Yuqi Chen[1] / Wensheng Guo[2] / Peter Kotanko[3] / Len Usvyat[4] / Yuedong Wang[1]

Joint Model for Mortality and Hospitalization

[1] Statistics & Applied Probability – University of California – Santa Barbara, Santa Barbara, CA, USA, E-mail: ychen@pstat.ucsb.edu, yuedong@pstat.ucsb.edu
[2] Department of Biostatistics and Epidemiology – Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, E-mail: wguo@mail.med.upenn.edu
[3] Research Division – Renal Research Institute, New York, NY, USA, E-mail: Peter.Kotanko@RRINY.COM
[4] Fresenius Medical Care North America, Waltham, MA, USA, E-mail: Len.Usvyat@fmc-na.com

Abstract:
Modeling hospitalization is complicated because the follow-up time can be censored due to death. In this paper, we propose a shared frailty joint model for survival time and hospitalization. A random effect semi-parametric proportional hazard model is assumed for the survival time and conditional on the follow-up time, hospital admissions or total length of stay is modeled by a generalized linear model with a nonparametric offset function of the follow-up time. We assume that the hospitalization and the survival time are correlated through a latent subject-specific random frailty. The proposed model can be implemented using existing software such as SAS Proc NLMIXED. We demonstrate the feasibility through simulations. We apply our methods to study hospital admissions and total length of stay in a cohort of patients on hemodialysis. We identify age, albumin, neutrophil to lymphocyte ratio (NLR) and vintage as significant risk factors for mortality, and age, gender, race, albumin, NLR, pre-dialysis systolic blood pressure (preSBP), interdialytic weight gain (IDWG) and equilibrated Kt/V (eKt/V) as significant risk factors for both hospital admissions and total length of stay. In addition, hospitalization admissions is positively associated with vintage.

Keywords: end stage rental disease, hemodialysis, mixed outcomes, random effect, spline
DOI: 10.1515/ijb-2016-0002

1 Introduction

Hospitalization is a main contributor to the total cost of care and identification of the related risk factors is of interest in many health care studies. The main difficulty in modeling hospitalization data is due to the fact that the frequency of hospitalization and the total length of hospitalizations are functions of follow-up time that can be informatively censored due to death. Since both the hospitalization outcome and time-to-death are related to the underlying health, it is desirable to jointly model them as bivariate outcomes. Mixed types of multivariate outcomes are common in many fields of science and social science. Various statistical models and methods have been proposed to deal with different types of mixed outcomes [1]. For example, Fitzmaurice and Laird [2] proposed regression models for continuous and binary outcomes. They focused on marginal regression models with a set of covariates and treated the association between continuous and binary response as a nuisance characteristic of the data. Sammel, Ryan, and Legler [3] proposed latent variable models for mixed discrete and continuous outcomes. They modeled the associations among the outcomes by an unobserved latent variable which depends on a set of covariates. Catalano [4] proposed a latent variable model for continuous and ordinal outcomes, and extended it to allow for clustering of the bivariate outcomes. Dunson and Herring [5] proposed latent variable models for mixed discrete outcomes including count, binary and discrete event time. A Bayesian approach was introduced for inference where conditionally-conjugate priors were chosen to facilitate posterior computation. However, these methods can not handle censored data which is needed for joint modeling of survival time and hospitalization in health studies.

Our research is motivated by the need for improvement in care for end-stage renal disease (ESRD) patients. Hemodialysis (HD) is the most frequently used treatment modality for ESRD patients. In general, HD patients suffer from multiple comorbidities, such as diabetes and cardiovascular diseases, resulting in frequent hospitalizations and substantial mortality. In spite of improvements over the years, hospitalization and mortality rates of ESRD patients on HD remain much higher than those of the general population [6]. In this article we are interested in identifying risk factors for hospitalization and mortality. The data come from an observational study of patients on HD in Fresenius Medical Care. Covariates at baseline and outcomes including survival time, hospital admissions and total length of hospital stay at follow-up were collected. Approximately 20 % of patients

Yuedong Wang is the corresponding author.

died during the follow-up period and observational times for hospitalization outcomes of these patients are censored due to death. Since both survival time and hospitalization are associated with the underlying health condition, it is likely that these outcomes from the same subject are correlated. Therefore, it is necessary to develop a joint model for survival time and hospitalization. Details of the data are given in Section 5.

In this article we propose a semi-parametric latent variable model for joint modeling of a survival time and an outcome from exponential family. The survival time is modeled by a semi-parametric proportional hazard model with a subject-specific random effect. The hospitalization related endpoint, such as the number of admissions, length of stay or whether a subject has ever been hospitalized, can be modeled by a generalized linear mixed effects model. Since the hospitalization outcome may only be observed before death, an offset function will be included in the generalized linear model to take into account the follow-up time. To allow a flexible relationship between the hospitalization endpoint and the follow-up time, we introduce a nonparametric smooth offset function that includes parametric functions, such as logarithm, as special cases. When the offset function is parametric, these models reduce to the standard generalized mixed effects models and parameters of interest may be interpreted in terms of the constant conditional means such as incident rate, mean duration and average probability. The smooth offset function allows deviation from this rigid assumption. The forms of the baseline hazard function and the offset function are usually unknown. They will be modeled non-parametrically using spline functions with non-negative and, when appropriate, monotone constraints. A latent random variable will be used to model potential correlation between survival time and hospitalization outcome from the same subject [7].

We note that there is a large body of literature on the joint modeling of survival hazard function and hospitalization rate. See for example Lancaster and Intrator [8], Wang, Qin, and Chiang [9], Huang and Wolfe [10], Liu, Wolfe, and Huang [11], Huang, Qin, and Wang [12], and the references therein. These studies treated hospitalizations as recurrent events and focused on modeling the intensity function of the recurrent process. In this article, our main interest is on the expected number of hospitalizations and expected total length of stays which account for a major part of the total cost of care. We also note that there have been various proposals on the joint modeling of survival time and longitudinal data [13, 14]. We are interested in identifying risk factors at the baseline for the bivariate cross-sectional outcomes of hospitalization and time-to-death in the follow-up. Therefore methods for the joint modeling of longitudinal and survival data do not apply to our situation.

The rest of this article is organized as follows. Section 2 introduces the semi-parametric latent variable model. Section 3 provides details about our estimation procedure. Section 4 and Section 5 present simulation results and applications to patients on HD. The article ends with a discussion in Section 6.

2 The semi-parametric latent variable model

2.1 The overall model

For subject i, we denote D_i as the death time, C_i as the censoring time, $T_i = \min\{C_i, D_i\}$ as the observed time, $\Delta_i = I(D_i < C_i)$ as the event indicator and $h_i(t)$ as the hazard function. Let Y_i be another outcome variable from exponential family. For example, it could be the number of hospitalizations or the total length of hospital stays of subject i. Let Z_i^D and Z_i^Y be covariates associated with the outcomes D_i and Y_i respectively. We will consider the following joint model:

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\beta' Z_i^D + \nu_i), \\ g\left(\mathrm{E}(Y_i|T_i,\nu_i)\right) &= w(T_i) + \alpha' Z_i^Y + \eta\nu_i, \end{aligned} \tag{1}$$

where h_0 is the baseline hazard, g is the link function, $\nu_i \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$ is a shared frailty for subject i, α, β and η are unknown parameters, and w is an offset function. The first equation in (1) is a Cox proportional hazard model for survival time while the second equation in (1) is a generalized linear model for Y_i. The shared frailty is introduced to model heterogeneity among subjects and correlation between D_i and Y_i within a subject. For simplicity we consider a normal distribution for the shared frailty. Extensions to other distributions are straightforward. The offset term $w(T_i)$ is introduced to account for the fact that Y_i is only observed prior to time T_i.

2.2 A spline model for the baseline hazard

The form of the baseline hazard function $h_0(t)$ is generally unknown in practice. We will assume that $h_0(t)$ is a smooth function and model it using B-spline basis functions:

$$h_0(t) = \sum_{k=1}^{K+1+L_h} d_k B_k(t|K, \tau_h),$$

where $B_k(t|K, \tau_h)$ denote the evaluation at t of the K-degree B-spline basis functions generated with L_h internal knots $\tau_h = \{t_{h1}, t_{h2}, \cdots, t_{hL_h}\}$. We will use the constraints $d_k \geq 0$ to enforce the non-negativity constraint of the function $h_0(t)$. The function $h_0(t)$ is decided by coefficients d_k as well as the number and locations of knots. The estimation of coefficients and the selection of knots will be discussed in Section 3.

2.3 A spline or monotone spline model for the offset function

When Y_i represents counts such as hospital admissions, one possible assumption is that Y_i is generated from a homogeneous Poisson process. Under this assumption and canonical link for Poisson data, the offset function $w(t) = \log(t)$. However in practice Y_i may be generated from a non-homogeneous Poisson process [15]. It is therefore desirable to leave the functional form of w unspecified. Again we model w nonparametrically using B-spline basis functions:

$$w(t) = \sum_{k=1}^{K+1+L_w} c_k B_k(t|K, \tau_w),$$

where $B_k(t|K, \tau_w)$ denote the evaluation at t of the K-degree B-spline basis functions generated with L_w internal knots $\tau_w = \{t_{w1}, t_{w2}, \cdots, t_{wL_w}\}$.

For Poisson data, it is natural to assume that the expectation of Y_i increase with the observational time T_i. In this case we assume that $w(t)$ is a smooth non-decreasing function. Ramsay [16] used integrated M-splines to fit a monotone spline. We will adopt a similar approach using integrated B-splines. Specifically, denote integrated B-splines as $I_k(t|K, \tau) = \int_0^t B_k(u|K, \tau)du$ for $k = 1, \ldots, K$. Since B_k's are non-negative, I_k's provide a set of non-decreasing basis functions. We model w using integrated B-spline basis functions:

$$w(t) = \sum_{k=1}^{K+1+L_w} c_k I_k(t|K, \tau_w) + c,$$

where c is an unknown constant and c_k's are coefficients with constraints $c_k \geq 0$.

3 Estimation

The full likelihood is

$$L = \prod_{i=1}^{n} \int f(Y_i|T_i, \Delta_i, \nu_i) l_i(T_i, \Delta_i|\nu_i) f_\nu(\nu_i) d\nu_i, \tag{2}$$

where n is the total number of subject, $f(Y_i|T_i, \Delta_i, \nu_i)$ is the conditional density of Y_i in the exponential family, $f_\nu(\nu_i)$ is the density function of the latent random variable ν, and

$$l_i(T_i, \Delta_i|\nu_i) = \left\{h_0(T_i) \exp(\beta' Z_i^D + \nu_i)\right\}^{\Delta_i} \exp\left\{-\int_0^{T_i} h_0(t) \exp(\beta' Z_i^D + \nu_i) dt\right\}.$$

Our goal is then to obtain parameter estimates by maximizing the likelihood. Since there is no closed form solution, we apply the Newton-Raphson methods to compute parameter estimates numerically. For stability, we apply the Newton-Raphson ridge optimization where a pure Newton step is used when the Hessian is positive definite and when the Newton step successfully increases the value of the likelihood, otherwise a multiple of the identity matrix is added to the Hessian matrix [17]. To calculate the gradient and Hessian matrix, we need to evaluate integrals derived from the likelihood function. The Gaussian quadrature method is used to

approximate these integrals. We estimate random effects ν_i by their empirical Bayes estimators $\hat{\nu}_i$ that maximize $f(y_i|T_i, \Delta_i, \nu_i) l_i(T_i, \Delta_i|\nu_i) f_\nu(\nu_i)$.

Numerically stable implementations of these methods can be obtained from a variety of publicly available softwares [18]. In our simulation and example, we employed SAS procedure Proc NLMIXED to perform the computation. Proc NLMIXED has an appealing feature which allows a user-specified log likelihood functions with respect to the random effects. See Littell et al. [17] for details on this procedure.

The number and location of knots are fixed in the above discussion. While increasing the number of knots has the capability to model a more flexible function, having too many knots will increase the complexity of the model and result in over-fitting. A data-driven procedure for the selection of number and location of knots is desirable. We allow $h_0(t)$ and $w(t)$ to have different numbers and locations of knots. In practice one may place knots evenly in a range or at equally spaced quantiles of data. We select the numbers of knots by minimizing the following AIC Akaike [19]:

$$\mathrm{AIC}(L_h, L_w) = -2\log L + 2(L_h + L_w + 8). \tag{3}$$

4 Simulations

We generate simulation samples from the following model

$$\begin{array}{ll} h_i(t|\nu_i) & = h_0(t)\exp(\beta Z_i + \nu_i), \\ \log(\mathrm{E}(Y_i|T_i, \nu_i)) & = w(T_i) + \alpha Z_i + \eta\nu_i, \end{array} \tag{4}$$

where Z_i's are iid random variables with $P(Z_i = 0) = P(Z_i = 1) = 0.5$, $\nu_i \overset{iid}{\sim} \mathrm{N}(0, 0.5)$, and conditional on T_i and ν_i, Y_i follows a Poisson distribution with mean $\exp(w(T_i) + \alpha Z_i + \eta\nu_i)$. The censoring time $C_i = \min\{E_i, 4\}$ where $E_i \overset{iid}{\sim} \mathrm{Exp}(0.1)$. The true parameters are set to be $(\alpha, \beta, \eta) = (0.5, 0.5, 1)$. We consider two baseline hazard functions, Exponential baseline $h_0(t) = 1/2$ and Weibull baseline $h_0(t) = t/2$, and two offset functions, linear function $w(t) = t/2$ and log function $w(t) = \log(t)$. The censoring rates in all 4 cases are about 20 %.

The baseline hazard $h_0(t)$ is estimated using cubic B-spline basis functions. The offset function $w(t)$ is estimated using cubic integrated B-spline basis functions under the monotone constraint. Interior knots are equally spaced within the time period $(0, 4]$, and the number of knots for $h_0(t)$ and $w(t)$ range from 2 to 4 respectively. The optimal combination of number of knots is selected by minimizing the AIC (3).

Simulation under each setting is repeated 500 times. For the estimation of parameters, we compute bias, mean squared error (MSE) and coverage probability of 95 % confidence intervals (CP). The 95 % confidence interval is constructed as the MLE plus-minus 1.96 times the standard errors obtained from the variance-covariance matrix. For the estimation of functions $h_0(t)$ and $w(t)$, we compute the integrated mean square error (IMSE)

$$\mathrm{IMSE}(\hat{f}) = \int_0^4 (\hat{f}(t) - f(t))^2 dt$$

for each replicate, where f is either h_0 or w.

Table 1: Bias, mean squared error (MSE) and coverage probability of 95 % confidence intervals (CP) based on the joint model when $h_0(t) = 1/2$ and $w(t) = t/2$.

$h_0(t) = 1/2$	$w(t) = t/2$	α	$\boldsymbol{\beta}$	η	σ^2
$n = 300$	Bias	0.007	0.045	−0.064	0.337
	MSE	0.017	0.037	0.066	0.65
	CP	0.938	0.981	0.809	0.965
$n = 500$	Bias	0.002	0.014	−0.008	0.149
	MSE	0.010	0.022	0.871	0.936
	CP	0.946	0.946	0.871	0.936
$n = 1{,}000$	Bias	0.002	0.008	0.003	0.063
	MSE	0.005	0.01	0.031	0.062
	CP	0.94	0.948	0.916	0.94

Table 2: Bias, mean squared error (MSE) and coverage probability of 95 % confidence intervals (CP) based on the joint model when $h_0(t) = 1/2$ and $w(t) = \log(t)$.

$h_0(t) = 1/2$	$w(t) = log(t)$	α	β	η	**σ2**
$n = 300$	Bias	0.033	0.084	0.106	0.779
	MSE	0.025	0.064	0.109	2.833
	CP	0.966	0.968	0.774	0.957
$n = 500$	Bias	0.016	0.046	0.03	0.381
	MSE	0.016	0.030	0.088	0.912
	CP	0.955	0.973	0.842	0.953
$n = 1{,}000$	Bias	0.004	0.017	0.005	0.127
	MSE	0.007	0.011	0.053	0.156
	CP	0.947	0.966	0.890	0.951

Table 3: Bias, mean squared error (MSE) and coverage probability of 95 % confidence intervals (CP) based on the joint model when $h_0(t) = t/2$ and $w(t) = t/2$.

$h_0(t) = t/2$	$w(t) = t/2$	α	β	$\eta\ \eta$	σ^2
$n = 300$	Bias	−0.003	0.	0.016	0.056
	MSE	0.011	0.025	0.06	0.075
	CP	0.968	0.963	0.925	0.951
$n = 500n = 500$	Bias	−0.006	0.008	0.007	0.044
	MSE	0.007	0.015	0.038	0.052
	CP	0.944	0.962	0.912	0.930
$n = 1000$	Bias	0.002	0.003	0.011	0.017
	MSE	0.003	0.008	0.022	0.025
	CP	0.950	0.946	0.942	0.928

Table 4: Bias, mean squared error (MSE) and coverage probability of 95 % confidence intervals (CP) based on the joint model when $h_0(t) = t/2$ and $w(t) = \log(t)$.

$h_0(t) = t/2$	$w(t) = log(t)$	α	β	η	σ^2
$n = 300$	Bias	0.033	0.064	0.025	0.346
	MSE	0.020	0.064	−0.025	0.346
	CP	0.958	0.973	0.859	0.936
$n = 500$	Bias	0.014	0.036	−0.014	0.227
	MSE	0.013	0.027	0.070	0.386
	CP	0.945	0.955	0.850	0.951
$n = 1000$	Bias	0.009	0.015	−0.009	0.117
	MSE	0.006	0.011	0.040	0.100
	CP	0.954	0.950	0.892	0.942

Table 5: Integrated Mean Square Error (IMSE) of the baseline hazard $h_0(t)$ and offset function $w(t)$ fitted by the joint model.

		$h_0(t)$	$w(t)$
$h_0(t) = t/2$	$n = 300$	0.078	0.079
$w(t) = t/2$	$n = 500$	0.050	0.052
	$n = 1000$	0.027	0.027
$h_0(t) = t/2$	$n = 300$	0.109	0.151
$w(t) = \log(t)$	$n = 500$	0.063	0.097
	$n = 1000$	0.033	0.052
$h_0(t) = t/2$	$n = 300$	0.665	0.066
$w(t) = t/2$	$n = 500$	0.456	0.043

	$n = 1000$	0.230	0.025
$h_0(t) = t/2$	$n = 300$	0.856	0.165
$w(t) = \log(t)$	$n = 500$	0.662	0.114
	$n = 1000$	0.340	0.057

Table 1–Table 5 summarize performances of parameter and function estimates under four simulation settings. Overall the proposed estimation procedure perform well: bias and MSE are small, and the coverages of 95 % confidence intervals are close to the nominal value except for η. The coverages of 95 % confidence intervals for η are below the nominal value. This is not surprising because η is associated with the variance within subject and only limited information contributes to its estimations. One way to improve the coverage probability is to construct a confidence region for both η and σ^2 since the two estimates are highly correlated. The performances improve as sample size increases.

As an illustration, Figure 1 shows the $5th$, $25th$, $50th$, $75th$ and $95th$ best estimates of $\hat{h}_0(t)$ and $\hat{w}(t)$ ordered by the IMSE under the simulation setting when $h_0(t) = t/2$, $w(t) = \log(t)$ and $n = 500$. Overall, the estimates are close to the true functions except for the baseline hazard with large t. The poor estimation of the baseline hazard with large t is likely caused by censoring.

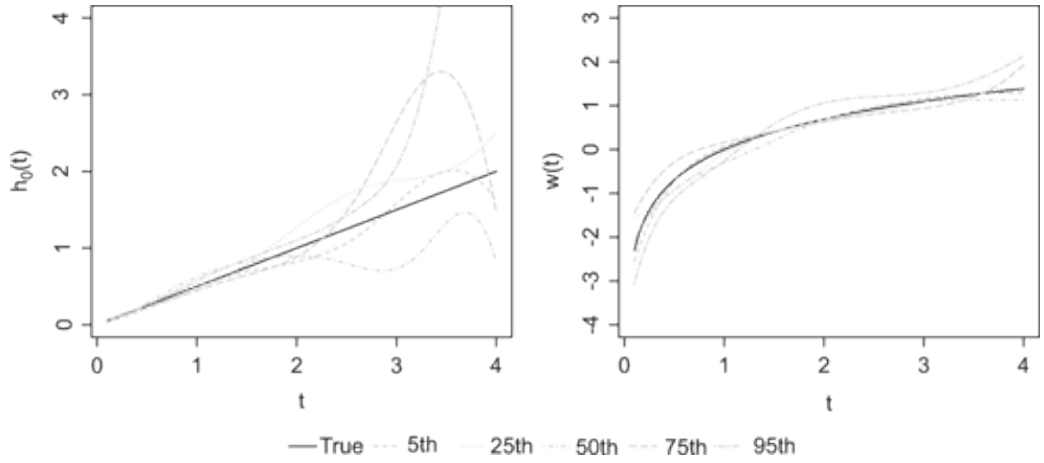

Figure 1: True function (solid lines) and estimates (dashed lines) of $h0(t) = t/2$ (left) and $w(t) = \log(t)$ (right) correspond to the $5th$, $25th$, $50th$, $75th$ and 95^{th} percentiles of the IMSE when $h0(t) = t/2$, $w(t) = \log(t)$ and $n = 500$.

We have also evaluated performance of our estimation procedure in a more complicated simulation setting. The data was generated from the following model

$$\begin{aligned} h_i(t|\nu_i) &= h_0(t)\exp(\beta_1 Z_{1i} + \beta_2 Z_{2i} + \nu_i), \\ \log(\mathrm{E}(Y_i|T_i,\nu_i)) &= w(T_i) + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \eta\nu_i, \end{aligned} \tag{5}$$

where Z_{1i}'s are iid random variables with $P(Z_{1i} = 0) = P(Z_{1i} = 1) = 0.5$, Z_{2i} is a continuous random variable generated from Uniform$(0, 1)$, and $\nu_i \overset{iid}{\sim} \mathrm{N}(0, 0.2)$. The sample size $n = 1000$ and the true parameters are set to be $(\alpha_1, \alpha_2, \beta_1, \beta_2, \eta) = (0.5, -1, 0.5, -1, 1)$. We consider Weibull baseline hazard $h_0(t) = t^2$ and $w(t) = log(t)$. The censoring rate is about 15 %.

We summarize bias, MSE and coverage probability of 95 % CP for the estimations of parameters in Table 6. The $5th$, $25th$, $50th$, $75th$ and $95th$ best estimated baseline hazard and offset function are shown in Figure 2. Overall the proposed estimation method performs well.

Table 6: Bias, mean squared error (MSE) and coverage probability of 95 % confidence intervals (CP) based on the joint model when $h_0(t) = t^2$ and $w(t) = \log(t)$.

$h_0(t) = t^2$	$w(t) = log(t)$	α_1	α_2	β_1	β_2	η	σ^2
$n = 1000$	Bias	−0.020	−0.027	0.037	−0.068	−0.113	0.146
	MSE	0.006	0.018	0.010	0.037	0.083	0.086
	CP	0.949	0.965	0.963	0.946	0.839	0.979

Table 7: Summary statistics of covariates.

	(Min, Max)	Mean (Std)
Age (year)	(1.00, 96.62)	62.39 (14.84)
BMI (kg/m^2)	(13.75, 49.51)	27.65 (6.46)
Albumin (g/dL)	(1.60, 4.74)	3.84 (0.37)
IDWG (%)	(0.41, 7.99)	3.48 (1.05)
PreSBP (mmHg)	(81.88, 219.29)	149.38 (18.86)
eKt/V	(0.68, 3.77)	1.46 (0.26)
NLR	(0.51, 31.18)	3.70 (2.32)
Vintage (year)	(0.08, 7.90)	2.56 (1.92)

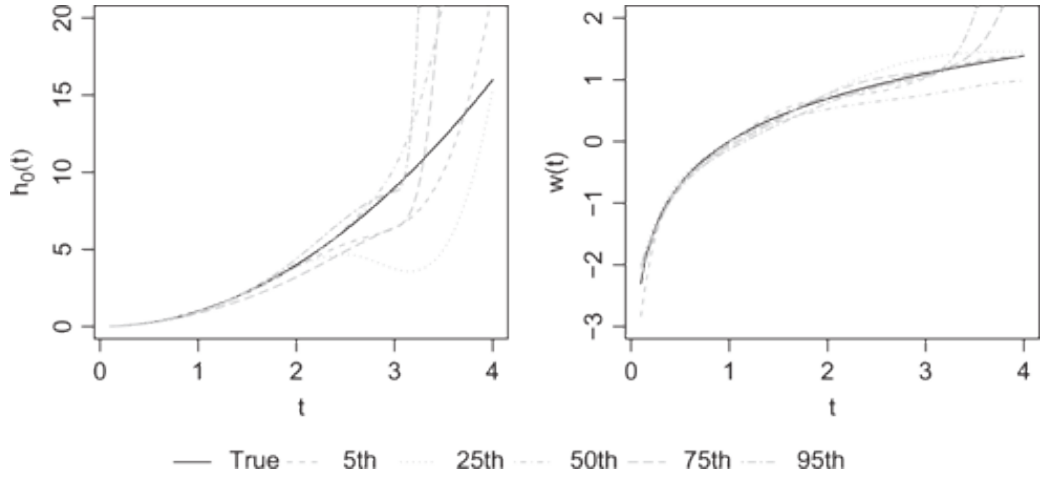

Figure 2: True function (solid lines) and estimates (dashed lines) of $h_0(t) = t^2$ (left) and $w(t) = \log(t)$ (right) correspond to the $5th$, $25th$, $50th$, $75th$ and $95th$ percentiles of the IMSE when $h_0(t) = t^2$, $w(t) = \log(t)$ and $n = 1000$.

5 Application

We now apply the proposed method to model mortality and hospitalization outcomes for patients on HD. Baseline covariates are collected from 1999 HD patients from 1 January 2007 to 31 December 2007. Survival time, the number of hospital admissions and total length of stay of these patients during the period of 1 January 2008 and 31 December 2009 are collected. 1078 (53.93 %) patients are male. 984 (49.22 %) patients are black, 834 (41.72 %) patients are white, the rest are from other races. Time-varying covariates are calculated as the averages in baseline period for each patient. The summary statistics for these covariates are listed in Table 7.

In previous studies, albumin and systolic blood pressure before dialysis (preSBP) have been found as significant risk factors for mortality [20–22]. Erdem, Kaya, Karatas, Dilek, and Akpolat [23] observed that HD patients with high neutrophil to lymphocyte ratio (NLR) levels have increased risk of short term mortality. Our preliminary analysis indicates that time in years since initiation of dialysis (vintage), inter-dialytic weight gain (IDWG) and a measure of dialysis capability eKt/V also have significant effect on mortality. In addition, we will include gender, race and BMI.

In modeling the hospitalization, the number of hospital admissions is usually the primary outcome which will be studied in Section 5.1 using a Poisson model. We are sometimes also interested in whether a patient has ever been hospitalized as a binary outcome. Since the probability of ever been hospitalized can be derived from the Poisson model, we omit the details of modeling the binary outcome in this paper. Given the subject has been hospitalized, a further goal is to identify the risk factors that lead to longer total length of stay which will be studied in Section 5.2 using a Gamma model. For simplicity we will consider the same set of covariates for all models.

5.1 Joint analysis of mortality and hospital admission

359 (17.96 %) patients died during the follow-up period. The number of hospital admissions in the data ranges from 0 to 37 with mean 2.53. We consider the following joint model:

$$\begin{aligned} h_i(t|\nu_i) &= h_0(t)\exp\{\beta_1 * Age_i + \beta_2 * Albumin_i + \beta_3 * PreSBP_i + \beta_4 * NLR_i \\ &\quad +\beta_5 * BMI_i + \beta_6 * Male_i + \beta_7 * IDWG_i + \beta_8 * eKt/V_i \\ &\quad +\beta_9 * Vintage_i + \beta_{10} * RaceWhite_i + \beta_{11} * RaceBlack + \nu_i\}, \\ g(E(Y_i|T_i,\nu_i)) &= w(T_i) + \alpha_1 * Age_i + \alpha_2 * Albumin_i + \alpha_3 * PreSBP_i + \alpha_4 * NLR_i \\ &\quad +\alpha_5 * BMI_i + \alpha_6 * Male_i + \alpha_7 * IDWG_i + \alpha_8 * eKt/V_i \\ &\quad +\alpha_9 * Vintage_i + \alpha_{10} * RaceWhite_i + \alpha_{11} * RaceBlack_i + \eta\nu_i, \end{aligned} \tag{6}$$

where Y_i represents the number of hospital admissions of patient i and is assumed to follow a Poisson distribution, and $\nu_i \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$.

As in the previous section we set the interior knots for baseline hazard and offset function equally spaced within the time period. The number of knots ranges from 2 to 4. Among all the combinations, the AIC selects 2 knots for the baseline hazard and 2 knots for the offset function.

We summarize the estimation results in Table 8. Tests are constructed based on asymptotic properties of the MLEs after selection of the knots. All covariates except BMI are significantly associated with the expected number of hospital admissions, while age, albumin, NLR, eKt/V and vintage are significantly associated with the hazard function. Overall age, NLR and vintage are positively associated with both hazard and the number of hospital admissions, while albumin and eKt/V are negatively associated with the outcomes. Furthermore, predialysis SBP and IDWG are positively associated with the number of hospital admissions, and female patients tend to have more hospital admissions.

The latent random variable is significant ($\hat{\sigma}^2 = 0.6008$, $p = 0.0057$), which supports the model with random effect. Furthermore $\hat{\eta}$ is significantly larger than 0 ($p < 0.0001$). It implies that the survival time and the number of hospital admissions are positive correlated. The estimated baseline function $h_0(t)$ and offset function $w(t)$ are shown in Figure 3 with 95 % point-wise confidence intervals. The confidence intervals are constructed based on asymptotic variances of the MLEs of coefficients associated with the B-spline bases. While our model allows for inhomogeneous Poison model, the logarithm function is close to the estimated offset function and well within the 95 % confidence intervals, suggesting that it is reasonable to model the offset function by the logarithm function in this case.

Table 8: Joint modeling of mortality and hospitalization of ESRD data.

	Covariates	Estimate	SE	p-value
Mortality	Age	0.0355	0.0048	< 0.0001
	Albumin	−1.2736	0.1681	< 0.0001
	PreSBP	−0.0004	0.0031	0.8990
	NLR	0.1061	0.0217	< 0.0001
	BMI	−0.0198	0.0106	0.0619
	Male	0.0748	0.1210	0.5365
	IDWG	0.0684	0.0625	0.2737
	eKt/V	−0.5936	0.2474	0.0165
	Vintage	0.1244	0.0307	< 0.0001
	Race(White)	0.1341	0.2151	0.5329
	Race(Black)	−0.2446	0.2184	0.2629
Hospitalization	Age	0.0089	0.0022	< 0.0001
	Albumin	−0.8126	0.0856	< 0.0001
	PreSBP	0.0072	0.0015	< 0.0001
	NLR	0.0776	0.0129	< 0.0001
	BMI	−0.0026	0.0049	0.5974
	Male	−0.1612	0.0600	0.0073
	IDWG	0.1018	0.0307	0.0009
	eKt/V	−0.2360	0.1170	0.0437
	Vintage	0.0386	0.0157	0.0140
	Race(White)	0.2634	0.1094	0.0162
	Race(Black)	0.3130	0.1069	0.0035
	σ^2	0.6008	0.2172	0.0057
	η	1.2225	0.2039	< 0.0001

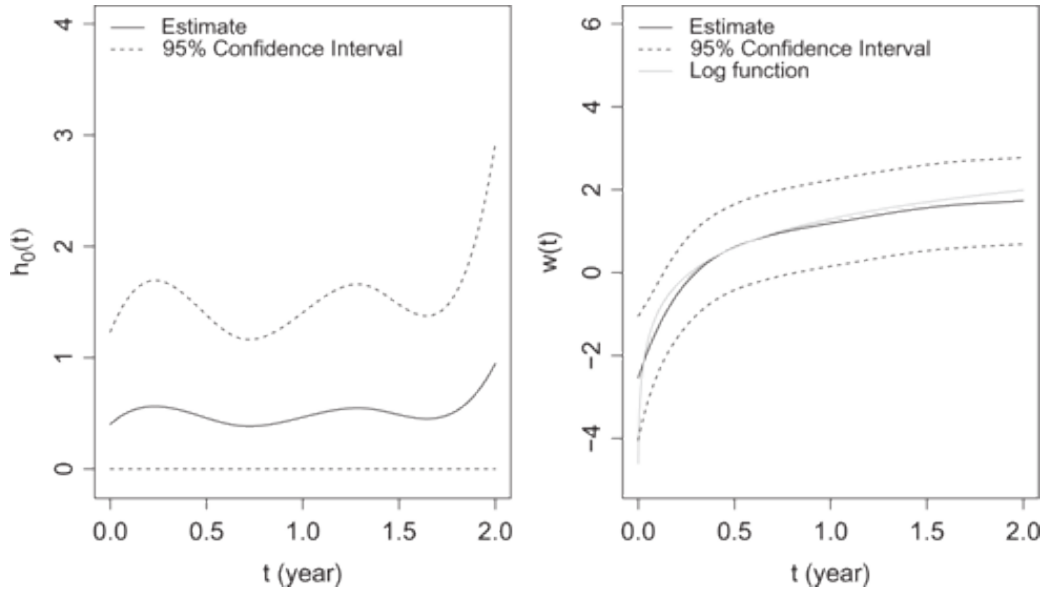

Figure 3: The estimated baseline function $h_0(t)$ and offset function $w(t)$ for the joint model of mortality and number of hospitalization.

5.2 Joint analysis of mortality and total length of stay

To further investigate the features of patients with hospitalizations, another interesting application is to model mortality and total length of hospital stay. We will focus on the patients who had positive length of stays (1396 patients). The total length of stay ranges from 1 to 368 with mean 26.13. We consider the following joint model:

$$\begin{aligned} h_i(t|\nu_i) &= h_0(t)\exp\{\beta_1 * Age_i + \beta_2 * Albumin_i + \beta_3 * PreSBP_i + \beta_4 * NLR_i \\ &\quad + \beta_5 * BMI_i + \beta_6 * Male_i + \beta_7 * IDWG_i + \beta_8 * eKt/V_i \\ &\quad + \beta_9 * Vintage_i + \beta_{10} * RaceWhite_i + \beta_{11} * RaceBlack + \nu_i\}, \\ g(\mu_i|T_i,\nu_i) &= w(T_i) + \alpha_1 * Age_i + \alpha_2 * Albumin_i + \alpha_3 * PreSBP_i + \alpha_4 * NLR_i \\ &\quad + \alpha_5 * BMI_i + \alpha_6 * Male_i + \alpha_7 * IDWG_i + \alpha_8 * eKt/V_i \\ &\quad + \alpha_9 * Vintage_i + \alpha_{10} * RaceWhite_i + \alpha_{11} * RaceBlack_i + \eta\nu_i, \end{aligned} \tag{7}$$

where Y_i represents the total length of stay of patient i and is assumed to follow a Gamma distribution, and $\nu_i \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$.

Similar process for knots selection applies, which results in 2 knots for the baseline hazard and 2 knots for the offset function. The estimation results are summarized in Table 9. All covariates except BMI and vintage are significantly associated with the expectation of total length of stay, while age, albumin, NLR and vintage are significantly associated with the hazard function. We note that conclusions about risk factors are consistent with those in the previous subsection except for race: the total length of hospital stays of white patients is not significantly different from that of other races while white patients have significantly larger number of hospitalizations than other races. The latent random variable is borderline significant ($\hat{\sigma}^2 = 0.2108$, $p = 0.0542$). The estimated baseline function $h_0(t)$ and offset function $w(t)$ are shown in Figure 4. The estimated offset function $w(t)$ is quite different from the logarithm function in this case.

Table 9: Joint modeling of mortality and hospitalization of ESRD data.

	Covariates	Estimate	SE	p-value
Mortality	Age	0.0302	0.0053	< 0.0001
	Albumin	−1.0237	0.1730	< 0.0001
	PreSBP	−0.0040	0.0035	0.2472
	NLR	0.0751	0.0228	0.0010
	BMI	−0.0204	0.0118	0.0843
	Male	0.0830	0.1340	0.5359
	IDWG	0.0939	0.0704	0.1826
	eKt/V	−0.4920	0.0704	0.0812

	Vintage	0.0944	0.0332	0.0045
	Race(White)	−0.0361	0.2303	0.8754
	Race(Black)	−0.3373	0.2335	0.1489
Length of Stay	Age	0.0070	0.0022	0.0017
	Albumin	−0.5335	0.0870	< 0.0001
	PreSBP	0.0047	0.0016	0.0036
	NLR	0.0484	0.0137	0.0004
	BMI	−0.0045	0.0051	0.3788
	Male	−0.1269	0.0626	0.0430
	IDWG	0.0740	0.0319	0.0205
	eKt/V	−0.2495	0.1234	0.0433
	Vintage	0.0275	0.0165	0.0955
	Race(White)	0.1338	0.1123	0.2336
	Race(Black)	0.2933	0.1110	0.0084
	σ^2	0.2108	0.1094	0.0542
	η	1.8883	0.5550	0.0007

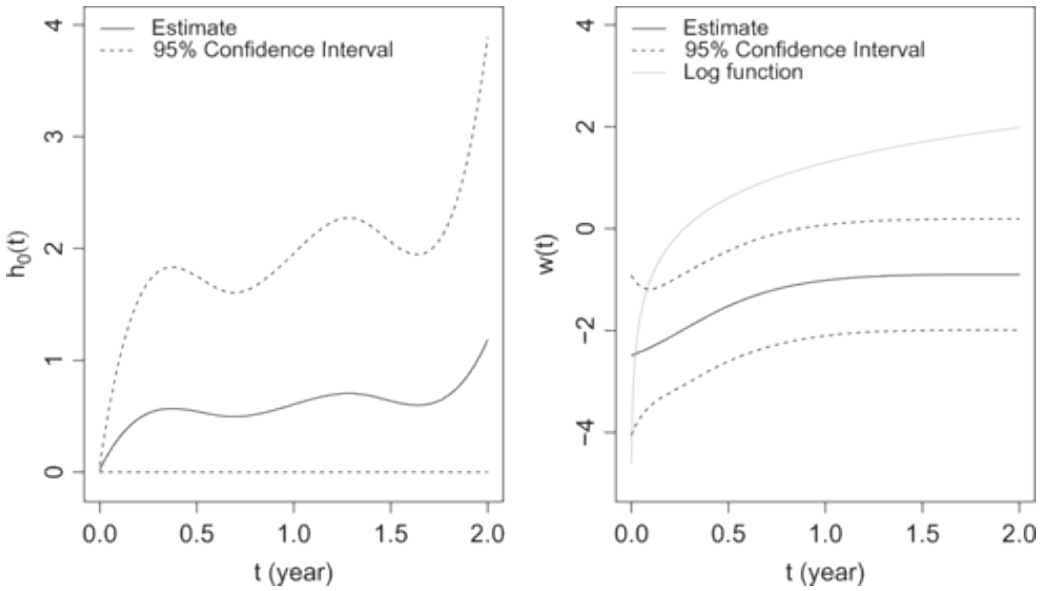

Figure 4: The estimated baseline function $h_0(t)$ and offset function $w(t)$ for the joint model of mortality and total length of stay.

6 Discussion

In this article, we propose a semi-parametric joint model for survival time and hospitalization. In particular, we consider the number of hospital admissions and total length of stay as hospitalization outcomes. A shared random effect is introduced to account for the within subject correlation between the two outcomes. The baseline hazard and offset functions are modeled non-parametrically through B-spline or monotone B-spline bases in order to gain flexibility. With fixed number of knots, the techniques to numerically obtain maximum likelihood estimation are presented. We have also discussed the AIC method for selecting the number of knots. Standard large sample properties of maximum likelihood estimation apply when knots are fixed. Simulation results indicate that the proposed estimation method performs well.

Throughout this article, we assume Normal distribution for the random effect. Our method can be easily generalized to other parametric distributions for the random effect. We used B-spline bases with non-negative coefficients to model the non-negative baseline hazard. An alternative approach is to model the logarithm of the baseline hazard using B-spline bases without constraints on coefficients. However the approach cannot be implemented using the SAS NLMIXED procedure since the likelihood involves an intractable integral. We have analyzed different aspects of the hospitalization separately. One future research is to build a joint model for survival time, hospital admission and length of stay. Our methodology may also be extended to the case of the zero-inflated Poisson model.

Acknowledgement

We thank the associated editor and two referees for constructive comments that substantially improved an earlier draft.

Funding

National Science Foundation, Grant DMS-1507620; National Institutes of Health, Grant R01GM104470.

References

[1] De Leon A, Chough KC. Analysis of mixed data: methods & applications. Boca Raton, FL: Chapman and Hall/CRC, 2013.
[2] Fitzmaurice G, Laird N. Regression models for a bivariate discrete and continuous outcome with clustering. J Am Stat Assoc. 1995;90:845–852.
[3] Sammel M, Ryan L, Legler J. Latent variable models for mixed discrete and continuous outcomes. J R Stat Soc Ser B (Stat Method). 1997;59:667–678.
[4] Catalano P. Bivariate modelling of clustered continuous and ordered categorical outcomes. Stat Med. 1997;16:883–900.
[5] Dunson D, Herring A. Bayesian latent variable models for mixed discrete outcomes. Biostatistics. 2005;6:11–25.
[6] Collins A, Foley R, Chavers B, Gilbertson D, Herzog C, Ishani A. US renal data system 2013 annual data report. American journal of kidney diseases. 2014;63(1 Suppl):A7.
[7] McCulloch C. Joint modelling of mixed outcome types using latent variables. Stat Methods Med Res. 2008;17:53–73.
[8] Lancaster T, Intrator O. Panel data with survival: hospitalization of HIV-positive patients. J Am Stat Assoc. 1998;93:46–53.
[9] Wang M-C, Qin J, Chiang C-T. Analyzing recurrent event data with informative censoring. J Am Stat Assoc. 2001;96:1057–1065.
[10] Huang X, Wolfe RA. A frailty model for informative censoring. Biometrics. 2002;58:510–520.
[11] Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. Biometrics. 2004;60:747–756.
[12] Huang C-Y, Qin J, Wang M-C. Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. Biometrics. 2010;66:39–49.
[13] Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. J Stat Softw. 2010;35:1–33.
[14] Tsiatis A, Davidian M. Joint modeling of longitudinal and time-toevent data: an overview. Stat Sin. 2004;14:809–834.
[15] Usvyat L, Kooman J, van der Sande F, Wang Y, Maddux F, Levin N. Dynamics of hospitalizations in hemodialysis patients: results from a large US provider. Nephrol Dial Transplant. 2014;29:442–448.
[16] Ramsay J. Monotone regression splines in action. Stat Sci. 1988;4:425–441.
[17] Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS for mixed models, 2nd Cary, NC: SAS Institute Inc., 2006.
[18] Press W, Teukolsky S, Vetterling W, Flannery B. Numerical recipes 3rd edition: the art of scientific computing. New York, NY: Cambridge University Press, 2007.
[19] Akaike H. Information theory and an extension of the maximum likelihood principle. Second Int Symp Inf Theory 1973:267–281.
[20] He J, Whelton P. Elevated systolic blood pressure and risk of cardiovascular and renal disease: overview of evidence from observational epidemiologic studies and randomized controlled trials. Am Heart J. 1999;138:S211–S219.
[21] Hsu C, McCulloch C, Iribarren C, Darbinian J, Go A. Body mass index and risk for end-stage renal disease. Ann Intern Med. 2006;144:21–28.
[22] Phelan P, O'Kelly P, Walshe J, Conlon P. The importance of serum albumin and phosphorous as predictors of mortality in ESRD patients. Ren Fail. 2008;30:423–429.
[23] Erdem E, Kaya C, Karatas A, Dilek M, Akpolat T. Neutrophil to lymphocyte ratio in predicting short-term mortality in hemodialysis patients. J Exp Clin Med. 2013;30:129–132.