A Joint modelling of socio-professional trajectories and cause-specific mortality

M. Karimi^{a,b,*}, G. Rey^a, A. Latouche^{c,d}

^a*Inserm-CépiDc, 80, Rue du Général Leclerc, Le Kremlin-Bicêtre, Cedex 94270, France*

^b*MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK*

^c*Conservatoire national des arts et métiers (Cnam), EA4629, Paris, France*

^d*Inserm U900, Institut Curie, Saint-Cloud, France*

Abstract

The association between life-course socio-professional trajectories and mortality has already been discussed in the literature. However, these socio-professional trajectories may be subject to informative censoring due to death. This loss to follow-up which is related to an individual's survival, should not be ignored and thus, it is of interest to model jointly these professional trajectories and their survival. The main focus has been made on continuous, binary or ordinal variables while much less attention has been paid to nominal categorical data. Therefore, an extension to the joint modelling of longitudinal nominal data and survival under a likelihood-based approach is proposed. A generalized linear mixed model is considered for modelling the longitudinal nominal data, in addition to two cause-specific proportional hazards model for the survival competing risks data. The association between longitudinal and survival outcomes is captured by assuming a multivariate Gaussian distribution for the joint distribution of the random effects of two sub-models. The proposed joint model provides a robust framework for estimating longitudinal membership probabilities, accounting for informative censoring caused by individual's death. Simulations are carried out to assess the performance of this joint model comparing with the results of the separate longitudinal and competing risks analysis. A disadvantage of joint models is that they are computationally intensive. To overcome this problem, an approach mimicking a meta-analysis strategy of individual participant data is suggested. The relevance of this approach is then illustrated on a large sample of the French salaried population, which contains all employment records between 1976 and 2002.

Keywords: Generalized Linear Mixed Model, Cause-specific Hazards, Joint Model, Membership Probability, Large-scale Data

*Corresponding author at: MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK

Email addresses: m.karimi@imperial.ac.uk (M. Karimi), gregoire.rey@inserm.fr (G. Rey), aurelien.latouche@cnam.fr (A. Latouche)

1. Introduction

The motivation of this work comes from an epidemiological question that is measuring the association between life-course socio-professional position and the occurrence of mortality by different causes of death. For this purpose, we have the Cosmop-DADS database, which contains individuals' annual occupational episodes between 1976 and 2002 linked to the causes of death. The data on employment episodes are classified based on various social characteristics without an obvious hierarchical order such as income. In earlier studies it was common to consider just a limited number of episodes, either an individual's employment at entry into the labour market or his/her position at mid-life age (Melchior et al., 2006; Davey Smith et al., 1997). To capture the evolution of socio-economic position and the transitions between social groups, the whole professional trajectory, which can be modelled as a sequence of consecutive professional positions occupied by an individual, should be taken into account. Model-based, such as Multistate models, and model-free such as Sequence Analysis approaches of life course analysis are traditionally used to study life course data, but these approaches consider completed trajectories only, and thus left-truncated and not randomly selected data (Eerola and Helske, 2012). Considering other characteristics of professional trajectories, such as cumulated time spent in each professional position and the number of transitions between professional episodes, in addition to the professional position itself, will aide in the understanding of professional trajectories process. To simultaneously consider the professional and mortality process, a naïve and common approach is to use the administrative employment positions and different characteristics of professional trajectories as time-dependent covariates in the extended Cox proportional hazards model (Karimi et al., 2015). However, the employment episodes are collected only for subjects under the study, and so are endogenous time-dependent covariates. Thus, analysis based on the extended Cox model (Andersen et al., 1993), which assumes that the covariates path is predictable, is not optimal for endogenous time-dependent covariates and might generate biases.

In this work, the focus is on the association between professional trajectories and death due to different causes of death. Death of each individual withdraws him/her from the study and results in informative censoring. Since the missing data conducted from this censoring is directly related to the individual's survival, the informative censoring is similar to the Missing Not At Random (MNAR) mechanism in longitudinal studies and should not be ignored. Consequently, it is of interest to model jointly the professional trajectories and the missingness process, in other words, the competing risks.

Joint analysis of longitudinal outcomes and survival data can be categorized into *pattern-mixture models*, *selection models* and *random effects models* (Diggle, 1998). Although mathematically all these models describe the joint distribution of longitudinal outcomes and survival data, they have different statistical interpretations. The pattern-mixture models focus on estimating the longitudinal trajectory conditional on the missingness process, in opposition to the selection models where the interest lies in estimating the

missingness process given the longitudinal outcomes. The third category is based on the assumption that both missingness and longitudinal processes have arisen from an underlying health condition. This health condition is defined by a random effect and hence, longitudinal and missingness processes are independent, conditional on this random effect. Most of the proposed joint models in this category, known as *shared-parameter models*, use a shared random effect between longitudinal process and missingness mechanism. An extension of this model was proposed by [Henderson et al. \(2000\)](#) which postulates a latent bivariate Gaussian process. They assume that the longitudinal and missingness processes are independent given the Gaussian process and covariates. As a result, the association between longitudinal and missingness processes can be described by the correlation between the two components of the bivariate Gaussian process. Since both processes are not sharing the same random effects, this extension allows random effects to influence the missingness process independently of the longitudinal process. We focus on this category, and we refer to this class of models as *joint models for longitudinal and time-to-event data*.

The first forms of the joint models were introduced by [Faucett and Thomas \(1996\)](#), and [Wulfsohn and Tsiatis \(1997\)](#), and since several extensions have been proposed in the literature. The fundamental idea of the so-called joint models is based on linking the survival model with a suitable model for the longitudinal measurements, usually a random effect model ([Laird and Ware, 1982](#)), in which the correlations between the repeated measures are not ignored, via a common unobserved structure, to capture the correlations between the two longitudinal and survival processes. Different approaches have been developed in the literature for the association structure in joint models. One may include the mixed model defined for the longitudinal outcomes as a covariate in the survival sub-model ([Faucett and Thomas, 1996](#); [Rizopoulos, 2012](#)). An alternative approach would be including the random effects directly in both longitudinal and survival sub-models with an assumed joint distribution for the random effects ([Henderson et al., 2000](#); [Elashoff et al., 2007](#); [Li et al., 2010](#)). Excellent overviews of these models can be found in the literature ([Tsiatis and Davidian, 2004](#); [Rizopoulos, 2013](#)). Most of these works have focused on continuous measurements or the quality of life measurements, on binary measurements or ordinal responses and there has been less attention to non-ordinal categorical longitudinal outcomes, as in Cosmop-DADS database. The recent extension of the joint modelling of categorical longitudinal data and time-to-event data, developed by [Murawska and Rizopoulos \(2013\)](#), is based on a Bayesian approach. However, the main issue of this method is the high calculation time, even for a small dataset. Given the structure of our large dataset, the Cosmop-DADS database, in this work we extend the work of [Li et al. \(2010\)](#), by proposing a joint model for nominal longitudinal data and competing risks data in a likelihood-based framework.

To model the nominal longitudinal data, we focus on mixed-effects regression models. This choice is based on the characteristics of these classes of models that include subject-specific random effects into the regression models, that take into account for the influence of subjects on their repeated observations. Therefore, in the joint model, we adopt a Generalized Linear Mixed Model (GLMM) for nominal responses

to model the longitudinal trajectories and two cause-specific proportional hazards models for competing risks survival data. The association between longitudinal and survival outcomes is captured by assuming a multivariate Gaussian distribution for the joint distribution of the random effects of two sub-models. The Expectation-Maximization (EM) algorithm is used to estimate the parameters in both longitudinal and survival sub-models, as well as the inverse of the empirical Fisher information from the profile likelihood for approximating the variance-covariance matrix of the parameter estimates. The proposed joint model also provides a robust framework to estimate longitudinal membership probabilities, accounting for the informative censoring.

So far, the existing joint models have been applied to sample sizes up to 2000 individuals, which is much smaller than the motivating database of this work. The second objective of this study is to propose an approach mimicking a meta-analysis, to address the calculation problems in joint models and large datasets, by extracting independent stratified samples from the large dataset, applying the joint model on each sample and then combining the results.

The remainder of this paper is organized as follows. In section 2, we describe our proposed joint model. Section 3 contains the likelihood function of the model and the parameter estimation procedure. In section 4, we conduct a simulation study to evaluate the performance of the proposed model. In section 5, we apply the joint modelling of longitudinal nominal data and cause-specific hazards on a large sample of the CosmopDADS, and where we calculated the marginal membership probabilities. We conclude the paper with a discussion in section 6.

2. Joint modelling framework

The proposed joint model is comprised of three components: the nominal longitudinal sub-model, the cause-specific sub-models and the variance-covariance matrix of random effects to describe the joint association of repeated values and competing risks data.

2.1. Nominal longitudinal sub-model

Let the longitudinal sub-model be specified as a GLMM (Hartzel et al., 2001), an extension of the Generalized Linear Models (GLM), incorporating random effects as well as fixed effects in the linear predictor. Adding random effects allows for multiple observations on each subject, and thus takes into account the correlation within the observations of each subject, by incorporating subject-specific random effects. Since the responses are nominal, the GLMM is presented as a baseline-category logit model with random effects that pairs each category with a reference category (Agresti, 2013).

Let n be the number of subjects in the study and m_i be the number of repeated values for each subject, $i = 1, \dots, n$. Let Y_{ij} denote the observed value for subject i on occasion j ($j = 1, \dots, m_i$). We assume that

the repeated values are nominal data with K modalities, $Y_{ij} = k \in \{1, \dots, K\}$ represents that the modality k is observed on occasion j for a given individual i . Let $\mathbf{X}_{ij}$ be a $p \times 1$ vector of predictors for the fixed effects and $\mathbf{W}_{ij}$ be a $q \times 1$ vector of predictors for the random effects. The linear predictor of the GLMM for nominal outcomes is defined as $\eta_{ijk} = \alpha_k + \mathbf{X}_{ij}'\beta_k + \mathbf{W}_{ij}'\mathbf{b}_{ik}$ and the probability, π_{ijk} , that the modality k is observed for the j -th value of a given individual i , conditional on the random effects $\mathbf{b}_i$, is given by:

$$\pi_{ijk} = P(Y_{ij} = k | \mathbf{X}_{ij}, \mathbf{W}_{ij}, b_{ik}) = \begin{cases} \frac{1}{1 + \sum_{h=1}^{K-1} \exp(\alpha_h + \mathbf{X}_{ij}'\beta_h + \mathbf{W}_{ij}'\mathbf{b}_{ih})} & \text{if } k = K, \\ \frac{\exp(\alpha_k + \mathbf{X}_{ij}'\beta_k + \mathbf{W}_{ij}'\mathbf{b}_{ik})}{1 + \sum_{h=1}^{K-1} \exp(\alpha_h + \mathbf{X}_{ij}'\beta_h + \mathbf{W}_{ij}'\mathbf{b}_{ih})} & \text{if } k = 1, \dots, K-1. \end{cases} \quad (1)$$

Then, $\alpha = (\alpha_1, \dots, \alpha_{K-1})'$ is the vector of intercepts with $\alpha_K = 0$. $\beta_k = (\beta_{k1}, \dots, \beta_{kp})'$ is a $p \times 1$ vector of the fixed effects parameters with $\beta_K = \mathbf{0}$. So $\exp(\beta_{ks})$, with β_{ks} the s -th element of β_k , can be interpreted as the increase in odds of falling into modality k versus modality K resulting from a one-unit increase in the s -th covariate, given observing the same random effect and holding other covariates constant. Let $\beta = (\beta'_1, \dots, \beta'_{K-1})'$. $\mathbf{b}_{ik} = (b_{ik1}, \dots, b_{ikq})'$ is a $q \times 1$ vector of the random effects for subject i in the k -th modality. The random effects $\mathbf{b}_{ik}$ are commonly assumed to follow a multivariate Gaussian distribution with mean zero $\mathbf{0}$ and the covariance matrix $\Sigma_{\mathbf{b}_k}$, $\mathbf{b}_{ik} \sim N_q(\mathbf{0}, \Sigma_{\mathbf{b}_k})$. Let $\mathbf{b}_i = (\mathbf{b}'_{i1}, \dots, \mathbf{b}'_{i,K-1})'$ be the $(K-1)q \times 1$ vector of random effects for subject i following a multivariate Gaussian distribution with mean zero $\mathbf{0}$ and the covariance matrix $\Sigma_{\mathbf{b}}$, $\mathbf{b}_i \sim N_{(K-1)q}(\mathbf{0}, \Sigma_{\mathbf{b}})$, defined as:

$$\Sigma_{\mathbf{b}} = \begin{pmatrix} \Sigma_{\mathbf{b}_1} & \Sigma_{\mathbf{b}_1\mathbf{b}_2} & \cdots & \Sigma_{\mathbf{b}_1\mathbf{b}_{K-1}} \\ \Sigma_{\mathbf{b}_2\mathbf{b}_1} & \Sigma_{\mathbf{b}_2} & \cdots & \Sigma_{\mathbf{b}_2\mathbf{b}_{K-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{\mathbf{b}_{K-1}\mathbf{b}_1} & \Sigma_{\mathbf{b}_{K-1}\mathbf{b}_2} & \cdots & \Sigma_{\mathbf{b}_{K-1}} \end{pmatrix}.$$

2.2. Cause-specific hazards sub-model

The survival sub-model is defined as a proportional hazards model for each cause-specific hazard (Prentice et al., 1978), incorporating a subject-specific random effect to capture the association between longitudinal and survival processes.

Let Z_i be a $l \times 1$ vector of covariates. Note that the vectors X_{ij} , W_{ij} and Z_i could consist of the same covariates. Let $T_i = \min(T_i^*, C_i)$ be the right-censored event time with C_i the censoring time of subject i and T_i^* the survival time of subject i , and $\epsilon_i = \delta_i \times D_i$ with $\delta_i = 1\{T_i^* \leq C_i\}$ the indicator of censorship and $D_i \in \{1, \dots, g\}$ indicating the failure type of subject i . We assume that the censoring mechanism is independent of the survival time. The sub-model for event d , $d = 1, \dots, g$, is specified as:

$$\begin{aligned}\lambda_d(t|\mathbf{Z}_i, u_i) &= \lim_{h \rightarrow 0} h^{-1} P(t \leq T_i < t+h, \epsilon_i = d | T_i \geq t, \mathbf{Z}_i, u_i) \\ &= \lambda_{0d}(t) \exp(\mathbf{Z}_i' \boldsymbol{\gamma}_d + \nu_d u_i),\end{aligned}\tag{2}$$

where $\lambda_d(t | \mathbf{Z}_i, u_i)$ is the instantaneous risk of failure from cause d at time t given the vector of covariates $\mathbf{Z}_i$ and the frailty u_i , $\lambda_{0d}(t)$ is an unspecified baseline hazard function for event d and $\boldsymbol{\gamma} = (\gamma_1', \dots, \gamma_g')'$ is a vector of fixed regression coefficients. In this formulation, the heterogeneities, that are not observed through the covariates $\mathbf{Z}_i$, are accounted for by the random effect u_i and the parameter ν_d . Parameters $\nu = (\nu_1, \dots, \nu_g)'$ is, therefore, the vector of coefficients of the random effects u_i with ν_1 set to 1 to ensure identifiability (Elashoff et al., 2008).

2.3. Variance-covariance matrix of random effects

Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})'$, $\mathbf{Y} = (\mathbf{Y}_1', \dots, \mathbf{Y}_n')'$, $\widetilde{\mathbf{T}}_i = (T_i, \epsilon_i)'$ and $\widetilde{\mathbf{T}} = (T_1, \epsilon_1, \dots, T_n, \epsilon_n)'$. The association between Y and $\widetilde{T}$ is modelled by assuming that the joint distribution of the random effects of the two sub-models, b_i and u_i follows a multivariate Gaussian distribution. We choose an unstructured variance-covariance matrix of the random effects as all homogeneous random effects assumptions are testable within this modelling. Furthermore, the estimated variance-covariance matrix is guaranteed to be positive definite (Huang et al., 2011) and thus, invertible. Also, the association between longitudinal process and survival data can be assessed by testing the hypothesis $\Sigma_{bu} = 0$, meaning that the joint model reduces to separate analysis if $\Sigma_{bu} = 0$. Following this assumption, all elements of Σ_{bu} , which represent the correlation between random effects b and u , together with parameters ν could measure the strength of the association between two longitudinal and competing risks processes.

$$\mathbf{a}_i = \begin{pmatrix} b_i \\ u_i \end{pmatrix} \sim N_{(K-1)q+1} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_b & \Sigma_{bu}' \\ \Sigma_{bu} & \sigma_u^2 \end{pmatrix} \right).\tag{3}$$

3. Likelihood function, Estimation and Inference

The inference proceeds under two main assumptions that given the covariates and the random effects a_i , 1) the longitudinal outcomes process is independent of the competing risks survival data and also 2) the longitudinal outcomes are independent of each other.

Let $\boldsymbol{\Psi} = (\alpha, \beta, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\Sigma}, \lambda_{01}(t), \dots, \lambda_{0g}(t))$ be the vector of all parameters in (1) and (2), where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\mathbf{a}_i$ defined in (3). We assume that the longitudinal outcomes are independent of the competing risks survival data, conditional on covariates and random effects. The joint distribution of $(Y, \widetilde{T})$ is completely determined by $f(Y | \mathbf{a}, \boldsymbol{\Psi})$, $f(\widetilde{T} | \mathbf{a}, \boldsymbol{\Psi})$ and $f(\mathbf{a} | \boldsymbol{\Psi})$ where $f(\cdot)$ stands for the

probability density function and $\mathbf{a} = (b, u)'$ represents the vector of random effects of the two sub-model. The observed-data likelihood function for Ψ , conditional on the observed data $(Y_i, \tilde{T}_i)$ for $i = 1, \dots, n$, is

$$\begin{aligned} L(\Psi|Y, \tilde{T}) &\propto \prod_{i=1}^n f(Y_i, \tilde{T}_i|\Psi) \\ &= \prod_{i=1}^n \int_{\mathbf{a}} f(Y_i|\tilde{T}_i, \mathbf{a}, \Psi) f(\tilde{T}_i|\mathbf{a}, \Psi) f(\mathbf{a}|\Psi) d\mathbf{a}. \end{aligned} \quad (4)$$

Since Y and $\tilde{T}$ are independent given the covariates under assumption 1, the equation (4) can be formulated as:

$$\begin{aligned} L(\Psi|Y, \tilde{T}) &\propto \prod_{i=1}^n \int_{\mathbf{a}} f(Y_i|\mathbf{a}, \Psi) f(\tilde{T}_i|\mathbf{a}, \Psi) f(\mathbf{a}|\Psi) d\mathbf{a} \\ &= \prod_{i=1}^n \int_{\mathbf{a}} \left[\prod_{j=1}^{m_i} \prod_{h=1}^K \{\pi_{ijh}\}^{I(Y_{ij}=h)} \right] \left\{ \prod_{d=1}^g \lambda_d(T_i|\mathbf{Z}_i, \mathbf{u}, \gamma_d, \nu_d)^{I(\epsilon_i=d)} \right\} \\ &\quad \times \exp \left[- \int_0^{T_i} \left\{ \sum_{d=1}^g \lambda_d(t|\mathbf{Z}_i, u_i, \gamma_d, \nu_d) \right\} dt \right] \\ &\quad \times \frac{1}{\sqrt{(2\pi)^{(K-1)q+1} |\Sigma|}} \exp \left(- \frac{1}{2} \mathbf{a}' \Sigma^{-1} \mathbf{a} \right) d\mathbf{a}. \end{aligned} \quad (5)$$

3.1. Parameter estimation

Following Wulfsohn and Tsiatis (1997), we used parametric Maximum Likelihood Estimation (MLE) approach to estimate all parameters, except the baseline hazard functions λ_{0d} which were estimated using non-parametric maximum likelihood. Maximizing the observed-data likelihood, equation (5), in the presence of integration over random effects $\mathbf{a}_i$ is difficult. As a result and for simplification, the complete-data likelihood conditional on the random effects will be considered.

$$\begin{aligned} L(\Psi|Y, \tilde{T}, \mathbf{a}) &\propto \prod_{i=1}^n \left[\prod_{j=1}^{m_i} \prod_{k=1}^K \{\pi_{ijk}\}^{I(Y_{ij}=k)} \right] \\ &\quad \times \left\{ \prod_{d=1}^g \lambda_d(T_i|\mathbf{Z}_i, u) ^{I(\epsilon_i=d)} \right\} \\ &\quad \times \exp \left[- \int_0^{T_i} \left\{ \sum_{d=1}^g \lambda_d(t|\mathbf{Z}_i, u_i) \right\} dt \right] \\ &\quad \times \frac{1}{\sqrt{(2\pi)^{(K-1)q+1} |\Sigma|}} \exp \left(- \frac{1}{2} \mathbf{a}_i' \Sigma^{-1} \mathbf{a}_i \right). \end{aligned} \quad (6)$$

The EM algorithm is employed to obtain the MLE of Ψ . This algorithm iterates between E-steps and

M-steps. The E-step computes the expected logarithm of the complete-data likelihood conditional on the observed data and the current estimates of the parameters, and the M-step estimates the new parameter by maximizing the expected log-likelihood mentioned above. The E-step requires multiple integrations over the random effects. As the assumed distribution for the random effects is normal, the Gauss-Hermite quadrature can be used to approximate the integrals, by a summation of a specified number of quadrature points Q for each dimension. In the maximization step, each cumulative baseline hazard function for cause d , $H_{0d}(t)$, is assumed to be a step function with jumps at observed event times due to cause d , $d = 1, \dots, g$. These two steps are repeated until the convergence criteria are met. More details on the EM algorithm are given in the Appendix.

Estimation of the standard errors could be based on a profile likelihood approach (Rizopoulos, 2012) in which some parameters of the likelihood function are eliminated by replacing them with their MLE as a function of the remaining parameters. This method allows the calculation of a smaller size matrix to be inverted for the Standard Errors estimation.

To estimate the Standard Errors, we followed the approach applied in Elashoff et al. (2007). The parameter vector $\Psi = (\alpha, \beta, \gamma, \nu, \Sigma, \lambda_{01}(t), \dots, \lambda_{0g}(t))$ was split into two components, $\Omega = (\alpha, \beta, \gamma, \nu, \Sigma)$ and $\Lambda = (\lambda_{01}(t), \dots, \lambda_{0g}(t))$. The variance-covariance matrix of Ω is approximated by inverting the empirical Fisher information obtained from the profile likelihood where the baseline hazards function have been profiled out. The observed information matrix of Ω is approximated by

$$\sum_{i=1}^n l^{(i)}(\hat{\Omega}|Y, \tilde{T}) l^{(i)}(\hat{\Omega}|Y, \tilde{T})', \quad (7)$$

given that $l^{(i)}(\hat{\Omega}|Y, \tilde{T})$ is the observed score vector from the profile likelihood on the i -th subject evaluated at $\hat{\Omega}$.

A program was developed in C programming language to estimate the parameters and the standard errors. To reduce the calculation time, OpenMP (Dagum and Menon, 1998) was used. OpenMP is a specification for a set of compiler directives, library routines, and environment variables that can be used to specify high-level parallelism in Fortran and C/C++ programs.

3.2. Estimation of marginal membership probabilities in the longitudinal sub-sample

In a mixed-effects multinomial logistic model with quantitative predictors, it is of interest to plot the estimated marginal probabilities. As explained in Hedeker (2003), the marginal probabilities, $\pi_k^m, k = 1, \dots, K$ can be estimated in two steps. In the first step, the *subject-specific* probabilities are calculated by replacing specific values of covariates and estimated parameters, $\hat{\Psi}$, in the equation (1). These probabilities are functions of the subject-specific random effects $\mathbf{b}_i$, $\hat{\pi}_k^{ss}(b)$. Then the marginal probabilities can be obtained by integrating over the random effects distribution of these subject-specific probabilities:

$$\hat{\pi}_k^m = \int_{\mathbf{b}} \hat{\pi}_k^{ss}(\mathbf{b}) f(\mathbf{b}) d\mathbf{b}, \quad \text{for } k = 1, \dots, K, \quad (8)$$

where $f(b)$ denotes the probability density function of the random effects. Since we assume that the random effects are normally distributed, we can resolve the integration by the numerical quadrature techniques, such as Gauss-Hermite quadrature. The advantage of this new marginal membership probability compared to those obtained directly from a mixed-effects model is that the new ones are accounted for competing risks, which should not be ignored. 5

Pointwise confidence intervals of these marginal probabilities can be estimated employing delta method as follows. Let $\Phi = (\phi_1, \dots, \phi_P)$ be the parameter vector of the longitudinal sub-model. Let Φ^0 be the true value of Φ and $\hat{\Phi}$ an estimation of Φ obtained by the MLE. Asymptotic property of the MLEs, provided in [Zeng and Cai \(2005\)](#), shows that as $n \rightarrow \infty$:

$$\mathbf{I}(\Phi^0)^{1/2}(\hat{\Phi} - \Phi^0) \xrightarrow{D} \mathcal{N}_P(\mathbf{0}, \mathbf{I}_P), \quad (9)$$

where $\mathbf{I}(\Phi^0)$ is the Fisher information matrix for Φ^0 , approximated by the equation (7), $\mathbf{I}_P$ is the identity matrix of rank P and $\xrightarrow{D}$ denotes convergence in distribution. Thus for large n , $\hat{\Phi} \stackrel{D}{\sim} \Phi^0 + \mathbf{I}(\Phi^0)^{-1} \mathcal{N}_P(0, \mathbf{I}_P)$, where $\stackrel{D}{\sim}$ is the shorthand for *is approximately distributed as*. Following *delta method*, the marginal probability π_k^m , $k = 1, \dots, K$ is approximately distributed as: 10

$$\pi_k^m[\hat{\Phi}] - \pi_k^m[\Phi^0] \stackrel{D}{\sim} \mathcal{N}_P(\mathbf{0}, \partial_{\Phi} \pi_k^m(\Phi^0) \mathbf{I}^{-1}(\Phi^0) (\partial_{\Phi} \pi_k^m(\Phi^0))'), \quad (10)$$

with $\partial_{\Phi} \pi_k^m$ being the derivative of the function π_k^m with respect to the parameter vector Φ .

We can apply this result to the multivariate case where 15

$$\mathbf{\Pi}(\Phi) = \begin{pmatrix} \pi_1^m(\Phi) \\ \pi_2^m(\Phi) \\ \vdots \\ \pi_K^m(\Phi) \end{pmatrix}. \quad (11)$$

Then

$$\mathbf{\Pi}(\hat{\Phi}) - \mathbf{\Pi}(\Phi^0) \stackrel{\mathcal{D}}{\sim} \mathcal{N}(\mathbf{0}, \partial_{\Phi} \mathbf{\Pi}(\Phi^0) \mathbf{I}^{-1}(\Phi^0) (\partial_{\Phi} \mathbf{\Pi}(\Phi^0))'), \quad (12)$$

with

$$\partial_{\Phi} \mathbf{\Pi}(\Phi) = \begin{pmatrix} (\nabla \pi_1^m(\Phi))' \\ (\nabla \pi_2^m(\Phi))' \\ \vdots \\ (\nabla \pi_K^m(\Phi))' \end{pmatrix}, \quad (13)$$

and

$$\nabla \pi_{\mathbf{k}}^m(\Phi) = \begin{pmatrix} \partial \pi_k^m(\Phi) / \partial \phi_1 \\ \partial \pi_k^m(\Phi) / \partial \phi_2 \\ \vdots \\ \partial \pi_k^m(\Phi) / \partial \phi_P \end{pmatrix}. \quad (14)$$

Given the equation (8) and by the so-called *differentiation under the integral sign*, the gradient of the k -th marginal probability is

$$\frac{\partial \pi_k^m(\Phi)}{\partial \phi_p} = \frac{\partial \int_{\mathbf{b}} \hat{\pi}_k^{ss}(\mathbf{b}) f(\mathbf{b}) d\mathbf{b}}{\partial \phi_p} = \int_{\mathbf{b}} \frac{\partial}{\partial \phi_p} (\hat{\pi}_k^{ss}(\mathbf{b}) f(\mathbf{b})) d\mathbf{b}, \quad \text{for } p = 1, \dots, P. \quad (15)$$

5 To obtain a simultaneous 95% confidence interval, we used bootstrap. To estimate this simultaneous confidence interval, we chose 100 sub-samples from the initial sample. The curve of the marginal probabilities of each is calculated separately for each sub-sample. The simultaneous 95% confidence interval for each marginal probability is then the region that contains 95% of them.

3.3. Joint modelling on large-scale data

10 The joint modelling of longitudinal data and survival outcomes becomes almost out of reach in large-scale data. The motivating sample in this study is substantially large compared to the existing applications of joint models on real databases. We, therefore, propose an approach mimicking a meta-analysis. However, we note that the proposed approach is not a standard meta-analysis, but it uses some characteristics of this

method. The rationale for this meta-analysis approach is inspired by the work of [Molenberghs et al. \(2005\)](#), in a different context and estimation methods, and is transferred to our setting.

In this approach, the initial sample is randomly stratified into S equal size sub-samples. The parameters of the joint model were estimated separately on each randomly stratified sub-sample, $\hat{\Psi}_1, \dots, \hat{\Psi}_S$, and then were combined. As the sub-samples are chosen randomly, the combination of estimations was done by taking the mean of the estimated parameters and therefore, the pooled estimation of the parameter is obtained by $\hat{\Psi}^* = \frac{1}{S} \sum_{s=1}^S \hat{\Psi}_s$ and their variance is estimated by $var(\hat{\Psi}^*) = \frac{1}{S^2} \sum_{s=1}^S var \hat{\Psi}_s$.

4. Simulation study

4.1. Design

We performed a simulation study to evaluate the performance of the proposed joint model, considering an association between longitudinal and survival processes in the presence of various censoring rates. In particular, two different scenarios with 500 samples each and imposed administrative censoring rates of 50% (scenario I) and 80% (scenario II) were considered. For both scenarios, first, the longitudinal data were simulated from sub-model (1) with 9 observation times and three categories for the categorical longitudinal responses, just as in the motivating example. Then, the survival data were generated from sub-model (2) with two competing events. In the end, to keep the desired censoring rates, the longitudinal measurements were censored at the simulated censoring/survival times.

For the longitudinal part, the data were simulated using sub-model (1) for 500 individuals. We assumed a GLMM with the linear predictor $\eta_{ijk} = \alpha_k + \mathbf{X}_{ij}'\beta_k + \mathbf{W}_{ij}'\mathbf{b}_{ik}$ where $\alpha = (\alpha_1, \alpha_2)' = (-1, 0)$ is the vector of intercepts and $K = 3$ was considered as the reference category. We considered the vector of predictors for the fixed effects as $\mathbf{X}_{ij} = (x_i, t_{ij})'$ where t_{ij} is the visit time and $x_i \sim \text{Bernoulli}(0.5)$ is the sex indicator with $\beta_1 = (\beta_{11}, \beta_{12})' = (0.2, 0.3)$ the vector of fixed effects parameters for the first category and $\beta_2 = (\beta_{21}, \beta_{22})' = (0.1, 0.5)$ the vector of fixed effects parameters for the second category. In both scenarios, at most 9 visit times were considered, meaning that less than 9 measures can be observed for an individual due to occurrence of an event. We aimed to have administrative censoring, therefore, in scenario I, the administrative censoring time is set to 0.7 and so the exact visit times are $t_{ij} = 0, 0.0875, \dots, 0.7$ and in scenario II, the administrative censoring time is set to 0.15 and the exact times are $t_{ij} = 0, 0.01875, \dots, 0.15$. These censoring times were chosen such that at the end an administrative censoring rate of 50% for scenario I and 80% for scenario II should have been obtained. We fixed the The vector of predictors for random effects was set to $W_{ij} = 1$, so that b_{ik} is the random intercept for subject i and category k . To insure the existence of an association between the longitudinal and time-to-event data, random effects b_i and u_i were simulated from a multivariate Gaussian distribution using the *mvrnorm* function of the R software, setting the mean vector to 0 and the covariance matrix Σ with $\Sigma_{b_1} = \Sigma_{b_2} = 1$, $\sigma_u^2 = 0.5$, $\Sigma_{b_1 b_2} = -0.5$

Table 1: Description of simulation scenarios

	Scenario I		Scenario II	
	Mean	Proportion	Mean	Proportion
Repeated values	8.2		6.4	
<i>Categories ($K = 3$)</i>				
1		21.47		21.99
2		43.50		45.71
3		35.03		32.20
<i>Event Type</i>				
0		82.73		50.85
1		10.33		27.99
2		6.94		21.16

and $\Sigma_{b_1u} = \Sigma_{b_2u} = -0.2$. All probabilities $\pi_{ijk} = P(Y_{ij} = k)$ were calculated and individuals' longitudinal categories using a multinomial experiment based on the calculated probabilities were generated (*rmultinom* function).

Two competing risks were simulated with scheme proposed in Beyersmann et al. (2009) with constant baseline hazards $\lambda_{01} = 0.15$, $\lambda_{02} = 0.25$, and the covariates vector $\mathbf{Z}_i = (z_i, x_i)'$ with $z_i \sim N(2, 1)$. Vector of the fixed regression coefficients was set to $\gamma_1 = (\gamma_{11}, \gamma_{12})' = (0.8, -1)$ for the first event and $\gamma_2 = (\gamma_{21}, \gamma_{22})' = (0.5, -1)$ for the second event and coefficient of the random effects u_i was taken to be $\nu_2 = 0.5$. Table 1 shows the rate of each event type in our simulation.

4.2. Results

Table 2 and Table 3 summarize the simulation results on 500 replications with the sample size $n = 500$. The parameters of the proposed joint model and the separate analysis, consisting of a nominal GLMM for longitudinal data and a Gaussian frailty model for each cause-specific hazard ratio, were estimated respectively. The estimated parameters, simulated bias, standard errors (SE), 95% confidence interval coverage probability (CP), root-mean-square error (RMSE) and the comparison of the mean square errors (MSE) of both joint and separate analysis are presented in Table 2 and Table 3.

The joint model provided almost unbiased estimates for all the parameters in both scenarios. The coverage probabilities reach their nominal value of 95%. On the contrary, the separate analysis produces much larger bias in the time trend β_{12} and β_{22} for the longitudinal sub-model. The non-ignorable missing values after death cannot be taken into account in the nominal GLMM alone, and therefore produces bias in the estimated time trend (β_{12} and β_{22}). The standard error of the random effect coefficient, ν_2 , is poorly estimated by the separate analysis. We observe that the joint model provides us much more accurate estimates of the coefficient of the random effect in the survival sub-model, ν_2 . The ratio of the MSE of the separate analysis (MSE_S) over the MSE of the joint model (MSE_J) for the parameter ν_2 is much larger

than 1, which indicates that combining longitudinal data and survival outcomes improves the estimates of the survival sub-model.

Overall, in the presence of an association between longitudinal outcomes and competing risks data, the joint model performs better than the separate analysis, as the MSE_J is smaller than or equal to the MSE_S , or the ratio of MSE_S over the MSE_J is larger than or close to one. Finally, we observe that in both joint models and separate analysis, the estimation of the variance of the random effect in the survival sub-model, σ_u^2 , is biased. Estimation of the variance of the survival random effect, σ_u^2 , requires a larger sample size.

To investigate the impact of misspecification of the normality assumption on the performance of the joint model, a second set of simulations were carried out, where the random effects were generated from a multivariate t -distribution with 5 degrees of freedom. The simulated data were then analysed based on the assumption of the joint model, meaning that the random effects are assumed to be distributed multivariate normally (results not shown). Like the previous simulation set, the separate analysis identifies much larger bias in the time trend β_{12} and β_{22} and also ν_2 . Comparing the two sets of simulations, it seems that the impact of the model misspecification on the standard errors of the estimates in the joint model is not large. Comparison between the mean square errors of the joint and the separate analysis shows that overall, the joint model produces more accurate estimations and performs better compared to the separate analysis.

Table 2: Comparison of the joint model and the separate analyses (sample size = 500, 50% administrative censoring)

Parameter	True	Separate Analysis Estimate	Bias	SE	CP	RMSE	Joint Analysis Estimate	Bias	SE	CP	RMSE	MSE_S/MSE_J
<i>Longitudinal</i>												
Fixed effects												
α_1	-1.000	-0.975	-0.025	0.135	0.960	0.138	-1.000	-0.000	0.138	0.957	0.138	1.000
α_2	0.000	0.016	-0.016	0.111	0.968	0.112	0.000	-0.000	0.112	0.975	0.112	1.000
β_{11}	0.200	0.159	0.041	0.151	0.962	0.156	0.191	0.009	0.153	0.953	0.153	1.043
β_{12}	0.300	0.410	-0.110	0.243	0.939	0.267	0.308	-0.008	0.252	0.957	0.252	1.120
β_{21}	0.100	0.081	0.019	0.133	0.960	0.134	0.099	0.001	0.134	0.961	0.134	1.001
β_{22}	0.500	0.600	-0.100	0.205	0.941	0.228	0.516	-0.016	0.213	0.955	0.214	1.143
Random effects												
Σ_{b_1}	1.000	0.986	0.014	0.195	0.954	0.195	1.003	-0.003	0.198	0.959	0.198	0.969
Σ_{b_2}	1.000	0.979	0.021	0.159	0.960	0.161	1.004	-0.004	0.161	0.967	0.161	0.992
<i>Survival</i>												
Fixed effects												
γ_{11}	0.800	0.833	-0.033	0.109	0.876	0.114	0.815	-0.015	0.110	0.885	0.111	1.058
γ_{12}	-1.000	-1.028	0.028	0.198	0.884	0.200	-0.997	-0.003	0.198	0.908	0.198	1.017
γ_{21}	0.500	0.522	-0.022	0.123	0.899	0.125	0.510	-0.010	0.119	0.736	0.119	1.108
γ_{22}	-1.000	-1.076	0.076	0.236	0.926	0.248	-1.038	0.038	0.226	0.926	0.230	1.171
Random effects												
ν_2	0.500	0.679	-0.179	1.296	0.867	1.308	0.498	0.002	0.494	0.818	0.494	7.008
σ_u^2	0.500	0.770	-0.270	0.313	0.644	0.414	0.543	-0.043	0.324	0.721	0.326	1.605
<i>Covariance</i>												
$\Sigma_{b_1 b_2}$	-0.500	-0.505	0.005	0.125	0.979	0.125	-0.503	0.003	0.127	0.973	0.127	0.974
$\Sigma_{b_1 u}$	-0.200	-	-	-	-	-	-0.196	-0.004	0.174	0.883	0.174	-
$\Sigma_{b_2 u}$	-0.200	-	-	-	-	-	-0.202	0.002	0.151	0.867	0.151	-

Table 3: Comparison of the joint model and the separate analyses (sample size = 500, 80% administrative censoring)

Parameter	True	Separate Analysis Estimate	Bias	SE	CP	RMSE	Joint Analysis Estimate	Bias	SE	CP	RMSE	MSE_S/MSE_J
<i>Longitudinal</i>												
Fixed effects												
α_1	-1.000	-0.997	-0.003	0.124	0.969	0.124	-1.003	0.003	0.125	0.975	0.125	0.984
α_2	0.000	0.000	-0.000	0.104	0.981	0.104	-0.005	0.005	0.104	0.981	0.105	0.981
β_{11}	0.200	0.191	0.009	0.136	0.935	0.137	0.203	-0.003	0.137	0.944	0.137	1.000
β_{12}	0.300	0.456	-0.156	0.963	0.956	0.975	0.309	-0.009	0.974	0.952	0.974	1.002
β_{21}	0.100	0.095	0.005	0.121	0.979	0.121	0.106	-0.006	0.122	0.979	0.122	0.984
β_{22}	0.500	0.629	-0.129	0.806	0.948	0.817	0.512	-0.012	0.815	0.944	0.815	1.005
Random effects												
Σ_{b_1}	1.000	0.993	0.007	0.166	0.965	0.166	1.000	0.000	0.168	0.971	0.168	0.976
Σ_{b_2}	1.000	0.993	0.007	0.137	0.973	0.137	0.999	0.001	0.137	0.975	0.137	1.000
<i>Survival</i>												
Fixed effects												
γ_{11}	0.800	0.873	-0.073	0.165	0.919	0.181	0.849	-0.049	0.163	0.919	0.170	1.133
γ_{12}	-1.000	-1.077	0.077	0.331	0.939	0.340	-1.061	0.061	0.326	0.948	0.332	1.049
γ_{21}	0.500	0.532	-0.032	0.244	0.969	0.247	0.509	-0.009	0.207	0.954	0.207	1.424
γ_{22}	-1.000	-1.109	0.109	0.476	0.985	0.488	-1.087	0.087	0.418	0.979	0.427	1.306
Random effects												
ν_2	0.500	0.443	0.057	2.866	1.000	2.866	0.399	0.101	0.964	0.971	0.969	8.748
σ_u^2	0.500	0.954	-0.454	0.416	0.687	0.615	0.714	-0.214	0.380	0.644	0.437	1.980
<i>Covariance</i>												
$\Sigma_{b_1 b_2}$	-0.500	-0.501	0.001	0.108	0.979	0.108	-0.498	-0.002	0.109	0.975	0.109	0.982
$\Sigma_{b_1 u}$	-0.200	-	-	-	-	-	-0.207	0.007	0.259	0.886	0.259	-
$\Sigma_{b_2 u}$	-0.200	-	-	-	-	-	-0.197	-0.003	0.224	0.877	0.224	-

5. Application to the Cosmop-DADS database

The Cosmop-DADS population is a sample of the French population, composed of people for whom vital status and date of death are available, employed at least once as salaried workers in the semi-public and private sectors between 1976 and 2002. All individuals, 957 299 men, and 798 291 women were followed-up to 2002, and the administrative censoring date was set on 31 December 2002.

Initially, the DADS covered five major classes: craftsmen and trade-related workers, upper class, intermediary occupations, clerk class and manual workers class. Any occupations in the craftsmen and trade-related workers class were excluded from the study scope because those in DADS are not representative of this class in the general population. Also, regional and local authorities were not fully covered by DADS declarations before 1987. Therefore, any occupations of this type were excluded from the professional scope of this study. Those professional episodes done by an individual in careers not covered by the professional scope of this study are considered as missing professional episodes. Besides, as inactivity or retirement are not covered in the DADS panel, missing professional episodes contain these two instances as well. Missing data is a common issue in almost every longitudinal study. Imputation methods were introduced in the literature to address this problem, by making assumptions on the missing data mechanism (Rubin, 1987). However, in the Cosmop-DADS database, missing professional episodes resulted from different scenarios. Considering that these missing professional episodes are a mixture of working outside the study scope, being inactive or retired, and in the absence of other complementary databases, building a sound imputation model is not feasible.

We illustrate an application of the proposed joint model using a large sample size of the Cosmop-DADS database¹. In Karimi et al. (2015), the considered population has a total of 22% and 30% missing professional episodes for men and women, respectively. Fifty-two percent of men and 61% of women had at least 1 missing professional episode. The issue of missing data is a significant problem and will need to be considered in future works. The proposed model in the present study cannot handle this type of missing data that is not caused by death since the association between socio-professional category and health is bidirectional, and deciphering missing data not caused by death from the trajectory is not a simple task. For these kinds of missing values, an external source of data could help to build an imputation strategy. Due to the lack of other sources of data and since treating these missing data is not the objective of this study, we considered a subset of this population that contains all individuals without missing professional episodes. For having a reasonable censoring rate and proper professional trajectories (with adequate length and number of transitions between professional categories), we considered individuals followed-up at least for 10 years, at the age of 34 or more at the entrance and with a transition rate higher than 0.1. Here,

¹It is worthy of note that the results obtained here are not comparable with those in Karimi et al. (2015) as the considered populations in these two studies have not the same characteristics.

Table 4: Description of the meta-analysis sub-samples

	Sub-sample 1	Sub-sample 2	Sub-sample 3	Sub-sample 4	Sub-sample 5	Sub-sample 6	Sub-sample 7	Sub-sample 8	Sub-sample 9	Sub-sample 10	Total (10 Sub-samples)
Men	1184	1184	1184	1184	1184	1184	1184	1184	1184	1196	11852
Death	101	101	101	101	101	101	101	101	101	111	1020
Cancer	48	48	48	48	48	48	48	48	48	56	488
Other causes	53	53	53	53	53	53	53	53	53	55	532
Women	982	982	982	982	982	982	982	982	982	988	9827
Death	24	24	24	24	24	24	24	24	24	26	242
Cancer	14	14	14	14	14	14	14	14	14	15	141
Other causes	10	10	10	10	10	10	10	10	10	11	101

the transition rate is defined as the ratio of the number of changes between professional categories and the total number of follow-up years. For instance, if an individual was working in total for 20 years, first in the manual workers class for five years and then in the clerk class for 15 years, his transition rate is $1/20 = 0.05$. Finally, there are 11 852 men and 9827 women included in the analysis, with 94% right censoring rate. We note that considering this particular subset requires careful interpretations, and therefore the results may not be extrapolated to the general population.

5.1. Results

The data that were analysed in this study was comprised of 20 000 subjects. Following [Molenberghs et al. \(2005\)](#), the size of sub-samples should not be too small. Therefore, we decided to divide the initial sample into 10 randomly stratified sub-samples with 2000 individuals. The sub-samples were randomly stratified by cause of death to keep the same distribution of causes of death in all sub-samples. Table 4 describes these sub-samples.

Two competing events, cancer mortality and death by other causes were considered in this analysis. By using profession as a time-dependent variable in cause-specific Cox proportional hazards model, close estimates were found for the upper class and the intermediary occupations ([Karimi et al., 2015](#)), thus, we combined these two categories. Three broad categories for profession were considered: upper class, clerk class, and manual workers class, where the upper class was used as the reference category. The longitudinal sub-model was adjusted for sex and age (mean-centered). The survival sub-model was adjusted for sex. To take into account the left-truncation, an individual's age at entry was also added as a covariate into the survival sub-model. The following joint model is fitted to the Cosmop-DADS data:

$$\begin{cases} \eta_{ijk} = \alpha_k + \text{Age}_{ij} \times \beta_{k1} + \text{Sex}_{ij} \times \beta_{k2} + \text{Sex}_{ij} \times \text{Age}_{ij} \times \beta_{k3} + b_{ik}, \\ \lambda_d(t|\mathbf{Z}_i, u_i) = \lambda_{0d}(t) \exp(\text{Sex} \times \gamma_{d1} + \text{Age at entry} \times \gamma_{d2} + \nu_d u_i). \end{cases} \quad (16)$$

For comparison, we also carried out the separate analysis of the longitudinal nominal data and competing risks data meaning that the random effects b_{ik} 's and u_i are independent.

The pooled estimates of the proposed joint model and the separate analysis of these 10 randomly stratified

sub-samples are summarized in Table 5. Comparing the results of the joint and separate analysis, in the longitudinal sub-model, differences for working in the manual workers are found to be small, and bigger differences are found for working in the clerk class. However, strong differences are found in the survival sub-model results between the joint and separate analysis. Focusing on the results obtained from the joint model analysis, in the longitudinal sub-model, the estimated intercept, α_2 , has the following interpretation: The estimated log-odds of the clerk class versus the upper class in the reference group, i.e., men at the mean age, 48 years, is 1.044 (95% CI = [1.006, 1.082]). As highlighted earlier, the exponential of a fixed-effect coefficient in the longitudinal sub-model is interpreted as the increase in odds of being in the modality of interest versus the reference modality resulting from a one-unit increase in that covariate, given observing the same random effect and holding the other covariates constant. Therefore, the estimated sex effect in the clerk class versus the upper class, -1.861, is interpreted as follows: Holding age constant, for those with the same random effect, among men the odds of working in the clerk class rather than in the upper class are $0.155 = \exp(-1.861)$ (95% CI = [0.147, 0.163]) times (84.5% lower than) the odds among women. This means that men are less likely to work in the clerk class versus the upper class than women. No difference was found between men and women for working in the manual workers class versus the upper class.

For women, with the same random effect, a one-year increase in age multiplies the odds of working in the manual workers class rather than working in the upper class by $0.914 = \exp(-0.090)$ (95% CI = [0.912, 0.915]), i.e., decreases this odds by 8.6%. On the other hand, among men, with the same random effect, a one-year increase in age multiplies the odds of working in the manual workers class rather than working in the upper class by $0.926 = \exp(-0.090 + 0.013)$ (95% CI = [0.922, 0.930]), i.e., decreases this odds by 7.4%. The relatively large $\hat{\sigma}_{11}^2 = 5.879$ and $\hat{\sigma}_{22}^2 = 3.325$, random effects of the longitudinal sub-model, reflects strong positive associations between the repeated values.

For illustrating the difference between men and women over time, it is of interest to plot the estimated marginal membership probabilities of each professional category. As explained earlier, each marginal probability and its 95% CI were calculated by integrating the subject-specific probabilities over the random effects distribution. Figure 1 plots the estimated probabilities of working in the upper class, working in the clerk class or working in the manual workers class, for men and women, and their 95% pointwise and simultaneous confidence intervals, as a function of age², based on the parameter estimates of both joint model and separate analysis.

Estimated marginal probabilities based on the joint and separate analysis are slightly different for all professional categories in men. The difference between these probabilities is much stronger for women. These differences can be explained by the fact that based on the joint model parameter estimations, marginal probabilities are accounted for the informative censoring as well. As shown in Figure 1, the probability

²At each age, sum of the estimated marginal membership probabilities is equal to 1.

Table 5: Pooled estimates of 10 stratified sub-samples

(a) Longitudinal sub-model

	Joint model		Separate model	
	Estimate	[95% CI]	Estimate	[95% CI]
Upper class	Reference category			
Manual workers class				
Intercept	-0.113	[-0.151,-0.075]	-0.110	[-0.157,-0.064]
Sex (Men=1, Women=0)	0.006	[-0.051, 0.063]	-0.074	[-0.127,-0.022]
Age	-0.090	[-0.092,-0.088]	-0.088	[-0.090,-0.085]
Sex by Age	0.013	[0.011, 0.016]	0.011	[0.008, 0.013]
Clerk class				
Intercept	1.044	[1.006, 1.082]	0.658	[0.615, 0.701]
Sex (Men=1, Women=0)	-1.861	[-1.914,-1.808]	-1.527	[-1.576,-1.477]
Age	-0.040	[-0.042,-0.038]	-0.040	[-0.048,-0.044]
Sex by Age	-0.018	[-0.020,-0.016]	-0.012	[-0.014,-0.010]

(b) Survival sub-model

	Joint model		Separate model	
	CSHR	[95% CI]	CSHR	[95% CI]
<i>Fixed effects</i>				
Cancer				
Sex	3.803	[3.103, 4.662]	2.373	[1.951, 2.887]
Age at entry	1.247	[1.222, 1.273]	1.178	[1.156, 1.200]
Other causes				
Sex	4.574	[3.672, 5.697]	3.351	[2.695, 4.168]
Age at entry	1.164	[1.143, 1.186]	1.163	[1.142, 1.184]
	Estimate	[95% CI]		
<i>Random effect</i>				
$\hat{\nu}_2$	0.210	[-0.027, 0.448]	0.251	[-0.394, 0.897]

(c) Estimated variance-covariance matrix of random effects

$$\begin{aligned}
\text{Joint model } \hat{\Sigma} &= \begin{pmatrix} \mathbf{5.879} [5.669, 6.089] & \mathbf{2.817} [2.694, 2.941] & 0.172 [-0.124, 0.468] \\ \mathbf{2.817} [2.694, 2.941] & \mathbf{3.325} [3.208, 3.442] & -0.050 [-0.278, 0.178] \\ 0.172 [-0.124, 0.468] & -0.050 [-0.278, 0.178] & \mathbf{4.936} [4.552, 5.320] \end{pmatrix} \\
\text{Separate model } \hat{\Sigma} &= \begin{pmatrix} \mathbf{5.964} [5.750, 6.178] & \mathbf{2.856} [2.733, 2.979] & 0 \\ \mathbf{2.856} [2.733, 2.979] & \mathbf{3.370} [3.253, 3.487] & 0 \\ 0 & 0 & \mathbf{0.728} [0.530, 0.926] \end{pmatrix}
\end{aligned}$$

bold indicates p-value < 0.05

Figure 1: Estimated marginal membership probabilities in professional category based on the results of the joint and separate models. The 95% pointwise and simultaneous confidence intervals are obtained from the estimated parameters of the joint model.

of working in the clerk class decreases for men over time, but stays constant over time for women. The probability of working in the clerk class is markedly higher among women than among men. There is a more pronounced effect of age in men than in women in the manual working class.

Results of the survival sub-model in a joint model approach indicate significant effects of both covariates, sex and age at entry for mortality by cancer and by other causes. The results show that men had an increased cause-specific mortality compared to women, CSHRs = 3.803 [3.103, 4.662] and 4.574 [3.672, 5.697], respectively, for mortality from cancer and other causes.

A multivariate Wald test was used to test whether Σ_{bu} is different from zero which corresponds to the null hypothesis $H_0 : \sigma_{b_1u} = \sigma_{b_2u} = 0$ (p -value= 0.56). In both joint and separate analysis, a Wald test was performed to investigate if ν_2 is significantly away from 0 (p -value= 0.08 for joint model and p -value= 0.44 for separate model). Even though the Wald tests are not significant, it is worth detailing the interpretation of the sign and the magnitude of the estimates of $\hat{\sigma}_{b_1u}$, $\hat{\sigma}_{b_2u}$ and $\hat{\nu}_2$. Notably, in a joint modelling framework a negative sign of the estimated covariance between the random intercept b_{i2} in the longitudinal sub-model and the random effect u_i in the survival sub-model, $\hat{\sigma}_{b_2u} = -0.050$, together with the positive $\hat{\nu}_2 = 0.210$, coefficient of the random effect in the survival sub-model, highlights that there are lower cause-specific hazards for both death by cancers and death by other causes for individuals working in the clerk class as compared to the upper class. On the contrary, a positive sign of the estimated covariance $\hat{\sigma}_{b_1u} = 0.172$ and the positive $\hat{\nu}_2$ highlights that there are higher cause-specific hazards for both death from cancer and death from other causes for individuals working in the manual workers class as compared to the upper class. In a separate framework, the coefficient of the random effect in the survival model, $\hat{\nu}_2 = 0.251$, indicates that individual's risk of other cause mortality is smaller than individual's risk of cancer mortality. Finally, as shown in Table 5, the estimated variance of the random effect u_i in the separate model is markedly smaller than that of the joint modelling framework, which suggests that taking into account individuals' professional trajectories will result in observing stronger variations in mortality across cancer and other causes mortality. These findings support the hypothesis of an association between socio-professional trajectories and mortality.

6. Discussion

In this work, we propose an extension of the joint modelling of longitudinal outcomes and competing risks data that handles longitudinal nominal data. To model the repeated observations collected in longitudinal studies, by taking into account their correlation, both GLMMs and marginal models as Generalized Estimating Equations (GEE) (Liang and Zeger, 1986) have been proposed in the literature. However, these analyses are limited in that they do not consider simultaneously longitudinal outcomes and time-to-event data. Conducting a joint analysis of longitudinal outcomes and time-to-event data allows modelling these two types of data together. Also, the joint model gives the possibility to estimate the marginal membership

probabilities accounting for informative censoring, by integrating the estimated subject-specific probabilities, for specific covariates values, over the random effects distribution. For the sake of simplicity and given the Cosmop-DADS set also, we considered time-independent covariates in the competing risk model, however, introducing time-dependent covariates in the survival model is feasible, which can be achieved by evaluating the value of the time-dependent covariate at each time point for the risk set at that time.

Parameter estimation was performed by MLE through an EM algorithm and a one-step Newton-Raphson method. Parameter estimation was conducted by MLE through an EM algorithm and a one-step Newton-Raphson method. A profile likelihood approach was used to obtain the standard errors estimates. This approach may result in an underestimation. An alternative approach one could use instead is a non-parametric bootstrap, which is based on re-sampling individuals at random. However, for large datasets, this approach will not be computationally attractive. Another method proposed by [Louis \(1982\)](#) relies on the calculation of the inverse of the observed Hessian. Since the baseline hazard functions λ_{0d} , $d = 1, \dots, g$, are unspecified, the dimension of the Hessian matrix increases as the sample size increases and thus, this method becomes time-consuming and computationally unattractive in the calculation and inversion of this matrix. A feasible alternative to this is an entirely parametric model for the baseline hazard. Simulation scenarios were carried out to show that the proposed joint model provides less biased estimates than the separate approach, even under high censoring rates. For numerical integration in the expectation step, Gauss-Hermite rules were applied. This technique uses Q points in each of m dimensions of random effects and approximates the multiple integrals with a summation over Q^m quadrature points. Practically, this is computationally feasible for integral dimensions up to six. So if the number of modalities of longitudinal outcomes exceeds five, this technique becomes intractable. One could aggregate the modalities, if possible, to reduce the dimension of integrals or as an alternative approach the use of Monte Carlo techniques is proposed. In the current parametrization of our joint model, the value of the longitudinal outcome is associated with the cause-specific hazard of an event of interest. It is also possible to estimate the association between other characteristics of professional trajectories, such as the cumulated time spent in each professional position and the number of transitions between professional episodes, by introducing additional random effects into the model. If the number of random effects exceeds six, then parameter estimation should be based on the Monte Carlo techniques.

A large sample size of administrative data on professional trajectories motivated this work. When the datasets are large the calculation of parameter estimates of the joint model is not feasible, therefore, we adopted a meta-analysis strategy. We applied the proposed joint model in parallel on randomly stratified sub-samples of the large dataset and then combined the results by taking the average of the estimations. The meta-analysis strategy provides a pragmatic approach to address the computational burden of joint models. We should point out that a practical choice for sub-sample sizes had been made in this study. Future works may focus systematically on the sub-sample sizes and their relationship to the quality of estimations.

Acknowledgements

This work was mainly supported by joint funding from the Ministry of Health, the General Directorate of Health and the Mission research of the Directorate of research, studies, evaluation and statistics, the National Health Insurance Fund of the Salaried Workers, the Social Security Scheme for Self-employed Workers, the National Solidarity Fund for Autonomy and the National Institute of Prevention and Education for Health, within the framework of the 2011 call for research launched by IReSP (Institute of Research in Public Health) (grant number A11226LS). This work was completed by the support of INCa (National Institute of Cancer) and Cancéropôle Ile-de-France (grant number 2012-1-PL SHS-05-INSERM 11-1).

Appendix. The EM Algorithm

The complete-data log-likelihood of the equation (6) is written as follows:

$$\begin{aligned}
l(\Psi|Y, \tilde{T}, \mathbf{a}) &= \log L(\Psi|Y, \tilde{T}, \mathbf{a}) \\
&= \sum_{i=1}^n \left[\left(\sum_{j=1}^{m_i} \sum_{k=1}^K I(Y_{ij} = k) \log(\pi_{ijk}) \right) \right. \\
&\quad + \left(\sum_{d=1}^g I(\epsilon_i = d) \log(\lambda_d(T_i|\mathbf{Z}_i, u)) \right) \\
&\quad + \left(- \int_0^{T_i} \left\{ \sum_{d=1}^g \lambda_d(t|\mathbf{Z}_i, u_i) \right\} dt \right) \\
&\quad \left. + \left(\log \left(\frac{1}{\sqrt{(2\pi)^{(K-1)q+1} |\Sigma|}} \right) \left(- \frac{1}{2} \mathbf{a}_i' \Sigma^{-1} \mathbf{a}_i \right) \right) \right],
\end{aligned} \tag{17}$$

where

$$\log(\pi_{ijk}) = \begin{cases} 0 - \log \left(1 + \sum_{h=1}^{K-1} \exp(\alpha_h + \mathbf{X}_{ij}' \beta_h + \mathbf{W}_{ij}' \mathbf{b}_{ih}) \right) & \text{if } k = K, \\ (\mathbf{X}_{ij}' \beta_k + \mathbf{W}_{ij}' \mathbf{b}_{ik}) - \log \left(1 + \sum_{h=1}^{K-1} \exp(\alpha_h + \mathbf{X}_{ij}' \beta_h + \mathbf{W}_{ij}' \mathbf{b}_{ih}) \right) & \text{if } k \neq K. \end{cases} \tag{18}$$

E-step

In the $(m+1)$ -th iteration of the E-step, we evaluate:

$$\begin{aligned}
E_{\mathbf{a}_i|Y_i, \tilde{T}_i, \Psi^{(m)}}[h(\mathbf{a}_i)] &= \int h(\mathbf{a}_i) f(\mathbf{a}_i|Y_i, \tilde{T}_i, \Psi^{(m)}) d\mathbf{a}_i \\
&= \frac{\int h(\mathbf{a}_i) f(Y_i, \tilde{T}_i, \mathbf{a}_i|\Psi^{(m)}) d\mathbf{a}_i}{f(Y_i, \tilde{T}_i|\Psi^{(m)})} \\
&= \frac{\int h(\mathbf{a}_i) f(Y_i|\mathbf{a}_i, \Psi^{(m)}) f(\tilde{T}_i|\mathbf{a}_i, \Psi^{(m)}) f(\mathbf{a}_i|\Psi^{(m)}) d\mathbf{a}_i}{\int f(Y_i|\mathbf{a}_i, \Psi^{(m)}) f(\tilde{T}_i|\mathbf{a}_i, \Psi^{(m)}) f(\mathbf{a}_i|\Psi^{(m)}) d\mathbf{a}_i},
\end{aligned} \tag{19}$$

for $h(a)$ being a function of $l(\Psi|Y, \tilde{T}, \mathbf{a})$.

M-Step

The M-step estimates the new parameter by maximizing the expected log-likelihood mentioned in equation (19):

$$\Psi^{(m+1)} = \arg \max_{\Psi} E_{\mathbf{a}|Y, \tilde{T}, \Psi^{(m)}}[l(\Psi|Y, \tilde{T}, \mathbf{a})]. \tag{20}$$

5 The closed-form maximum likelihood estimates of the baseline hazards function are updated as follows:

$$\lambda_{0d}^{(m+1)}(t_{dj}) = \frac{n_{dj}}{\sum_{r \in R(t_{dj})} \exp\left(\mathbf{Z}_r' \gamma_d^{(m)}\right) E\left(\exp\left(\nu_d^{(m)} u_r\right)\right)}, \tag{21}$$

where $R(t_{dj})$ is the risk set at time t_{dj} and n_{dj} is the number of failures due to cause d at time t_{dj} , and the cumulative baseline hazard function for cause d , $H_{0d}(t)$, is assumed to be a step function with jumps at observed event times due to cause d , $d = 1, \dots, g$:

$$H_{0d}^{(m+1)}(t_{dq}) = \sum_{j=1}^q \lambda_{0d}^{(m+1)}(t_{dj}), \tag{22}$$

where q_d is the number of distinct failure times due to the d th cause, $t_{d1} \leq \dots \leq t_{dq_d}$ for $d = 1, \dots, g$.

10 The variance-covariance matrix Σ is updated as following:

$$\begin{aligned}
\Sigma_{b_k}^{(m+1)} &= \frac{1}{n} \sum_{i=1}^n E(b_{ik}b'_{ik}), \\
\Sigma_{b_k b_j}^{(m+1)} &= \frac{1}{n} \sum_{i=1}^n E(b_{ik}b'_{ij}), \\
\sigma_u^{2(m+1)} &= \frac{1}{n} \sum_{i=1}^n E(u_i^2), \\
\Sigma_{bu}^{(m+1)} &= \frac{1}{n} \sum_{i=1}^n E(b_i u_i),
\end{aligned} \tag{23}$$

where E stands for $E_{\mathbf{a}_i|Y_i, \tilde{T}_i, \Psi^{(m)}}$.

Since no closed-form solution exists for score equations of α , β , γ and ν , a one-step Newton-Raphson method is required to update the parameter estimations at each iteration:

$$\alpha_k^{(m+1)} = \alpha_k^{(m)} - \frac{S_{\alpha_k}^{(m)}}{I_{\alpha_k}^{(m)}}, \tag{24}$$

for $k = 1, \dots, K-1$, with $S_{\alpha_k}^{(m)}$ and $I_{\alpha_k}^{(m)}$ being

$$\begin{aligned}
S_{\alpha_k} &= \frac{\partial l(\Psi|Y, \tilde{T}, \mathbf{a})}{\partial \alpha_k} = \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{\partial}{\partial \alpha_k} \left[\sum_{k=1}^K I(Y_{ij} = k) \log(\pi_{ijk}) \right] \\
&= \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{h=1}^K I(Y_{ij} = h) E\{I(h = k) - \pi_{ijk}\},
\end{aligned} \tag{25}$$

and

5

$$I_{\alpha_k}^{(m)} = \frac{\partial S_{\alpha_k}}{\partial \alpha_k} = \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{h=1}^K I(Y_{ij} = h) E\{\pi_{ijk}^2 - \pi_{ijk}\}. \tag{26}$$

$$\beta_k^{(m+1)} = \beta_k^{(m)} - I_{\beta_k}^{(m)-1} S_{\beta_k}^{(m)}, \tag{27}$$

for $k = 1, \dots, K-1$, with $S_{\beta_k}^{(m)}$ and $I_{\beta_k}^{(m)}$ being

$$\begin{aligned}
S_{\beta_k} &= \frac{\partial l(\Psi|Y, \tilde{T}, a)}{\partial \beta_k} = \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{\partial}{\partial \beta_k} \left[\sum_{k=1}^K I(Y_{ij} = k) \log(\pi_{ijk}) \right] \\
&= \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{h=1}^K \left[I(Y_{ij} = h) E\{I(h = k) - \pi_{ijk}\} X_{ij} \right],
\end{aligned} \tag{28}$$

and

$$I_{\beta_k}^{(m)} = \frac{\partial S_{\beta_k}}{\partial \beta_k} = \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{h=1}^K \left[I(Y_{ij} = h) E\{\pi_{ijk}^2 - \pi_{ijk}\} X_{ij} X'_{ij} \right]. \tag{29}$$

$$\gamma_d^{(m+1)} = \gamma_d^{(m)} + I_{\gamma_d}^{(m)-1} S_{\gamma_d}^{(m)}, \tag{30}$$

for $d = 1, \dots, g$, with $S_{\gamma_d}^{(m)}$ and $I_{\gamma_d}^{(m)}$ being

$$\begin{aligned}
S_{\gamma_d} &= \frac{\partial l(\Psi|Y, \tilde{T}, a)}{\partial \gamma_d} = \sum_{i=1}^n \left\{ I(\epsilon_i = d) Z_i - \sum_{t_{dj} \leq T_i} \lambda_{0d}^{(m+1)}(t_{dj}) \exp(Z'_i \gamma_d^{(m)}) \times \right. \\
&\quad \left. E(\exp(\nu_d^{(m)} u_i)) Z_i \right\},
\end{aligned} \tag{31}$$

and

$$\begin{aligned}
I_{\gamma_d} &= -\frac{\partial S_{\gamma_d}}{\partial \gamma_d} = \sum_{i=1}^n \sum_{t_{dj} \leq T_i} \lambda_{0d}^{(m+1)}(t_{dj}) \exp(Z'_i \gamma_d^{(m)}) \times \\
&\quad E(\exp(\nu_d^{(m)} u_i)) Z_i Z'_i.
\end{aligned} \tag{32}$$

$$\nu_d^{(m+1)} = \nu_d^{(m)} + I_{\nu_d}^{(m)-1} S_{\nu_d}^{(m)}, \tag{33}$$

for $d = 1, \dots, g$, with $S_{\nu_d}^{(m)}$ and $I_{\nu_d}^{(m)}$ being

$$S_{\nu_d} = \frac{\partial l(\Psi|Y, \tilde{T}, a)}{\partial \nu_d} = \sum_{i=1}^n \left\{ I(\epsilon_i = d) E(u_i) - \sum_{t_{dj} \leq T_i} \lambda_{0d}^{(m+1)}(t_{dj}) \exp(Z_i' \gamma_d^{(m)}) \times \right. \\ \left. E(u_i \exp(\nu_d^{(m)} u_i)) \right\}, \quad (34)$$

and

$$I_{\nu_d} = -\frac{\partial S_{\nu_d}}{\partial \nu_d} = \sum_{i=1}^n \sum_{t_{dj} \leq T_i} \lambda_{0d}^{(m+1)}(t_{dj}) \exp(Z_i' \gamma_d^{(m)}) \times \\ E(u_i^2 \exp(\nu_d^{(m)} u_i)). \quad (35)$$

- Agresti, A., 2013. Categorical Data Analysis, 3rd Edition. Wiley-Blackwell, Hoboken, NJ, USA.
- Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer-Verlag, New York, USA.
- Beyersmann, J., Latouche, A., Buchholz, A., Schumacher, M., 2009. Simulating competing risks data in survival analysis. *Stat Med* 28 (6), 956–971. 5
- Dagum, L., Menon, R., 1998. "OpenMP": an industry standard API for shared-memory programming. *IEEE Comput Sci Eng* 5 (1), 46–55.
- Davey Smith, G., Hart, C., Blane, D., Gillis, C., Hawthorne, V., 1997. Lifetime socioeconomic position and mortality: prospective observational study. *BMJ* 314 (7080), 547–552. 10
- Diggle, P. J., 1998. Dealing with missing values in longitudinal studies. In: Everitt, B. S., Dunn, G. (Eds.), *Recent advances in the Statistical Analysis of Medical Data*. Arnold, London, pp. 203–228.
- Eerola, M., Helske, S., 2012. Statistical analysis of life history calendar data. *Stat Methods Med Res* 0 (0), 1–27.
- Elashoff, R. M., Li, G., Li, N., 2007. An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Stat Med* 26 (14), 2813–2835. 15
- Elashoff, R. M., Li, G., Li, N., 2008. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* 64 (3), 762–771.
- Faucett, C. L., Thomas, D. C., 1996. Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: a Gibbs Sampling Approach. *Stat Med* 15 (15), 1663–1685.
- Hartzel, J., Agresti, A., Caffo, B., 2001. Multinomial logit random effects models. *Stat Modelling* 1 (2), 81–102. 20
- Hedeker, D., 2003. A mixed-effects multinomial logistic regression model. *Stat Med* 22 (9), 1433–1446.
- Henderson, R., Diggle, P., Dobson, A., 2000. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1 (4), 465–480.
- Huang, X., Li, G., Elashoff, R. M., Pan, J., 2011. A general joint model for longitudinal measurements and competing risks survival data. *Lifetime Data Anal* 17 (1), 80–100. 25
- Karimi, M., Geoffroy-Perez, B., Fouquet, A., Latouche, A., Rey, G., 2015. Socioprofessional trajectories and mortality in France, 1976-2002: a longitudinal follow-up of administrative data. *J Epidemiol Community Health* 69 (4), 339–346.
- Laird, N. M., Ware, J. H., 1982. Random-effects models for longitudinal data. *Biometrics* 38 (4), 963–974.
- Li, N., Elashoff, R. M., Li, G., Saver, J., 2010. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Stat Med* 29 (5), 546–557. 30

- Liang, K.-Y., Zeger, S. L., 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73 (1), 13–22.
- Louis, T. A., 1982. Finding the observed information matrix when using the em algorithm. *J R Stat Soc Series B Methodol* 44 (2), 226–233.
- Melchior, M., Berkman, L. F., Kawachi, I., Krieger, N., Zins, M., Bonenfant, S., Goldberg, M., 2006. Lifelong socioeconomic trajectory and premature mortality (35-65 years) in France: findings from the GAZEL Cohort Study. *J Epidemiol Community Health* 60 (11), 937–944.
- Molenberghs, G., Verbeke, G., Iddi, S., 2005. Pseudo-likelihood methodology for partitioned large and complex samples. *Stat Probab Lett* 81 (7), 892–901.
- Murawska, M., Rizopoulos, D., 2013. Extensions in joint modeling of survival and longitudinal outcomes. Ph.D. thesis, Erasmus University Rotterdam.
- Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., Breslow, N. E., 1978. The analysis of failure times in the presence of competing risks. *Biometrics* 34 (4), 541–554.
- Rizopoulos, D., 2012. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R, 1st Edition. Chapman & Hall/CRC.
- Rizopoulos, D., 2013. Joint modeling of longitudinal and time-to-event data: Challenges and future directions. In: Torelli, N., Pesarin, F., Bar-Hen, A. (Eds.), *Advances in Theoretical and Applied Statistics*. Springer, pp. 199–209.
- Rubin, D. B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Tsiatis, A. A., Davidian, M., 2004. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sinica* 14, 809–834.
- Wulfsohn, M. S., Tsiatis, A. A., 1997. A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics* 53 (1), 330–339.
- Zeng, D., Cai, J., 2005. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann Stat* 33 (5), 2132–2163.