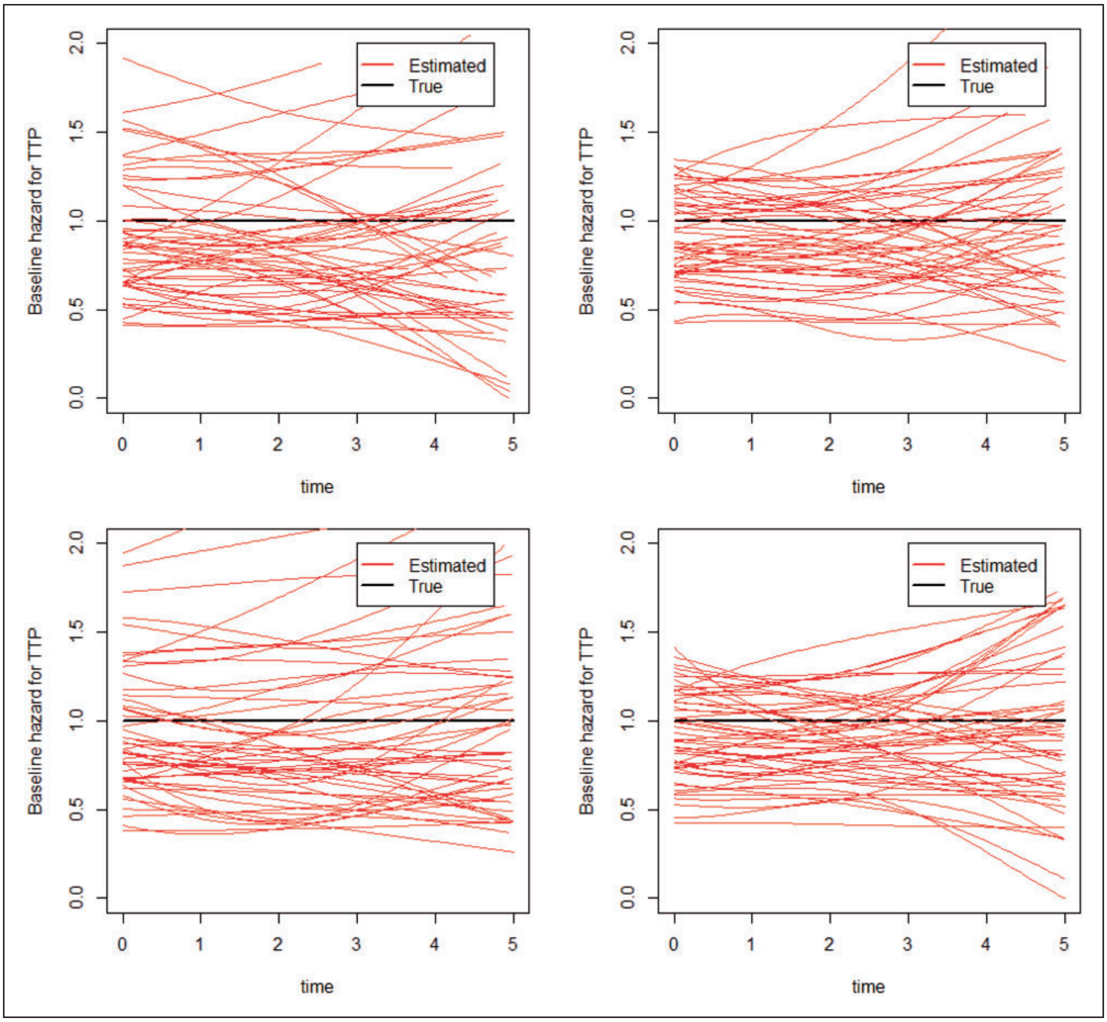

Figure 1. Simulation results for estimating the baseline hazard $r_0(x)$ based on 50 replications. Case (a): $\beta_1 = 1$, $\beta_2 = 1$, $\theta = 2$, $r_0(x) = 1$ and $\lambda_0(y) = 1$; $N_i = 100$ (upper left), $N_i = 200$(upper right). Case (b): $\beta_1 = -1$, $\beta_2 = -1$, $\theta = 2$, $r_0(x) = 1$ and $\lambda_0(y) = 1$; $N_i = 100$ (lower left), $N_i = 200$(lower right)

$\lambda_{ij}(y|X_{ij} = x, Z_{ij}, u_i)$ is remarkably higher than the predictive hazard of death without relapse $\lambda_{ij}(y|X_{ij} > x, Z_{ij}, u_i)$. Figure 3 also shows that the degree of the increase in the risk owning to cancer relapse is much greater than the effect owning to the change in *CXCL12* expression. In fact, given the relapse information, the effect of *CXCL12* expression on death seems to be attenuated as the two hazard functions for *CXCL12* $= -1$ and *CXCL12* $= +1$ cross one another. These results highlight the importance of preventing relapse for 2000 days to prolong patients' OS. Note that a similar risk prediction scheme of death according to relapse information at a given prediction time was previously developed in other joint models.[35,36] Such a prediction scheme is possible only when intra-subject dependence is modeled as in the proposed model.

When we perform the method of Rondeau et al.[6] by ignoring the intra-subject dependence (i.e. assuming $\theta = 0$ as in Section 3.2), the maximized penalized likelihood value reduces substantially (from −8604.093 to −8744.023). Nevertheless, the resulting RRs of *CXCL12* are very similar to those obtained under the proposed method (Table 5). This phenomenon suggests the robustness of the regression estimates in the method of Rondeau et al.[6] against misspecification of the intra-subject dependence.

6 Generalization to other data settings

The proposed methodologies discussed in Section 3 focuses on the semicompeting risks data under the setting of Rondeau et al.[6] In this section, we show that the present methodologies are reduced or extended to other

Table 5. The joint analysis of recurrence (TTP) and death (OS) for the meta-analytic data (four studies, 1003 patients) for ovarian cancer patients of Ganzfried et al.[19]

	Proposed method: Estimate (95% CI)	Method of Rondeau et al.[6]: Estimate (95% CI)
RR[a] for relapse (TTP) : $\exp(\beta_1)$	1.22 (1.13–1.32)	1.24 (1.14–1.35)
RR[a] for death (OS) : $\exp(\beta_2)$	1.18 (1.08–1.29)	1.17 (1.07–1.29)
Heterogeneity: $\eta = Var_\eta(\ u_i\ )$	0.033 (0.006–0.186)	0.028 (0.004–0.180)
Copula parameter: θ	2.35 (1.90–2.90)	0.00 (assumed fixed)
RR for death after relapse: $\theta + 1$	3.35 (2.90–3.90)	1.00 (assumed fixed)
Kendall's tau: $\tau = \theta/(\theta + 2)$	0.54 (0.49–0.59)	–
Maximum penalized log-likelihood	−8604.093	−8744.023

[a]RR (Relative Risk) of *CXCL12* expression on the hazards are examined. "RR>1" indicates that patients with high *CXCL12* expression have poor survival outcomes. The smoothing parameters are estimated as $\kappa_1 = 2.76 \times 10^{16}$ and $\kappa_2 = 3.45 \times 10^{16}$ for both methods.

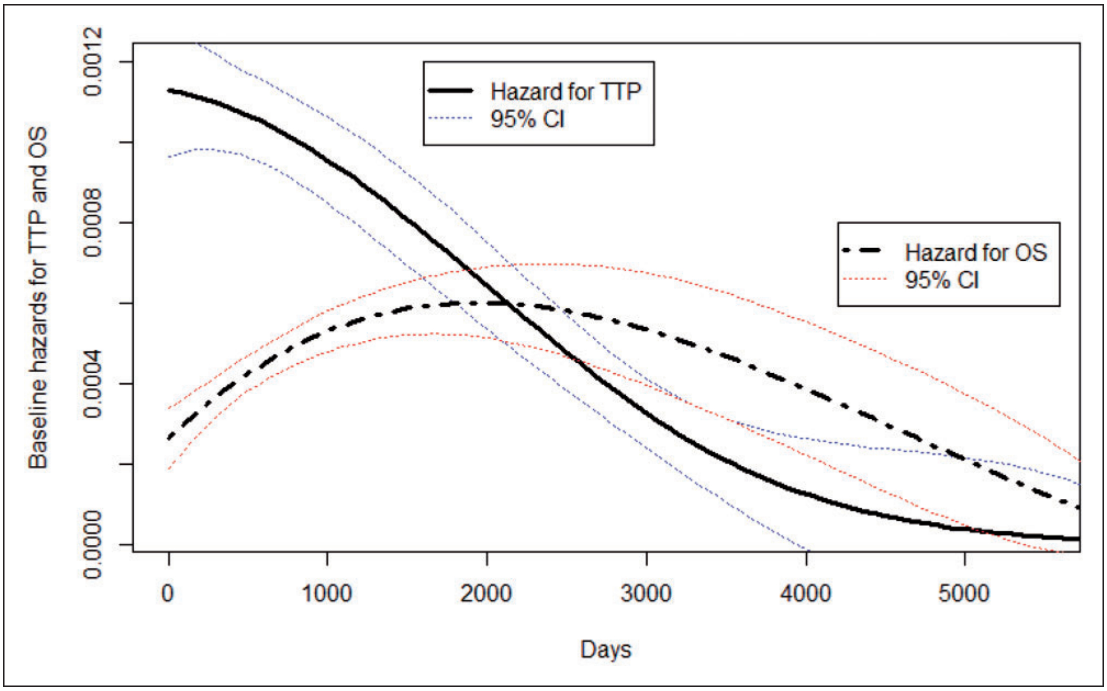

Figure 2. Baseline hazard functions for TTP (recurrence) and OS (death) based on the meta-analytic data of ovarian cancer patients. The dotted lines (red or blue color) show the 95% confidence intervals.

important directions, including the analysis of (i) standard semicompeting risks data, (ii) clustered competing risks data, (iii) standard competing risks data, (iv) clustered semicompeting risks data with left-truncation, and (v) recurrent event data.

As in Section 2, we let X_{ij} be TTP, D_{ij} be OS, and C_{ij} be censoring time for $j = 1, 2, \ldots, N_i$ and $i = 1, 2, \ldots, G$. The difference of the settings comes from different ways to set the study size N_i and different sampling schemes for a triplet $(X_{ij},\ D_{ij},\ C_{ij})$. For instance, in the semicompeting risks setting, a pair $(X_{ij},\ D_{ij})$ is available if a sample exhibits the order $X_{ij} < D_{ij} < C_{ij}$. In the competing risks setting, however, only the first occurring event $\min(X_{ij},\ D_{ij})$ is available for the same sample.

6.1 Standard semicompeting risks data

If data consist of a single study ($G = 1$) with the study size $N_1 = N$ and there is no frailty ($\eta = 0$), the data structure $(T_j,\ T_j^*,\ \delta_j,\ \delta_j^*) \equiv (T_{1j},\ T_{1j}^*,\ \delta_{1j},\ \delta_{1j}^*)$, $j = 1, 2, \ldots, N$ follows the standard semicompeting risks framework of Fine

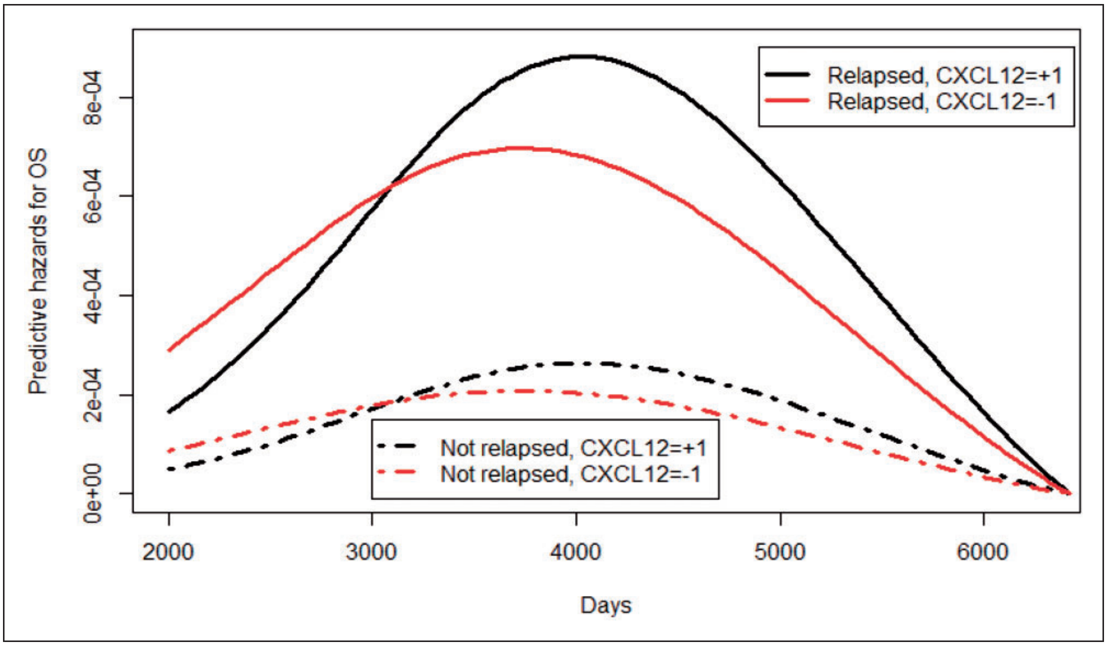

Figure 3. Predictive hazard functions for OS (death) when relapse information is given at time $x = 2000$ days. Four hazard functions are plotted according to the given relapse information (relapsed vs. not relapsed at time $x = 2000$ days) and *CXCL12* gene expression ($Z_{ij} = +1$ vs. $Z_{ij} = -1$). The expressions for the hazard functions are

$$\lambda_{ij}(\, y|X_{ij} = 2000,\, Z_{ij},\, u_i\,) = (\,\theta + 1\,)\lambda_{ij}(\, y|X_{ij} > 2000,\, Z_{ij},\, u_i\,),$$

and

$$\lambda_{ij}(\, y|X_{ij} > 2000,\, Z_{ij},\, u_i\,) = u_i^{\alpha}\lambda_0(y)\exp(\beta_2 Z_{ij})\,\frac{\exp\{\,\theta\Lambda_{ij}(\,y\,|u_i)\,\}}{\exp\{\,\theta R_{ij}(2000\,|u_i)\,\} + \exp\{\,\theta\Lambda_{ij}(\,y\,|u_i)\,\} - 1},$$

evaluated at $u_i = 1$, where all the parameters are estimated (see Appendix C for details).

et al.[16] Under this simplified context, the proposed log-likelihood (4) reduces to that of Chen,[37] as shown in Appendix D (Supplementary material online [available at: http://smm.sagepub.com/]). While Chen[37] suggests the NPMLE, we alternatively propose the penalized likelihood estimation. Other estimation schemes are also available (e.g. Hsieh et al.[38]).

6.2 Clustered competing risks data

Suppose that we observe $T_{ij} = \min(X_{ij},\ D_{ij},\ C_{ij})$, the first-occurring event time among TTP, OS and censoring, and the event-type indicators $\delta_{ij} = \mathbf{I}(T_{ij} = X_{ij})$ and $\delta_{ij}^{*} = \mathbf{I}(T_{ij} = D_{ij})$. This is the competing risks setup with two failure types (progression and death) subject to independent right-censoring. Although death might still be observed even after progression, possible treatment interventions at the time of progression may confound the original interpretation of death.[1,4] In this sense, death and progression censor each other, leading to the competing risks data (Table 6). This implies that the competing risks data has less information than the semicompeting risks data.

Given a copula parameter θ, the log-likelihood is expressed as

$$\begin{aligned}\ell\,(\alpha,\eta,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,r_0,\lambda_0|\theta) = \sum_{i=1}^{G}\Bigg[&\sum_{j=1}^{N_i}\{\delta_{ij}\log r_{ij}(T_{ij}) + \delta_{ij}^{*}\log\lambda_{ij}(T_{ij})\}\\ &+\log\int_0^{\infty} u_i^{m_i+\alpha m_i^{*}}\Bigg\{\prod_{j=1}^{N_i}\psi_\theta[u_iR_{ij}(T_{ij}),u_i^{\alpha}\Lambda_{ij}(T_{ij})]^{\delta_{ij}}\psi_\theta^{*}[u_iR_{ij}(T_{ij}),u_i^{\alpha}\Lambda_{ij}(T_{ij})]^{\delta_{ij}^{*}}\\ &\times D_\theta[u_iR_{ij}(T_{ij}),u_i^{\alpha}\Lambda_{ij}(T_{ij})]\Bigg\}f_\eta(u_i)\mathrm{d}u_i\Bigg].\end{aligned} \tag{6}$$

Table 6. Three mutually exclusive cases under competing risks.

First occurring event	T_{ij}	δ_{ij}	δ^*_{ij}	Likelihood contribution
Progression	X_{ij}	1	0	$\Pr(X_{ij} = T_{ij}, D_{ij} > T_{ij})$
Death	D_{ij}	0	1	$\Pr(X_{ij} > T_{ij}, D_{ij} = T_{ij})$
Censoring	C_{ij}	0	0	$\Pr(X_{ij} > T_{ij}, D_{ij} > T_{ij})$

When performing inference based on the penalized likelihood as in equation (5), we must assume that the copula parameter θ is known. This is because competing risks data may provide little information about the dependence between the competing events.[20] We suggest a sensitivity analysis that examines a range of plausible θ, as commonly done in the competing risks literature.[39–43] Our simulation results reveal that the penalized likelihood inference of Sections 3.3–3.4 exhibits sound statistical performance if the given parameter θ is correctly specified (not shown).

6.3 Standard competing risks data

As a simplified setting of Section 6.2, we assume that the data consist of a single study ($G = 1$) containing N subjects without frailty ($\eta = 0$). Then, the data $(T_j,\ \delta_j,\ \delta^*_j) \equiv (T_{1j},\ \delta_{1j},\ \delta^*_{1j})$, $j = 1, 2, \ldots, N$ follows the standard competing risks setups. Under this simplified setting, the log-likelihood (6) reduces to that of Chen,[42] as shown in Appendix E (Supplementary material online [available at: http://smm.sagepub.com/]).

Under the standard competing risks setting, several copula models with marginal Cox proportional hazards have been considered.[10,41,42,44] Escarela and Carriere[44] proposed the parametric likelihood inference under Weibull marginal distributions. Given a copula parameter θ, Chen[42] proposed the NPMLE for marginal inference. In our approach, we alternatively propose the penalized likelihood estimation. Due to the unidentifiability of θ, we suggest sensitivity analysis as mentioned in Section 6.2.

One can avoid the identifiability issue by fitting the Cox model on the sub-distribution hazard.[45] For this reason, the Cox model on the sub-distribution is more appealing than the Cox model on the marginal hazard, and is frequently used in biostatistics; see Bakoyannis and Touloumi[46] for applications to medicine and Binder et al.[47] for applications to bioinformatics. Do Ha et al.[48] developed statistical inference methods for the clustered competing risks data with the sub-distribution hazard approach which can also be used for meta-analysis or multi-center trials.

6.4 Clustered semicompeting risks data with left truncation

Left truncation usually occurs if the time scale of TTP and OS is age, where the inference focuses on the age-specific hazard. In this case, left-truncation (L_{ij}) corresponds to entry age, and the available samples are subject to the constraint $L_{ij} \le T_{ij}$. Hence, with left truncation, the observed data are $(L_{ij},\ T_{ij},\ T^*_{ij},\ \delta_{ij},\ \delta^*_{ij})$, subject to $L_{ij} \le T_{ij}$, for $j = 1, 2, \ldots, N_i$ and $i = 1, 2, \ldots, G$. The modified expression of the log-likelihood under left-truncation becomes

$$\begin{aligned}\ell\,(\alpha, \eta, \theta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r_0, \lambda_0) = \sum_{i=1}^{G} \Bigg[& \sum_{j=1}^{N_i} \{\delta_{ij} \log r_{ij}(T_{ij}) + \delta^*_{ij} \log \lambda_{ij}(T^*_{ij})\} \\ & + \log \int_0^\infty \Bigg\{ u_i^{m_i + \alpha m^*_i} \prod_{j=1}^{N_i} \psi_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T^*_{ij})]^{\delta_{ij}} \psi^*_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T^*_{ij})]^{\delta^*_{ij}} \\ & \times \Theta_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T^*_{ij})]^{\delta_{ij}\delta^*_{ij}} D_\theta[u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T^*_{ij})] \Bigg\} f_\eta(u_i) \mathrm{d}u_i \\ & - \log \int_0^\infty \prod_{j=1}^{N_i} D_\theta[u_i R_{ij}(L_{ij}), u_i^\alpha \Lambda_{ij}(L_{ij})] f_\eta(u_i) \mathrm{d}u_i \Bigg].\end{aligned}$$

See Appendix A (Supplementary material online [available at: http://smm.sagepub.com/]) for the derivation.

Under the independent copula $C_\theta(v,\ w) = vw$, we have $D_\theta(s, t) = \exp(-s - t)$ and $\psi_\theta(s, t) = \psi^*_\theta(s, t) = \Theta_\theta(s, t) = 1$. Then, the preceding expression is exactly the same as the log-likelihood obtained in Appendix 2 of Rondeau et al.[6]

6.5 Recurrent event data

The semicompeting risks framework of Section 2 and the proposed method in Section 3 can handle recurrent events data under the gap time scale.[22,24] For each subject $i \quad (i = 1, 2, \ldots, G)$, let X_{ij} be the gap time between $(j-1)$-th and j-th event times for $j = 1, 2, \ldots, N_i$. With the information about the terminal event, the induced gap times for death D_{ij} and censoring C_{ij} are similarly defined. The data consists of $(T_{ij},\ T^*_{ij},\ \delta_{ij},\ \delta^*_{ij})$, where $T_{ij} = \min(X_{ij},\ D_{ij},\ C_{ij})$, $\delta_{ij} = \mathbf{I}(T_{ij} = X_{ij})$, $T^*_{ij} = \min(D_{ij},\ C_{ij})$ and $\delta^*_{ij} = \mathbf{I}(T^*_{ij} = D_{ij})$. The gap times X_{ij}, $j < N_i$, are uncensored $(\delta_{ij} = 1)$ while the gap time X_{iN_i} is censored ($\delta_{ij} = 0$) by either death or censoring.

Under the recurrent event setting, the interpretation of the joint frailty-copula model largely differs from the meta-analytic setting. First, the frailty u_i in equation (1) represents the effect of unmeasured covariates at patient level. Thus, this frailty introduces patient-level dependence among recurrences and patient-level dependence between recurrence and death. Second, the copula model (2) describes the residual dependence induced by unmeasured recurrence-specific covariates. That is, even after covariates $\mathbf{Z}_{ij}$ and frailty u_i are given, a pair of gap times $(X_{ij},\ D_{ij})$ is still dependent. This dependence would be ignored if one could obtain sufficient amount of recurrence-specific covariates in each recurrence step j.

For demonstration, we analyze $G = 403$ patients with colorectal cancer who had operations in a hospital in Spain. The data are originally studied by González et al.,[49] and are now available in R *frailtypack* package.[22] The patients are followed up from the date of surgery to either the study end or the time of death whichever comes first. During the follow-up, patients may have several readmissions (recurrences) related to colorectal cancer. The number of recurrences varies from $N_i = 1$ (no readmission) to $N_i = 23$ (22 readmissions).

The major goal of González et al.[49] is to investigate the effect of gender on readmission times. They used the Cox proportional hazards model with the gender as a covariate and the frailty as a source of dependence among readmissions (in gap times) for each patient. Their analysis showed that men have higher hazard for readmissions relative to women (RR = 1.61, 95% CI: 1.21–2.15). We reexamined the conclusion of González et al.[49] under the joint frailty-copula model of readmissions and death in which the fraily accounts for the subject-level dependence, and the copula accounts for the recurrence-level dependence.

The results of fitting the proposed method under the Clayton copula are summarized in Table 7. We see that the gender effect of men on readmissions (RR = 1.66, 95% CI: 1.26–2.20) is very similar to the original result of González et al.[49] The similar results are obtained by the method of Rondeau et al.[6] (Table 7, using *joint.Cox* package or *frailtypack* package). Hence, our results confirm the conclusion of González et al.[49] under our joint frailty-copula model. Figure 4 depicts that the baseline hazard for readmission is consistently higher than that of death. This joint assessment of the two event risks may not be straightforward by the usual Cox regression analysis of González et al.[49]

The proposed method shows significant amount of the estimated frailty variance ($\eta = 1.16$, 95% CI: 0.93–1.45). This reveals the presence of heterogeneity between patients associated with unmeasured (omitted) covariates. The

Table 7. The joint analysis of readmission and death for the colorectal cancer data of González et al.[49] (403 patients).

	Proposed method with the Clayton copula: Using *joint.Cox*	Method of Rondeau et al.[6]: Using *joint.Cox*	Method of Rondeau et al.[6]: Using *frailtypack*
	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
RR[a] for readmission: $\exp(\beta_1)$	1.66 (1.26–2.20)	1.65 (1.24–2.19)	1.82 (1.36–2.42)
RR[a] for death: $\exp(\beta_2)$	1.88 (0.84–4.23)	1.79 (0.80–4.02)	1.45 (0.89–2.35)
Heterogeneity: $\eta = Var_\eta(\ u_i\ )$	1.16 (0.93–1.45)	1.14 (0.91–1.42)	1.01 (0.82–1.20)
α	3.5 (fixed)	3.5 (fixed)	1.35 (0.94–1.76)
Copula parameter: θ	0.57 (0.35–0.94)	0.00 (fixed)	0.00 (fixed)
RR for death after readmission: $\theta + 1$	1.57 (1.35–1.94)	1.00 (fixed)	1.00 (fixed)
Kendall's tau: $\tau = \theta/(\ \theta + 2\ )$	0.22 (0.14–0.31)	–	–
Maximum penalized log-likelihood	−5541.957	−5558.205	–

[a] RR (Relative Risk) of men (relative to women) on the hazards are examined. "RR>1" indicates that men have more readmissions or poor survival outcomes over women do. The smoothing parameters are estimated as $\kappa_1 = 3.4 \times 10^{13}$ and $\kappa_2 = 6.9 \times 10^{13}$ for both methods by using R *joint.Cox* package.

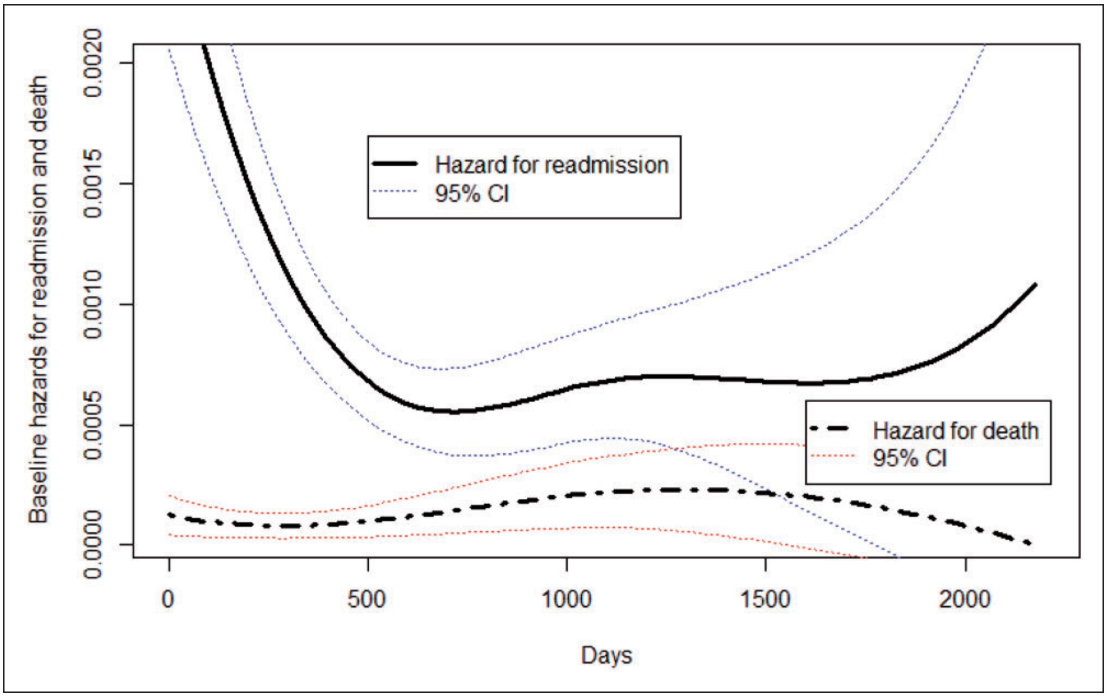

Figure 4. Baseline hazard functions for readmission and death (OS) based on the colorectal cancer patients of González et al.[49] The dotted lines (red or blue color) show the 95% confidence intervals.

method of Rondeau et al.[6] yields very similar estimates for the frailty variance. This variance induces dependence among successive readmissions as well as dependence between readmissions and deaths at patient-level.

The estimated copula parameter in the proposed method suggests that there exists weak residual dependence between readmission and death at recurrence level (Kendall's $\tau = 0.22$, 95% CI = 0.14–0.31). One possible cause of this residual dependence is the use of the same set of covariates for all the recurrence steps, i.e. $\mathbf{Z}_{ij} = \mathbf{Z}_{i1}$ for $j = 1, 2, \ldots, N_i$. This residual dependence could be removed, for instance, by incorporating time-dependent covariates which are updated at the last discharge date. In the absence of such covariates, the proposed copula model adjusts for the residual dependence and improves the log-likelihood value significantly over the model of Rondeau et al.[6] (Table 7).

The proposed method yields the copula parameter $\theta = 0.57$ (95% CI = 0.35–0.94), which implies that each readmission occurring after discharge increases the risk of death by 1.57 ($= \theta + 1$) times (see Appendix C, Supplementary material online, for details [available at: http://smm.sagepub.com/]). This recurrence-level risk prediction of death accounting for the prior readmission event is an important advantage of fitting the copula model.

7 Conclusion and discussion

We propose a copula-based joint model between time to tumour progression (TTP) and overall survival (OS) for meta-analysis with semicompeting risks. As the present paper focuses on meta-analyses combining heterogeneous studies, existing semicompeting risks approaches under bivariate survival models are not straightforwardly applied. In this respect, we build our new model on the basis of the joint frailty model of Rondeau et al.[6] that incorporate the study-specific frailty for meta-analysis. Our copula approach further extends the model of Rondeau et al.[6] by taking into account for intra-subject (patient-level) dependence between TTP and OS via copulas. As seen from our formulations of the Clayton copula model, the flexibility and mathematical convenience of copulas allow one to perform meta-analyses under the complex joint models with minimal computational cost. The simulations show that the penalized likelihood inference with spline approximations to the baseline hazards, originally developed by Rondeau et al.,[6] can perform well under our broader class of models. Remarkably, the simulations reveal the accuracy of the proposed methods even under a small number of studies. Since many meta-analyses for medical research consist of a small number studies (e.g. five studies), the methodologies can be safely applied to these real cases.

The ovarian cancer data meta-analysis revealed a significant amount of intra-subject dependence ($\tau = 0.54$, 95% CI = 0.49–0.59) between TTP and OS. The estimated value of τ can be used to test the intra-subject independence assumption imposed in the model of Rondeau et al.[6] If the amount of τ is not significantly different from zero, then the simpler model of Rondeau et al.[6] is recommended. Another way to test the intra-subject independence assumption is based on the likelihood ratio statistics with reference to the chi-square distribution with one degree of freedom (the case of one-parameter copulas).

With the significant amount of intra-subject dependence between TTP and OS found in the ovarian cancer data, the proposed copula model allowed a dynamic prediction of death using prior occurrence of relapse at patient-level. In particular, our copula model revealed that relapse occurring before death increased the risk of death by 3.35 times (95% CI = 2.90–3.90). Importantly, this change of risk due to relapse was much greater than the effect owning to *CXCL12* gene expression (Figure 3). This finding motivates the development of a more formal dynamic prediction tool as an accurate way to predict death. This topic will be studied in the future.

Modeling for the intra-subject association between TTP and OS is important in many different ways: to understand the disease progression mechanisms,[5] to perform a dynamic prediction of death from prior relapse,[35,36] and to help in the validation of the individual-level surrogacy.[3,9] With these demands in medical research, the proposed copula approach offers a tailored statistical model such that the analysis results can inform medical researchers of the amount of association between TTP and OS. Our simulations (Tables 3–4) show that the statistical inference for the intra-subject association between TTP and OS is made reliably.

The joint information of TTP and OS is not always recorded in publicly available data. For instance, Sabatier et al.[30] performed a meta-analysis combining six studies that analyzed the effect of *ECRG4* gene expression on both TTP (breast cancer relapse) and OS in breast cancer patients (see Table S1 of their paper). Among their six available studies, one study (GSE1456, $n = 159$) offers both TTP (breast cancer relapse) and OS information, one study (GSE3494, $n = 251$) offers OS information only, the study (GSE4922, $n = 249$) offers PFS information and the study (GSE21653, $n = 266$) offers PFS information only. Sabatier et al.[30] conducted separate meta-analyses on PFS and OS based on the available subsets of studies. The joint analysis of TTP and OS are more desired by fitting a single joint model to all the studies. However, it is a challenging topic to develop appropriate adjustments to the likelihood under the missing mechanisms.

Supplementary material

Supplementary materials include Appendix A (Derivation of the log-likelihood function), Appendix B (Cubic spline bases), Appendix C (Prediction and cross-ratio function), Appendix D (Log-likelihood under the standard semicompeting risks data without clustering) and Appendix E (Log-likelihood of the standard competing risks data without clustering).

Acknowledgements

The authors thank the editor and two anonymous reviewers for their helpful comments that substantially improve the manuscript.

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 work is supported by the research grant funded by the Ministry of Science and Technology of Taiwan (MOST 103-2118-M-008-MY2).

References

1. Pazdur R. Endpoints for assessing drug activity in clinical trials. *Oncologist* 2008; **13**: 19–21.

2. Shi Q and Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int J Clin Oncol* 2009; **14**: 102–111.
3. 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.
4. 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.
5. Sherrill B, Amonker M, Wu Y, et al. Relationship between effects on time-to-disease progression and overall survival in studies of metastatic breast cancer. *Br J Cancer* 2008; **99**: 1542–1548.
6. 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 Meth Med Res* 2015; **24**: 711–729.
7. Cheema PK and Burkes RL. Overall survival should be the primary endpoint in clinical trials for advanced non-small cell lung cancer. *Curr Oncol* 2013; **20**: e150–e160.
8. Cox DR. Regression models and life-tables (with discussion). *J Royal Stat Soc Ser B* 1972; **34**: 187–220.
9. Burzykowski T, Molenberghs G and Buyse M (eds) *The Evaluation of Surrogate Endpoints*. New York: Springer, 2005.
10. Emura T and Chen YH. Gene selection for survival data under dependent censoring, a copula-based approach. *Stat Meth Med Res* 2016; **25**: 2840–2857.
11. Hougaard P. *Analysis of Multivariate Survival Data*. New York: Springer, 2000.
12. Duchateau L and Janssen P. *The Frailty Model*. New York: Springer, 2008.
13. Crowder MJ. *Multivariate Survival Analysis and Competing Risks*. Boca Raton: CRC Press, 2012.
14. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-analysis*. Wiley, 2009.
15. 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.
16. Fine JP, Jiang H and Chappell R. On semi-competing risks data. *Biometrika* 2001; **88**: 907–920.
17. Emura T. R joint.Cox: Penalized likelihood estimation under the joint Cox models between TTP and OS for meta-analysis. *CRAN*, R package version 2.0: 2015-06-18. CRAN: The Comprehensive R Archive Network.
18. Popple A, Durrant LG, Spendlove I, et al. The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. *Br J Cancer* 2012; **106**: 1306–1313.
19. 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.
20. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci USA* 1975; **72**: 20–22.
21. Kryczek I, Wei S, Keller E, et al. Stroma-derived factor (SDF-1/CXCL12) and human tumor pathogenesis. *Am J Physiol-Cell Physiol* 2007; **292**: C987–C995.
22. Rondeau V, Gonzalez JR, Mazroui Y, et al. R frailtypack: general frailty models: shared, joint and nested frailty models with prediction. *CRAN*, R package version 2.7.5: 2015-03-06. CRAN: The Comprehensive R Archive Network.
23. Geerdens C, Claeskens G and Janssen P. Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty. *Biostatistics* 2013; **14**: 433–446.
24. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, et al. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics* 2007; **8**: 708–721.
25. Nelsen RB. *An introduction to Copulas*, 2nd ed. New York: Springer Series in Statistics, Springer-Verlag, 2006.
26. O' Sullivan F. Fast computation of fully automated log-density and log-hazard estimation. *SIAM J Sci Stat Comput* 1988; **9**: 363–379.
27. 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.
28. Rondeau V, Commenges D and Joly P. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis* 2003; **9**: 139–153.
29. Ramsay J. Monotone regression spline in action. *Statis Sci* 1988; **3**: 425–461.
30. 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.
31. Jenssen TK, Kuo WP, Stokke T, et al. Association between gene expressions in breast cancer and patient survival. *Human Genetics* 2002; **111**: 411–420.
32. Matsui S. Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinformatics* 2006; **7**: 156.
33. Emura T, Chen YH and Chen HY. Survival prediction based on compound covariate under Cox proportional hazard models. *PLoS ONE* 2012; **7**: e47627.
34. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**: 141–151.
35. 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.
36. 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.

37. Chen YH. Maximum likelihood analysis of semicompeting risks data with semiparametric regression models. *Lifetime Data Anal* 2012; **18**: 36–57.
38. Hsieh JJ, Wang W and Ding AA. Regression analysis based on semicompeting risks data. *J Royal Stat Soc: Ser B (Statistical Methodology)* 2008; **70**: 3–20.
39. Rivest LP and Wells MT. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J Mult Anal* 2001; **79**: 138–155.
40. Braekers R and Veraverbeke N. A copula-graphic estimator for the conditional survival function under dependent censoring. *Can J Stat* 2005; **33**: 429–447.
41. Huang X and Zhang N. Regression survival analysis with an assumed copula for dependent censoring. *Biometrics* 2008; **64**: 1090–1099.
42. Chen YH. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J Royal Stat Soc Ser B* 2010; **72**: 235–251.
43. Stapline ND, Kimber AC, Collett D, et al. Dependent censoring in piecewise exponential survival models. *Stat Meth Med Res* 2015; **24**: 325–341.
44. Escarela G and Carriere JF. Fitting competing risks with an assumed copula. *Stat Meth Med Res* 2003; **12**: 333–349.
45. Fine JP and Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; **94**: 548–560.
46. Bakoyannis G and Touloumi G. Practical methods for competing risks data: a review. *Stat Meth Med Res* 2012; **21**: 257–272.
47. Binder H, Allignol A, Schumacher M, et al. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 2009; **25**: 890–896.
48. Do Ha I, Christian NJ, Jeong JH, et al. Analysis of clustered competing risks data using subdistribution hazard models with multivariate frailties. *Stat Meth Med Res* 2016; **25**: 2488–2505.
49. González JR, Fernandez E, Moreno V, et al. Sex differences in hospital readmission among colorectal cancer patients. *J Epidemiol Commun Health* 2005; **59**: 506–511.

Appendices to: a joint frailty-copula model between tumour progression and death for meta-analysis

Takeshi Emura Graduate Institute of Statistics, National Central University, Jhongda Road, Jhongli City Taoyuan 32001, Taiwan
Masahiro Nakatochi Center for Advanced Medicine and Clinical Research, Nagoya University Hospital, Japan
Kenta Murotani Center for Clinical Research, Aichi Medical University, Japan
Virginie Rondeau INSERM CR897 (Biostatistic), Université Bordeaux Segalen, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France

Corresponding author: Takeshi Emura, Graduate Institute of Statistics, National Central University, Jhongda Road, Jhongli City Taoyuan 32001, Taiwan,
Email: emura@stat.ncu.edu.tw , takeshiemura@gmail.com Tel +886-3-4227151 # 65452

Appendix A: Derivation of the log-likelihood function

Given a frailty u_i, the likelihood for observed data $\mathbf{T}_i = (\, T_{i1}, ..., T_{iN_i}\,)$, $\mathbf{T}_i^* = (\, T_{i1}^*, ..., T_{iN_i}^*\,)$, $\boldsymbol{\delta}_i = (\, \delta_{i1}, ..., \delta_{iN_i}\,)$, and $\boldsymbol{\delta}_i^* = (\, \delta_{i1}^*, ..., \delta_{iN_i}^*\,)$ is

$$\begin{aligned}
L(\, \mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^* \mid u_i\,) &= \prod_{j=1}^{N_i} \Pr(\, X_{ij} = T_{ij}, D_{ij} = T_{ij}^* \mid u_i\,)^{\delta_{ij}\delta_{ij}^*} \Pr(\, X_{ij} = T_{ij}, D_{ij} > T_{ij}^* \mid u_i\,)^{\delta_{ij}(1-\delta_{ij}^*)} \\
&\qquad \times \Pr(\, X_{ij} > T_{ij}, D_{ij} = T_{ij}^* \mid u_i\,)^{(1-\delta_{ij})\delta_{ij}^*} \Pr(\, X_{ij} > T_{ij}, D_{ij} > T_{ij}^* \mid u_i\,)^{(1-\delta_{ij})(1-\delta_{ij}^*)} \\
&= \prod_{j=1}^{N_i} \{\, u_i r_{ij}(T_{ij}) u_i^{\alpha} \lambda_{ij}(T_{ij}^*) D_\theta^{[1,1]}[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]\,\}^{\delta_{ij}\delta_{ij}^*} \times \{\, u_i r_{ij}(T_{ij}) D_\theta^{[1,0]}[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]\,\}^{\delta_{ij}-\delta_{ij}\delta_{ij}^*} \\
&\qquad \times \{\, u_i^{\alpha}\lambda_{ij}(T_{ij}^*) D_\theta^{[0,1]}[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]\,\}^{\delta_{ij}^*-\delta_{ij}\delta_{ij}^*} \times \{\, D_\theta[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]\,\}^{1-\delta_{ij}-\delta_{ij}^*+\delta_{ij}\delta_{ij}^*} \\
&= \left\{ \prod_{j=1}^{N_i} r_{ij}(T_{ij})^{\delta_{ij}} \lambda_{ij}(T_{ij}^*)^{\delta_{ij}^*} \right\} \left\{ \; u_i^{m_i+\alpha m_i^*} \prod_{j=1}^{N_i} \psi_\theta[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]^{\delta_{ij}} \psi_\theta^*[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]^{\delta_{ij}^*} \right. \\
&\qquad \left. \times \Theta_\theta[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,]^{\delta_{ij}\delta_{ij}^*} D_\theta[\, u_i R_{ij}(T_{ij}), u_i^{\alpha}\Lambda_{ij}(T_{ij}^*)\,] \; \right\}.
\end{aligned}$$

Integrating out the unobserved frailty, the contribution of i-th study to the likelihood is

$$
\begin{aligned}
L(\,\mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^*\,) &= \int_0^\infty L(\,\mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^* \mid u_i\,) f_\eta(u_i) du_i \\
&= \prod_{j=1}^{N_i} r_{ij}(T_{ij})^{\delta_{ij}} \lambda_{ij}(T_{ij}^*)^{\delta_{ij}^*} \int_0^\infty \Bigg\{ u_i^{m_i+\alpha m_i^*} \prod_{j=1}^{N_i} \psi_\theta[\,u_i R_{ij}(T_{ij}), u^\alpha \Lambda_{ij}(T_{ij}^*)\,]^{\delta_{ij}} \psi_\theta^*[\,u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)\,]^{\delta_{ij}^*} \\
&\qquad \times \Theta_\theta[\,u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)\,]^{\delta_{ij}\delta_{ij}^*} D_\theta[\,u_i R_{ij}(T_{ij}), u_i^\alpha \Lambda_{ij}(T_{ij}^*)\,] \Bigg\} f_\eta(u_i) du_i.
\end{aligned}
$$

Equation (4) follows by taking logarithm and summing up for $i = 1, 2, ..., G$.

If data is subject to left-truncation, it contains left-truncation times $\mathbf{L}_i = (\,L_{i1}, ..., L_{iN_i}\,)$. Accordingly, observed data $(\,\mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^*\,)$ is obtained under left-truncation constraints $\mathbf{L}_i \le \mathbf{T}_i$ (i.e., $L_{ij} \le T_{ij}$ for $j = 1, 2, ..., N_i$). The truncation probability is

$$
\begin{aligned}
\Pr(\,\mathbf{L}_i \le \mathbf{T}_i\,) &= \int_0^\infty \prod_{j=1}^{N_i} \Pr(\,L_{ij} \le T_{ij} \mid u_i)\; f_\eta(u_i) du_i \\
&= \int_0^\infty \prod_{j=1}^{N_i} \Pr(\,L_{ij} \le C_{ij}, L_{ij} \le X_{ij}, L_{ij} \le D_{ij} \mid u_i)\; f_\eta(u_i) du_i \\
&= \prod_{j=1}^{N_i} \Pr(\,L_{ij} \le C_{ij}\,) \int_0^\infty \prod_{j=1}^{N_i} \Pr(\,L_{ij} \le X_{ij}, L_{ij} \le D_{ij} \mid u_i) f_\eta(u_i) du_i.
\end{aligned}
$$

Accounting for the truncation constraints, the contribution to the likelihood is modified as

$$
\begin{aligned}
L(\,\mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^* \mid \mathbf{L}_i \le \mathbf{T}_i) &= \frac{L(\,\mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^*\,)}{\Pr(\,\mathbf{L}_i \le \mathbf{T}_i\,)} \\
&= \frac{L(\,\mathbf{T}_i, \mathbf{T}_i^*, \boldsymbol{\delta}_i, \boldsymbol{\delta}_i^*\,)}{\prod_{j=1}^{N_i} \Pr(\,L_{ij} \le C_{ij}\,) \int_0^\infty \prod_{j=1}^{N_i} D_\theta[\,u_i R_{ij}(L_{ij}), u_i^\alpha \Lambda_{ij}(L_{ij})\,] f_\eta(u_i) du_i}.
\end{aligned}
$$

This immediately produces the modified log-likelihood under left-truncation.

Appendix B: Cubic spline bases

We use five basis functions to approximate the hazards via $r_0(t)=\sum_{\ell=1}^{5} g_\ell M_\ell(t)$ and $\lambda_0(t)=\sum_{\ell=1}^{5} h_\ell M_\ell(t)$. For a knot sequence $\xi_1<\xi_2<\xi_3$ with an equally spaced mesh $\Delta=\xi_2-\xi_1=\xi_3-\xi_2$, let $z_i(t)=(t-\xi_i)/\Delta$ for $i=1$, 2, and 3. Define M-spline basis functions

$$M_1(t)=-\frac{4\mathbf{I}(\xi_1\le t<\xi_2)}{\Delta}z_2(t)^3,\qquad M_5(t)=\frac{4\mathbf{I}(\xi_2\le t<\xi_3)}{\Delta}z_2(t)^3,$$

$$M_2(t)=\frac{\mathbf{I}(\xi_1\le t<\xi_2)}{2\Delta}\{\,7z_1(t)^3-18z_1(t)^2+12z_1(t)\,\}-\frac{\mathbf{I}(\xi_2\le t<\xi_3)}{2\Delta}z_3(t)^3,$$

$$M_3(t)=\frac{\mathbf{I}(\xi_1\le t<\xi_2)}{\Delta}\{\,-2z_1(t)^3+3z_1(t)^2\,\}+\frac{\mathbf{I}(\xi_2\le t<\xi_3)}{\Delta}\{\,2z_2(t)^3-3z_2(t)^2+1\,\},$$

$$M_4(t)=\frac{\mathbf{I}(\xi_1\le t<\xi_2)}{2\Delta}z_1(t)^3+\frac{\mathbf{I}(\xi_2\le t<\xi_3)}{2\Delta}\{\,-7z_2(t)^3+3z_2(t)^2+3z_2(t)+1\,\}.$$

Also define the integral form $I_\ell(t)=\int_{\xi_1}^{t}M_\ell(u)du$, called the I-spline basis function, written as

$$I_1(t)=1-z_2(t)^4\mathbf{I}(\xi_1\le t<\xi_2),\qquad I_5(t)=z_2(t)^4\mathbf{I}(\xi_2\le t<\xi_3),$$

$$I_2(t)=\{\,\frac{7}{8}z_1(t)^4-3z_1(t)^3+3z_1(t)^2\,\}\mathbf{I}(\xi_1\le t<\xi_2)+\{\,1-\frac{1}{8}z_3(t)^4\,\}\mathbf{I}(\xi_2\le t<\xi_3),$$

$$I_3(t)=\{\,-\frac{1}{2}z_1(t)^4+z_1(t)^3\,\}\mathbf{I}(\xi_1\le t<\xi_2)+\{\,\frac{1}{2}+\frac{1}{2}z_2(t)^4-z_2(t)^3+z_2(t)\,\}\mathbf{I}(\xi_2\le t<\xi_3),$$

$$I_4(t)=\frac{1}{8}z_1(t)^4\mathbf{I}(\xi_1\le t<\xi_2)+\{\,\frac{1}{8}-\frac{7}{8}z_2(t)^4+\frac{1}{2}z_2(t)^3+\frac{3}{4}z_2(t)^2+\frac{1}{2}z_2(t)\,\}\mathbf{I}(\xi_2\le t<\xi_3).$$

Figure A depicts the M- and I-spline basis functions with knots $\xi_1=1$, $\xi_2=2$, and $\xi_3=3$.

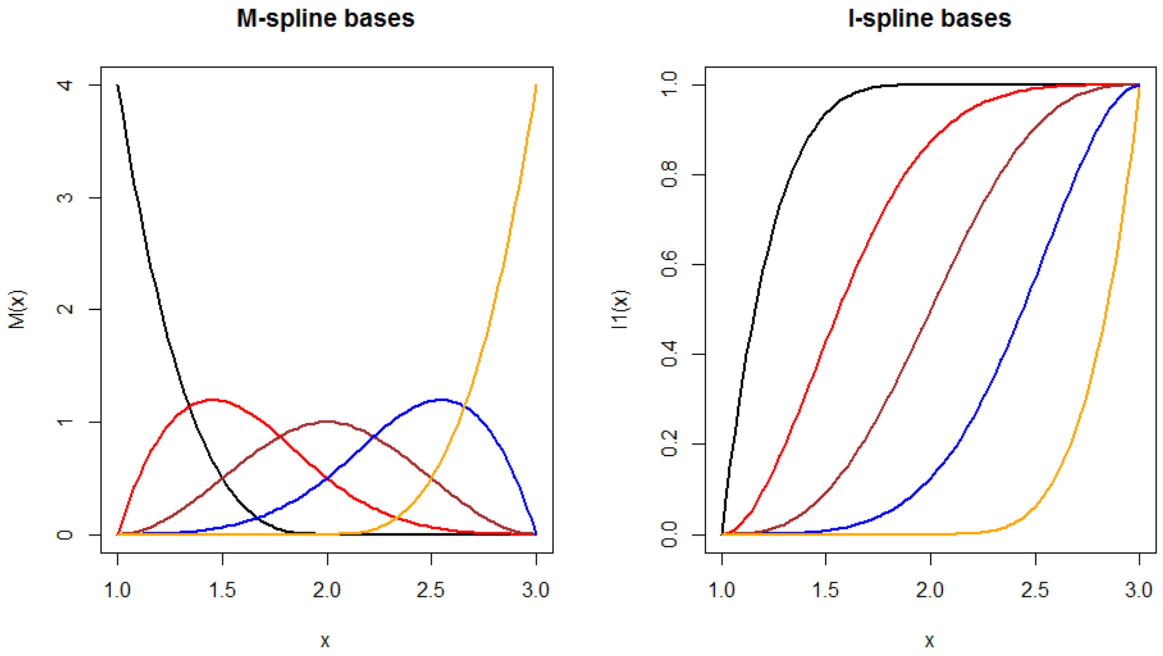

Figure A. The plots for five M-spline basis functions (left-panel) and I-spline basis functions (right-panel) with equally spaced knots $\xi_1 = 1$, $\xi_2 = 2$, and $\xi_3 = 3$.

The second derivatives of the five basis functions are

$$\ddot{M}_1(t) = -\frac{24}{\Delta^3} z_2(t)\mathbf{I}(\xi_1 \le t < \xi_2), \quad \ddot{M}_5(t) = \frac{24}{\Delta^3} z_2(t)\mathbf{I}(\xi_2 \le t < \xi_3)$$

$$\ddot{M}_2(t) = \left\{ \frac{21}{\Delta^3} z_1(t) - \frac{18}{\Delta^3} \right\}\mathbf{I}(\xi_1 \le t < \xi_2) - \frac{3}{\Delta^3} z_3(t)\mathbf{I}(\xi_2 \le t < \xi_3)$$

$$\ddot{M}_3(t) = \left\{ -\frac{12}{\Delta^3} z_1(t) + \frac{6}{\Delta^3} \right\}\mathbf{I}(\xi_1 \le t < \xi_2) + \left\{ \frac{12}{\Delta^3} z_2(t) - \frac{6}{\Delta^3} \right\}\mathbf{I}(\xi_2 \le t < \xi_3)$$

$$\ddot{M}_4(t) = \frac{3}{\Delta^3} z_1(t)\mathbf{I}(\xi_1 \le t < \xi_2) + \left\{ -\frac{21}{\Delta^3} z_2(t) + \frac{3}{\Delta^3} \right\}\mathbf{I}(\xi_2 \le t < \xi_3).$$

It follows that

$$\int \ddot{M}_1(t)^2 dt = \frac{192}{\Delta^5}, \int \ddot{M}_2(t)^2 dt = \frac{96}{\Delta^5}, \int \ddot{M}_3(t)^2 dt = \frac{24}{\Delta^5}, \int \ddot{M}_4(t)^2 dt = \frac{96}{\Delta^5}, \int \ddot{M}_5(t)^2 dt = \frac{192}{\Delta^5},$$

$$\int \ddot{M}_1(t)\ddot{M}_2(t)dt = -\frac{132}{\Delta^5},\ \int \ddot{M}_1(t)\ddot{M}_3(t)dt = \frac{24}{\Delta^5},\ \int \ddot{M}_1(t)\ddot{M}_4(t)dt = \frac{12}{\Delta^5},\ \int \ddot{M}_1(t)\ddot{M}_5(t)dt = 0,$$

$$\int \ddot{M}_2(t)\ddot{M}_3(t)dt = -\frac{24}{\Delta^5},\ \int \ddot{M}_2(t)\ddot{M}_4(t)dt = -\frac{12}{\Delta^5},\ \int \ddot{M}_2(t)\ddot{M}_5(t)dt = \frac{12}{\Delta^5},$$

$$\int \ddot{M}_3(t)\ddot{M}_4(t)dt = -\frac{24}{\Delta^5},\ \int \ddot{M}_3(t)\ddot{M}_5(t)dt = \frac{24}{\Delta^5},\ \int \ddot{M}_4(t)\ddot{M}_5(t)dt = -\frac{132}{\Delta^5},$$

where the range of integral is $(\ \xi_1, \xi_3\ ]$. Then, the penalization term is explicitly computed as matrix algebras with

$$\int \ddot{r}_0(t)^2 dt = \sum_{k=1}^{5}\sum_{\ell=1}^{5} g_k g_\ell \int \ddot{M}_k(t)\ddot{M}_\ell(t)dt = \frac{1}{\Delta^5}\mathbf{g}'\begin{bmatrix} 192 & -132 & 24 & 12 & 0 \\ -132 & 96 & -24 & -12 & 12 \\ 24 & -24 & 24 & -24 & 24 \\ 12 & -12 & -24 & 96 & -132 \\ 0 & 12 & 24 & -132 & 192 \end{bmatrix}\mathbf{g}.$$

Appendix C: Prediction and cross-ratio function

We consider prediction of death at time $y \geq x$ conditional on events occurring time x:

1) Predictive hazard of death **with relapse** at x:

$$\lambda_{ij}(\ y \mid X_{ij} = x, Z_{ij}, u_i\ ) = \lim_{h\to 0} \Pr(\ D_{ij} < y + h \mid D_{ij} \geq y,\ X_{ij} = x, Z_{ij}, u_i\ )/h, \qquad y \geq x$$

2) Predictive hazard of death **without relapse** at x:

$$\lambda_{ij}(\ y \mid X_{ij} \geq x, Z_{ij}, u_i\ ) = \lim_{h\to 0} \Pr(\ D_{ij} < y + h \mid D_{ij} \geq y,\ X_{ij} \geq x, Z_{ij}, u_i\ )/h, \qquad y \geq x$$

Under the joint-frailty copula model, their formulas are

$$\lambda_{ij}(\ y \mid X_{ij} = x, Z_{ij}, u_i\ ) = u_i^{\alpha}\lambda_0(y)\exp(\beta_2 Z_{ij})\frac{D_\theta^{[1,1]}[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]}{D_\theta^{[1,0]}[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]}, \qquad y \geq x,$$

$$\lambda_{ij}(\ y \mid X_{ij} \geq x, Z_{ij}, u_i\ ) = u_i^{\alpha}\lambda_0(y)\exp(\beta_2 Z_{ij})\frac{D_\theta^{[0,1]}[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]}{D_\theta[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]}, \qquad y \geq x.$$

The hazard ratio is

$$\frac{\lambda_{ij}(\ y \mid X_{ij} = x, Z_{ij}, u_i\ )}{\lambda_{ij}(\ y \mid X_{ij} \geq x, Z_{ij}, u_i\ )} = \frac{D_\theta[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]D_\theta^{[1,1]}[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]}{D_\theta^{[1,0]}[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]D_\theta^{[0,1]}[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ]}.$$

$$\equiv \Theta_\theta[\ u_i R_{ij}(x), u_i^{\alpha}\Lambda_{ij}(y)\ ], \qquad y \geq x$$

This is the so-called “cross-ratio” function (Oakes, 1986) which is interpreted as

- $\Theta_\theta > 1$; intra-subject positive association (relapse increases the risk of death),
- $0 < \Theta_\theta < 1$; intra-subject negative association (relapse decreases the risk of death),
- $\Theta_\theta = 1$; intra-subject independence (relapse is not related to death).

Under the Clayton copula, it is easy to show that the cross-ratio function is a constant,

$$\Theta_\theta[\, u_i R_{ij}(x), u_i^{\alpha} \Lambda_{ij}(y)\,] = 1 + \theta \,.$$

Indeed, the Clayton model is derived as the constant odds ratio model (Clayton, 1978).

Appendix D: Log-likelihood under the standard semicompeting risks data without clustering

The proposed log-likelihood of Equation (4) reduces to

$$\begin{aligned}\ell(\,\theta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r_0, \lambda_0\,) = & \sum_{j=1}^{N} \{\, \delta_j \log r_j(T_j) + \delta_j^* \log \lambda_j(T_j^*)\,\} \\ & + \sum_{j=1}^{N} \{\, \delta_j \log \psi_\theta[\, R_j(T_j), \Lambda_j(T_j^*)\,] + \delta_j^* \log \psi_\theta^*[\, R_j(T_j), \Lambda_j(T_j^*)\,]\,\} \\ & + \sum_{j=1}^{N} \{\, \delta_j \delta_j^* \log \Theta_\theta[\, R_j(T_j), \Lambda_j(T_j^*)\,] + \log D_\theta[\, R_j(T_j), \Lambda_j(T_j^*)\,]\,\},\end{aligned}$$

where $R_j(t) = R_0(t)\exp(\boldsymbol{\beta}_1'\mathbf{Z}_j)$, $\Lambda_j(t) = \Lambda_0(t)\exp(\boldsymbol{\beta}_2'\mathbf{Z}_j)$, $r_j = dR_j / dt$, and $\lambda_j = d\Lambda_j / dt$. The preceding log-likelihood is equivalent to that derived by Chen (2012) under the mode

$$\Pr(\, X_j > x\,, D_j > y\,) = C_\theta[\, \exp\{\, -R_j(x)\,\}, \exp\{\, -\Lambda_j(y)\,\}\,]\,. \qquad \text{(C)}$$

Appendix E: Log-likelihood of the standard competing risks data without clustering

The log-likelihood of Equation (6) reduces to

$$\begin{aligned}\ell(\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r_0, \lambda_0 \mid \theta\,) = & \sum_{j=1}^{N} \{\, \delta_j \log r_j(T_j) + \delta_j^* \log \lambda_j(T_j\,)\,\} \\ + \sum_{j=1}^{N} \{\, \delta_j \log \psi_\theta[\, R_j(T_j), \Lambda_j(T_j\,)\,] & + \delta_j^* \log \psi_\theta^*[\, R_j(T_j), \Lambda_j(T_j\,)\,]\,\} + \sum_{j=1}^{N} \log D_\theta[\, R_j(T_j), \Lambda_j(T_j\,)\,],\end{aligned}$$

where all the mathematical symbols follow Appendix C. The preceding expression is the log-likelihood proposed by Chen (2010) under the joint model (C).

References

ChenYH. Maximum likelihood analysis of semicompeting risks data with semiparametric regression models. *Lifetime Data Anal* 2012; **18**:36–57.

Chen YH. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula, *Journal of the Royal Statistical Society, Ser. B* 2010; **72:** 235-51.

Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**(1), 141-151.

Oakes D. Semi-parametric inference in a model for association in bivariate survival data, *Biometrika* 1986; **73**, 353-361.