UC Irvine UC Irvine Previously Published Works

Title

Multilevel joint modeling of hospitalization and survival in patients on dialysis

Permalink

https://escholarship.org/uc/item/27t7c0pw

Journal

Stat, 10(1)

ISSN

2049-1573

Authors

Kürüm, Esra Nguyen, Danh V Li, Yihao <u>et al.</u>

Publication Date

2021-12-01

DOI

10.1002/sta4.356

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed



WILEY

Multilevel joint modeling of hospitalization and survival in patients on dialysis

Accepted: 23 December 2020

Esra Kürüm¹ | Danh V. Nguyen² | Yihao Li³ | Connie M. Rhee^{2,4} | Kamyar Kalantar-Zadeh^{2,4} | Damla Şentürk³

¹Department of Statistics, University of California, Riverside, 92521, California, USA ²Department of Medicine, University of California Irvine, Orange, 92868, California, USA

³Department of Biostatistics, University of California, Los Angeles, 90095, California, USA ⁴Harold Simmons Center for Chronic Disease Research and Epidemiology, University of California Irvine School of Medicine, Orange, 92868, California, USA

Correspondence

Esra Kürüm, Department of Statistics, University of California, Riverside, CA 92521, USA. Email: esra.kurum@ucr.edu

Funding information

National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R01 DK092232 More than 720,000 patients with end-stage renal disease in the United States require life-sustaining dialysis treatment. In this population of typically older patients with a high morbidity burden, hospitalization is frequent at a rate of about twice per patient-year. Aside from frequent hospitalizations, which is a major source of death risk, overall mortality in dialysis patients is higher than other comparable populations, including Medicare patients with cancer. Thus, understanding patient- and facility-level risk factors that jointly contribute to longitudinal hospitalizations and mortality is of interest. Towards this objective, we propose a novel methodology to jointly model hospitalization, a binary longitudinal outcome, and survival, based on multilevel data from the United States Renal Data System (USRDS), with repeated observations over time nested in patients and patients nested in dialysis facilities. In our approach, the outcomes are modeled through a common set of multilevel random effects. In order to accommodate the USRDS data structure, we depart from the literature on joint modeling of longitudinal and survival data by including multilevel random effects and multilevel covariates, at both the patient and facility levels. An approximate Expectation-Maximization algorithm is developed for estimation and inference where fully exponential Laplace approximations are utilized to address computational challenges.

KEYWORDS

biostatistics, longitudinal data, SUBJECT AREAS, survival analysis, TOPICS, TOPICS

1 | INTRODUCTION

In the United States, there were over 726,000 individuals with end-stage-renal disease (ESRD) at the end of 2016 among whom 70% were on dialysis, a life-sustaining treatment (United States Renal Data System, 2018). Dialysis patients have a substantially higher level of mortality risk compared to most other morbid populations, including Medicare populations with cancer, diabetes, or cardiovascular disease. Furthermore, because of the nature of dialysis treatment and overall comorbid conditions, dialysis patients are more frequently hospitalized, about twice per patient-year (United States Renal Data System, 2018). Thus, for the dialysis population, mortality and longitudinal hospitalizations are correlated patient outcomes, and it is of interest to examine the relative contribution of risk factors to this correlated outcome pair, after initiation of dialysis. Potential risk factors include both patient-level risk factors and dialysis facility-level risk factors, such as facility staffing (e.g., the ratio of nurses to patients). Recent works in modeling cause-specific (e.g., cardiovascular-related) and all-cause longitudinal hospitalizations in the dialysis population have considered approaches that partly conditioned on survival (Estes, Nguyen, Dalrymple, Mu, & Senturk, 2016; Li et al., 2018), which provide useful perspectives when the primary focus is on patients' longitudinal hospitalization trajectories after initiation of dialysis. Our work here focuses on joint modeling of longitudinal hospitalization and survival outcomes in this population, in particular for multilevel/hierarchical data. In the model development, we address new joint modeling challenges when facing complex hierarchical data with multilevel covariates and computational challenges of high-dimensional random effects (REs).

We propose a novel multilevel joint model (MJM) that accounts for three-level hierarchical data, with longitudinal measurements, hospitalizations over time, nested within subjects, and subjects further nested within dialysis facilities where they receive regular care. MJM accommodates the hierarchical structure of the data from the United States Renal Data System (USRDS), a large national database, through multilevel REs and

multilevel risk factors affecting both survival and longitudinal hospitalization outcomes. In particular, at the subject level, these include patient demographics and baseline comorbidities. At the facility level, facility staffing, such as the ratio of nurses to patients, may impact patient outcomes.

We note that joint modeling of longitudinal and survival outcomes has been extensively studied; see Tsiatis and Davidian (2004) for an excellent review. A common technique in modeling the dependency between the two outcomes is through the use of REs, often referred to as frailties (Henderson, Diggle, & Dobson, 2000; Hsieh, Tseng, & Wang, 2006; Liu, Ma, & O'Quigley, 2008; Rizopoulos, Molenberghs, & Lesaffre, 2017; Rizopoulos, Taylor, Van Rosmalen, Steyerberg, & Takkenberg, 2015; Rizopoulos, Verbeke, & Molenberghs, 2008; Njagi, Rizopoulos, Molenberghs, Dendale, & Willekens, 2013; Tsiatis & Davidian, 2001; Song, Davidian, & Tsiatis, 2002; Wulfsohn & Tsiatis, 1997). Most of the literature on joint modeling consider a two-level hierarchy, that is, longitudinal data nested within subjects, where a subject-specific RE is typically used to model the dependency between the longitudinal outcome and survival. The few works (Liu et al., 2008) that consider a three-level data structure (with longitudinal outcomes and only subject-level risk factors. More specifically, the work by Liu et al. (2008) does not model the direct effect of the longitudinal outcome on survival but rather builds a dependency in modeling of the two outcomes via multilevel REs and utilizes Gauss quadrature techniques in estimation which do not scale up to the large complex data structure of USRDS data with a large number of facilities. The proposed MJM departs from previous literature in considering multilevel REs (at both the subject-level and dialysis facility-level) and multilevel risk factors in modeling the dependency between a generalized longitudinal outcome (hospitalization) and survival, where the direct effects of the longitudinal outcome on survival is targeted. In addition, a feasible estimation and inference framework is proposed based on fully exponential Laplace approximations that are scalable to estimation in large (USRDS) population data.

Estimation for the proposed MJM is based on an EM algorithm, where the subject- and facility-level REs connecting the two outcomes are considered missing. The expectation step (E-step) estimates the posterior mean and variance of the REs, whereas the maximization step (M-step) maximizes the joint likelihood with respect to model parameters, given the REs. The MJM for our three-level hierarchical data, with longitudinal measurements nested in subjects and subjects clustered in dialysis facilities, leads to a high-dimensional vector of REs (of order $n_i + 1$) at the facility level with the facility-level RE as well as subject-level REs for n; patients receiving dialysis at the ith facility. This is a major computational challenge which has hindered the estimation of joint models with hierarchical data and multilevel REs and is compounded when the size of the data is large. Our analysis of the USRDS data includes over 292,000 observations on ~34,000 patients in >500 facilities, where the number of patients within a facility, denoted by n_i, ranges from 50 to 162. Although the Laplace approximation method has less computational burden than other numerical integration methods such as Gauss guadrature or Monte Carlo approaches in approximating integration of high-dimensional REs, the error associated with the approximation can get large in sparse longitudinal applications with a small number of repeated measurements within subjects. Because the USRDS data have subjects with only a few (<5) repetitions during the follow-up period, we adopt the fully exponential Laplace approximations (Rizopoulos, Verbeke, & Lesaffre, 2009; Tierney, Kass, & Kadane, 1989) to address this major computational challenge, which has been shown to lead to reliable estimation and lower order approximation errors than the standard Laplace approximation in modeling sparse longitudinal outcomes with few repeated measurements within a subject. Our previous work shows reliability and efficacy of the fully exponential Laplace approximations in addressing integration over high-dimensional REs in the context of generalized multilevel varying coefficient models (Li et al., 2018).

Distinct from Rizopoulos et al. (2009), who considered the exponential Laplace approximations for joint modeling with a continuous longitudinal outcome, we demonstrate the use of exponential Laplace approximations in joint modeling of a generalized (e.g., binary) longitudinal outcome with survival, and for a higher level (three-level) hierarchical data which are much larger than data considered in JM contexts. Furthermore, we investigate the appropriateness of the use of model-based standard errors (SEs) in MJM and provide practical guidance.

The remainder of the paper is organized as follows. The proposed MJM and the EM algorithm utilizing fully exponential Laplace approximations and SEs of model parameters are presented in Section 2. Simulation studies to examine the efficacy of estimation and comparison of SE estimates based on likelihood- and bootstrap-based approaches are provided in Section 3. We also compare our proposed MJM with a simplified joint model that ignores the correlation at the highest level of the hierarchy (i.e., at the facility-level) in Section 3. In Section 4, we illustrate the proposed MJM to jointly model longitudinal hospitalization risk and survival using the USRDS data. We conclude with a brief discussion in Section 5.

2 | MJM, ESTIMATION, AND INFERENCE

2.1 | Model formulation

To obtain the joint distribution of the binary longitudinal and survival outcomes, we begin by defining the submodels for each outcome. For the longitudinal submodel, denote the binary longitudinal outcome as $Y_{ij}(t)$ for subject (patient) *j* in cluster (facility) *i* at time *t*. For the USRDS data, the outcome $Y_{ij}(t)$ is defined as the indicator of at least one hospitalization in a 3-month follow-up window with midpoint *t*, for the *j*th patient, *j* = 1, ..., *n_i*, receiving dialysis at the *i*th facility *i* = 1, ..., *n*. Let $X_{ij} = (X_{ij1}, ..., X_{ijp})^T$ and $Z_{i(j)} = \{Z_{i(j)1}, ..., Z_{i(j)q}\}^T$ denote the subject-and facility-level predictor vectors with the corresponding regression coefficients $\beta = (\beta_1, ..., \beta_p)^T$ and $\psi = (\psi_1, ..., \psi_q)^T$, respectively. The USRDS facility-level characteristics, such as the nurse-to-patient ratio, are reported yearly. Hence, $Z_{i(j)}$ denotes those characteristics reported in the calendar year prior to the *j*th patient initiating dialysis and, therefore, carry a second subject index.

The proposed submodel for the longitudinal outcome is a linear mixed effects model:

$$m_{ij}(t) = E\{Y_{ij}(t) | \mathbf{X}_{ij}, \mathbf{Z}_{i(j)}, b_{ij}, \xi_i\} = g^{-1}\{\mathbf{X}_{ij}^{\mathsf{T}} \boldsymbol{\beta} + \mathbf{Z}_{i(i)}^{\mathsf{T}} \boldsymbol{\psi} + \gamma t + b_{ij} + \xi_i\},\$$

where $g(\cdot)$ is the canonical logit link with $g(p) = \log\{p/(1-p)\}$ and b_{ij} and ξ_i denote the subject- and facility-level REs such that $b_{ij} \sim \mathcal{N}(0, \sigma_b^2)$ and $\xi_i \sim \mathcal{N}(0, \sigma_{\xi}^2)$. We assume that the subject- and facility-level REs are independent. Note that the parameter estimates and SEs in joint modeling are reported to be robust to misspecification of the distribution of the REs (Hsieh et al., 2006; Rizopoulos et al., 2008; Song et al., 2002). We make two important remarks regarding the above submodel for the longitudinal process: (1) although our motivating problem involves a binary outcome process, namely, hospitalization, the model is applicable to a generalized outcome; (2) for simplicity of exposition, we describe a common longitudinal model with a linear time effect and a random intercept, although the technical estimation and inference procedures that will be subsequently described can directly accommodate more general formulations of time-dynamic effects.

In the survival submodel, the true and observed event (death) times are denoted by T_{ij}^* and T_{ij} , respectively, where the observed event time is defined as the minimum of the potential censoring time C_{ij} and T_{ij}^* . In addition, $\delta_{ij} = I(T_{ij} \leq C_{ij})$ denotes the event indicator, where $l(\cdot)$ is the indicator function. For the survival submodel, we adopt a proportional hazards model with the hazard of death at time *t*, accounting the hospitalization history up to time *t*, defined as

$$\begin{aligned} h_{ij}\{t|\mathcal{M}_{ij}(t), \mathbf{X}_{ij}, \mathbf{Z}_{i(j)}\} &= \lim_{\Delta_t \to 0} \Pr\{t \leq T_{ij}^* < t + \Delta_t | T_{ij}^* \geq t, \mathcal{M}_{ij}(t), \mathbf{X}_{ij}, \mathbf{Z}_{i(j)}\} \\ &= h_0(t) \exp\{\mathbf{X}_{ij}^\mathsf{T} \boldsymbol{\zeta} + \mathbf{Z}_{i(j)}^\mathsf{T} \boldsymbol{\eta} + \alpha m_{ij}(t)\}, \end{aligned}$$
(1)

where $\mathcal{M}_{ij}(t) = \{m_{ij}(s), 0 \le s < t\}$ is the history of the true unobserved longitudinal process up to the time point $t, \zeta = (\zeta_1, \dots, \zeta_p)^T$ and $\eta = (\eta_1, \dots, \eta_q)^T$ are the multilevel covariate effects on survival, $h_0(\cdot)$ is the baseline hazard function, and α is the regression coefficient that quantifies the effect of the longitudinal outcome on the risk of an event. Thus, in terms of the survival function, we have

$$S_{ij}\{t|\mathcal{M}_{ij}(t), \mathbf{X}_{ij}, \mathbf{Z}_{i(j)}\} = \Pr\{T_{ij}^* > t|\mathcal{M}_{ij}(t), \mathbf{X}_{ij}, \mathbf{Z}_{i(j)}\}$$

$$= \exp\left[-\int_{0}^{t} h_0(s) \exp\{\mathbf{X}_{ij}^\mathsf{T}\boldsymbol{\zeta} + \mathbf{Z}_{i(j)}^\mathsf{T}\boldsymbol{\eta} + \alpha m_{ij}(s)\}ds\right].$$
(2)

The definitions of the hazard and survival functions indicate that the instantaneous probability of death at time *t* depends on the current value of the longitudinal outcome, the hospitalization risk score (1) at time *t*, whereas the survival function depends on the entire history of the hospitalization risk up to time *t*, namely, $M_{ij}(t)$.

Hence, the joint distributions of the survival and longitudinal outcomes, specifically

$$p(T_{ij},\delta_{ij},Y_{ij},b_{ij},\xi_i;\theta) = p(T_{ij},\delta_{ij}|b_{ij},\xi_i;\theta) p(Y_{ij}|b_{ij},\xi_i;\theta) p(b_{ij},\xi_i;\theta),$$
(3)

are connected through the multilevel REs b_{ij} and ξ_i , which not only account for the association between the two outcomes but also explain the correlation between longitudinal measurements within a subject. In (3), Y_{ij} denotes the $n_{ij} \times 1$ vector of the longitudinal outcome for the *j*th subject within the *i*th facility, $\theta = (\theta_t^T, \theta_y^T, \theta_b, \theta_{\xi})^T$ denotes the full parameter vector with survival parameters $\theta_t = (\zeta^T, \eta^T, \alpha, \theta_{h_0})^T$ and longitudinal parameters $\theta_y = (\beta^T, \psi^T, \gamma)^T$, and θ_{h_0} contains the parameters in modeling the baseline hazard function $h_0(\cdot)$, $\theta_b = \sigma_b^2$ and $\theta_{\xi} = \sigma_{\xi}^2$. The density of the observed event time T_{ij} in (3) given the multilevel REs is given as

$$p(T_{ij}, \delta_{ij} | b_{ij}, \xi_i; \theta) = h_{ij} \{ T_{ij} | \mathcal{M}_{ij}(T_{ij}); \theta \}^{\delta_{ij}} S_{ij} \{ T_{ij} | \mathcal{M}_{ij}(T_{ij}); \theta \}$$
$$= \left[h_0(T_{ij}) \exp\{ \mathbf{X}_{ij}^\mathsf{T} \boldsymbol{\zeta} + \mathbf{Z}_{i(j)}^\mathsf{T} \boldsymbol{\eta} + \alpha m_{ij}(T_{ij}) \} \right]^{\delta_{ij}}$$
$$\times \exp\left[-\int_{0}^{T_{ij}} h_0(s) \exp\{ \mathbf{X}_{ij}^\mathsf{T} \boldsymbol{\zeta} + \mathbf{Z}_{i(j)}^\mathsf{T} \boldsymbol{\eta} + \alpha m_{ij}(s) \} ds \right]$$

(5)

In addition, the joint density for the longitudinal outcome and the REs in (3) is

$$\begin{split} p(\mathbf{Y}_{ij}|b_{ij},\xi_i;\theta) \, p(b_{ij},\xi_i;\theta) &= \left[\prod_{k=1}^{n_{ij}} p\{\mathbf{Y}_{ij}(t_{ijk})|b_{ij},\xi_i;\theta_{\mathbf{y}}\}\right] \, p(b_{ij},\xi_i;\theta) \\ &= \left(\prod_{k=1}^{n_{ij}} \frac{\exp\left[\{\mathbf{X}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \mathbf{Z}_{i(j)}^{\mathsf{T}}\boldsymbol{\psi} + \gamma \, t_{ijk} + b_{ij} + \xi_i\}\mathbf{Y}_{ij}(t_{ijk})\right]}{1 + \exp\{\mathbf{X}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \mathbf{Z}_{i(j)}^{\mathsf{T}}\boldsymbol{\psi} + \gamma \, t_{ijk} + b_{ij} + \xi_i\}} (2\pi\sigma_b^2)^{-1/2} \exp\left\{-b_{ij}^2/(2\sigma_b^2)\right\}\right) \\ &\times (2\pi\sigma_{\varepsilon}^2)^{-1/2} \exp\left\{-\xi_i^2/(2\sigma_{\varepsilon}^2)\right\}, \end{split}$$

where t_{ijk} , $k = 1, ..., n_{ij}$, denote the midpoints of the n_{ij} 3-month intervals in the follow-up period of the *j*th patient from the *i*th facility. The longitudinal outcomes, $Y_{ij}(t_{ijk})$, $k = 1, ..., n_{ij}$, within a subject are assumed to be independent conditional on the multilevel REs.

2.2 | Estimation and inference

We propose an approximate EM algorithm (Dempster, Laird, & Rubin, 1977), in which we treat multilevel REs as missing data. The proposed EM iterates between the E-step, targeting the REs and the M-step, maximizing the approximate expected complete likelihood to estimate $\theta = (\theta_{\star}^{T}, \theta_{v}^{T}, \theta_{b}, \theta_{\varepsilon})^{T}$.

Let $\ell(\mathbf{u}, \theta)$ denote the complete joint log-likelihood:

$$\ell(\mathbf{u},\theta) = \sum_{i=1}^{n} \ell_i(\mathbf{u}_i,\theta) = \sum_{i=1}^{n} \log L_i(\mathbf{u}_i,\theta) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log p(T_{ij},\delta_{ij},Y_{ij},b_{ij},\xi_i;\theta),$$
(4)

where $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_n)^T$, $\mathbf{u}_i = (b_{i1}, \dots, b_{in_i}, \xi_i)^T$, and $L_i(\mathbf{u}_i, \theta)$ and $p(T_{ij}, \delta_{ij}, Y_{ij}, b_{ij}, \xi_i; \theta)$ denote the likelihood contribution of the *i*th facility and the *j*th subject within the *i*th facility, respectively. Using (3), $\ell_i(\mathbf{u}_i, \theta)$ can be defined as

$$\begin{split} \ell_{i}(\mathbf{u}_{i},\theta) &= \sum_{j=1}^{n_{i}} \log p(\mathsf{T}_{ij},\delta_{ij},\mathsf{Y}_{ij},b_{ij},\xi_{i};\theta) \\ &= \sum_{j=1}^{n_{i}} \left\{ \log p(\mathsf{T}_{ij},\delta_{ij}|b_{ij},\xi_{i};\theta) + \log p(\mathsf{Y}_{ij}|b_{ij},\xi_{i};\theta) + \log p(b_{ij},\xi_{i};\theta) \right\} \\ &= \sum_{j=1}^{n_{i}} \left\{ \int \delta_{ij} \left[\log h_{0}(\mathsf{T}_{ij}) + \{\mathsf{X}_{ij}^{\mathsf{T}}\boldsymbol{\zeta} + \mathsf{Z}_{i(j)}^{\mathsf{T}}\boldsymbol{\eta} + \alpha m_{ij}(\mathsf{T}_{ij})\} \right] \\ &- \int_{0}^{\mathsf{T}_{ij}} h_{0}(s) \exp\{\mathsf{X}_{ij}^{\mathsf{T}}\boldsymbol{\zeta} + \mathsf{Z}_{i(j)}^{\mathsf{T}}\boldsymbol{\eta} + \alpha m_{ij}(s)\} ds \right) \\ &+ \left[\left\{ \sum_{k=1}^{n_{i}} g(m_{ijk}) \mathsf{Y}_{ijk} + \log(q_{ijk}) - \frac{b_{ij}^{2}}{2\sigma_{b}^{2}} - \frac{1}{2} \log(2\pi\sigma_{b}^{2}) \right\} - \frac{\xi_{i}^{2}}{2}\sigma_{\xi}^{2} - \frac{1}{2} \log(2\pi\sigma_{\xi}^{2}) \right] \right], \end{split}$$

where $m_{ij}(t) = g^{-1} \{ \mathbf{X}_{ij}^{\mathsf{T}} \boldsymbol{\beta} + \mathbf{Z}_{i(j)}^{\mathsf{T}} \boldsymbol{\psi} + \gamma t + b_{ij} + \xi_i \}$, $m_{ijk} = m_{ij}(t_{ijk})$, $Y_{ijk} = Y_{ij}(t_{ijk})$ and $q_{ijk} = 1 - m_{ijk}$. The incomplete likelihood used in defining the expected value and variance of the REs is given by $L(\theta) = \sum_{i=1}^{n} \int L_i(\mathbf{u}_i, \theta) du_i$.

The steps of the proposed EM algorithm are outlined below.

- 1. The initial values for all the parameters, denoted by $\theta^0 = (\theta_t^{0T}, \theta_y^{0T}, \theta_b^0, \theta_{\xi}^0)^T$, are set to initial estimates from separate model fits to the longitudinal and the survival outcome. More specifically, a generalized multilevel linear mixed effect model is fitted to the longitudinal outcome, and a Cox model (Cox, 1972) is fitted to the survival outcome.
- 2. E-step: in the ℓ th iteration, the estimates of the posterior mean and variance of the REs in $\mathbf{u}_i = (b_{i1}, \dots, b_{in_i}, \xi_i)^T$ are obtained via fully exponential Laplace approximations, leading to the approximated expected complete likelihood.
- 3. M-step: the expected complete likelihood is maximized to obtain closed form solutions for the current estimates of θ_b and θ_{ξ} ($\theta_b = \sigma_b^2$ and $\theta_{\xi} = \sigma_{\xi}^2$). The approximate expected complete likelihood is maximized to obtain the rest of the current estimates of $\theta^{\gamma\sigma} = (\theta_t, \theta_y)^T$ via a Newton-Raphson algorithm.
- The algorithm iterates between Steps 2 and 3 until the difference between two consecutive complete log-likelihood values are less than a predefined tolerance level *ε*.

2.2.1 | E-step and fully exponential laplace approximation

The posterior mean \mathbf{u}_{i0} and variance \mathbf{v}_{i0} of \mathbf{u}_i are defined as

$$\mathbf{u}_{i0} = \frac{\int \mathbf{u}_i L_i(\mathbf{u}_i, \theta) d\mathbf{u}_i}{\int L_i(\mathbf{u}_i, \theta) d\mathbf{u}_i} \quad \text{and} \quad \mathbf{v}_{i0} = \frac{\int (\mathbf{u}_i - \mathbf{u}_{i0}) (\mathbf{u}_i - \mathbf{u}_{i0})^{\mathsf{T}} L_i(\mathbf{u}_i, \theta) d\mathbf{u}_i}{\int L_i(\mathbf{u}_i, \theta) d\mathbf{u}_i}.$$
 (6)

Note that the integrals in (6) are taken with respect to the potentially high-dimensional REs vector \mathbf{u}_i (of dimensions 51–163 in our data application) and their closed form does not exist. For approximating the high-dimensional integrals in (6), we employ the fully exponential Laplace approximation (Tierney et al., 1989). The fully exponential Laplace approximation can only be applied to strictly positive functions, and the integrands in the estimation of \mathbf{u}_{i0} might not always satisfy this condition. Therefore, adopting the approach of Rizopoulos et al. (2009), we estimate the posterior mean and variance through targeting $E\{\exp(\mathbf{c}^{\mathsf{T}}\mathbf{u}_i)\}$ (where $\mathbf{c} = (c_1, \ldots, c_{n_i+1})^{\mathsf{T}}$ is a constant vector), which is always positive. More specifically, \mathbf{u}_{i0} and \mathbf{v}_{i0} are obtained using the cumulant generating function $\log [E\{\exp(\mathbf{c}^{\mathsf{T}}\mathbf{u}_i)\}]$ via $\mathbf{u}_{i0} = \partial \log [E\{\exp(\mathbf{c}^{\mathsf{T}}\mathbf{u}_i)\}] /\partial \mathbf{c}^{\mathsf{T}}|_{\mathbf{c}=0}$ and $\mathbf{v}_{i0} = \partial^2 \log [E\{\exp(\mathbf{c}^{\mathsf{T}}\mathbf{u}_i)\}] /\partial \mathbf{c}^{\mathsf{T}}\partial \mathbf{c}|_{\mathbf{c}=0}$. The fully exponential Laplace approximation is performed in two steps: (1) the mode of \mathbf{u}_i is estimated by maximizing the approximate complete likelihood via a Newton-Raphson algorithm, and (2) the mode from the first step is used to obtain \mathbf{u}_{i0} and \mathbf{v}_{i0} via differentiating the cumulant-generating function and evaluating at $\mathbf{c} = 0$.

In the first step, the mode $\hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i^{(c)}|_{\mathbf{c}=0}$, where $\hat{\mathbf{u}}_i^{(c)} = \operatorname{argmax}_{\mathbf{u}_i} \{\ell_i(\mathbf{u}_i, \theta) + \mathbf{c}^T \mathbf{u}_i\}$, is obtained. Maximization is implemented via a safeguarded Newton-Raphson algorithm where at the (*it* + 1)th iteration, $\hat{\mathbf{u}}_i$ is updated through

$$\hat{\mathbf{u}}_i^{it+1} = \hat{\mathbf{u}}_i^{it} - s(\boldsymbol{\Sigma}_i^{it})^{-1} \mathcal{J}(\hat{\mathbf{u}}_i^{it}), \tag{7}$$

with $\Sigma_i^{it} = \Sigma_i^{(c)}|_{(\mathbf{c},\mathbf{u}_i)=(0,\hat{u}_i^{it})}, \Sigma_i^{(c)} = -\partial^2 \{\ell_i(\mathbf{u}_i,\theta) + \mathbf{c}^T \mathbf{u}_i\}/\partial \mathbf{u}_i^T \partial \mathbf{u}_i = -\partial^2 \ell_i(\mathbf{u}_i,\theta)/\partial \mathbf{u}_i^T \partial \mathbf{u}_i, \mathcal{J}(\hat{\mathbf{u}}_i^{it}) = -\partial \ell_i(\mathbf{u}_i,\theta)/\partial \mathbf{u}_i^T|_{\mathbf{u}_i=\hat{\mathbf{u}}_i}, \text{ and } s \text{ denoting the step size used along the Newton-Raphson updating direction. The estimated mode from the first step is used in targeting the posterior mean and variance of <math>\mathbf{u}_i$ in the second step by differentiating the cumulant-generating function and evaluating at $\mathbf{c} = 0$:

$$\mathbf{u}_{i0} = \hat{\mathbf{u}}_i - \frac{1}{2} \operatorname{tr}(\mathcal{V}), \qquad \mathbf{v}_{i0} = \boldsymbol{\Sigma}_i^{-1} - \frac{1}{2} \operatorname{tr} \left\{ -\mathcal{V}\mathcal{V}^{\mathsf{T}} + \boldsymbol{\Sigma}_i^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_i^{(c)}}{\partial \mathbf{c}^{\mathsf{T}} \partial \mathbf{c}} \Big|_{(\mathbf{c}, \mathbf{u}_i) = (0, \hat{\mathbf{u}}_i)} \right\},$$
(8)

where $\mathcal{V} = \Sigma_i^{-1} \{\partial \Sigma_i^{(c)} / \partial \mathbf{c}^T\}|_{(\mathbf{c}, \mathbf{u}_i) = (0, \hat{\mathbf{u}}_i)}, \Sigma_i = \Sigma_i^{(c)}|_{\mathbf{c}=0}$, and $\hat{\mathbf{u}}_i$ and Σ_i^{-1} denote $\hat{\mathbf{u}}_i^{it}$ and the inverse of Σ_i^{it} from the last iteration of the Newton-Raphson algorithm, respectively. Details of the fully exponential Laplace algorithm are presented in the supporting information.

After estimating the posterior mean and variance of \mathbf{u}_i , the expectation of the complete joint likelihood is approximated in the E-step. Let $\theta^* = (\theta_t^{*T}, \theta_y^{*T}, \theta_b^*, \theta_{\xi}^*)^T$ denote the current parameter estimates with $\theta_t^{*T} = (\boldsymbol{\zeta}^{*T}, \boldsymbol{\eta}^{*T}, \alpha^*, \theta_{h_0}^*)$, $\theta_y^{*T} = (\boldsymbol{\beta}^{*T}, \boldsymbol{\psi}^{*T}, \gamma^*)$, $\theta_b^* = \sigma_b^{*2}$, and $\theta_{\xi}^* = \sigma_{\xi}^{*2}$. Because the closed form expression for $\sum_{i=1}^{n} E\{\ell_i(\mathbf{u}_i, \theta) | Y_i, T_i, \delta_{ij}, \mathbf{X}_i, \mathbf{Z}_{i(j)}, \theta^*\}$ is intractable, we approximate the expected log-likelihood via a second degree Taylor's expansion around \mathbf{u}_{ϕ}^* :

$$\begin{split} &\sum_{i=1}^{n} \ell_{i}^{\prime} (\mathbf{u}_{i0}^{*}, \theta^{*}) + \ell_{i}^{\prime} (\mathbf{u}_{i0}^{*}, \theta^{*}) E(\mathbf{u}_{i} - \mathbf{u}_{i0}^{*}) - \frac{1}{2} E(\mathbf{u}_{i} - \mathbf{u}_{i0}^{*})^{\mathsf{T}} \Sigma_{i}^{*} E(\mathbf{u}_{i} - \mathbf{u}_{i0}^{*}) \\ &= \sum_{i=1}^{n} \left[\sum_{j=1}^{n_{i}} \left(\delta_{ij}^{*} \left[\log h_{0}^{*}(T_{ij}) + \{ \mathbf{X}_{ij}^{\mathsf{T}} \zeta^{*} + \mathbf{Z}_{i(j)}^{\mathsf{T}} \boldsymbol{\eta}^{*} + \alpha^{*} m_{ij}^{*}(T_{ij}) \} \right] \\ &- \int_{0}^{T_{ij}} h_{0}^{*}(s) \exp\{ \mathbf{X}_{ij}^{\mathsf{T}} \zeta^{*} + \mathbf{Z}_{i(j)}^{\mathsf{T}} \boldsymbol{\eta}^{*} + \alpha^{*} m_{ij}^{*}(s) \} ds \\ &+ \left\{ \sum_{k=1}^{n_{i}} Y_{ijk} \{ g(m_{ijk}^{*}) \} + \log(q_{ijk}^{*}) - \frac{(b_{0ij}^{*})^{2} + v_{b,ij0}^{*}}{2\sigma_{b}^{*2}} - \frac{1}{2} \log(2\pi\sigma_{b}^{*2}) \right\} \\ &- \frac{(\xi_{i0}^{*})^{2} + v_{\xi,i0}^{*}}{2\sigma_{\xi}^{*2}} - \frac{1}{2} \log(2\pi\sigma_{\xi}^{*2}) - \frac{v_{b,ij0}^{*} + 2r_{ij0}^{*} + v_{\xi,i0}^{*}}{2} B_{ij}^{*} \right) \right], \end{split}$$

where $m_{ij}^*(t) = g^{-1} \{ \mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \mathbf{Z}_{i(j)}^T \boldsymbol{\psi}^* + \gamma^* t + b_{ij0}^* + \xi_{i0}^* \}$, $m_{ijk}^* = m_{ij}^*(t_{ijk})$, $q_{ijk}^* = 1 - m_{ijk}^*$, and B_{ij}^* is B_{ij} (defined in the supporting information) evaluated at θ^* , $\Sigma_i = -\partial^2 \ell_i / \partial \mathbf{u}_i^T \partial \mathbf{u}_i |_{\mathbf{u}_i = \hat{\mathbf{u}}_i, \theta = \theta^*}$, and $E(\mathbf{u}_i - \mathbf{u}_{i0}^*) = 0$. Furthermore, in (9), $\mathbf{u}_{i0}^* = (b_{i10}^*, \dots, b_{inj0}^*, \xi_{i0}^*)^T$, $\mathbf{v}_{\xi,i0}^*$, and r_{ij0}^* denote the estimated posterior mean of \mathbf{u}_i , posterior variance of the subject- and facility-level REs, and posterior covariance of subject- and facility-level REs based on the current parameter estimates, respectively.

2.2.2 | M-step

For estimation of the variance components, σ_b^2 and σ_{ξ}^2 , the incomplete log-likelihood $\ell(\theta) = \log L(\theta)$ is maximized directly, by setting the score functions to zero. The score functions of the incomplete log-likelihood with respect to σ_b^2 and σ_{ξ}^2 can be given as

$$V(\sigma_b^2) = \frac{\partial \ell(\theta)}{\partial \sigma_b^2} = \sum_{i=1}^n \frac{\partial}{\partial \sigma_b^2} \log \left\{ \int L_i(\mathbf{u}_i, \theta) du_i \right\} = \sum_{i=1}^n \int \sum_{j=1}^n \left(\frac{b_{ij}^2}{2\sigma_b^2} - \frac{1}{\sigma_b^2} \right) \mathcal{P}(\mathbf{u}_i) du_i = \sum_{i=1}^n V_i(\sigma_b^2)$$

and

$$V(\sigma_{\xi}^{2}) = \frac{\partial \ell(\theta)}{\partial \sigma_{\xi}^{2}} = \sum_{i=1}^{n} \frac{\partial}{\partial \sigma_{\xi}^{2}} \log \left\{ \int L_{i}(\mathbf{u}_{i}, \theta) du_{i} \right\} = \sum_{i=1}^{n} \int \left(\frac{\xi_{i}^{2}}{2} \sigma_{\xi}^{2} - \frac{1}{\sigma_{\xi}^{2}} \right) \mathcal{P}(\mathbf{u}_{i}) du_{i} = \sum_{i=1}^{n} V_{i}(\sigma_{\xi}^{2}),$$

where $\mathcal{P}(\mathbf{u}_i) = L_i(\mathbf{u}_i, \theta) / \int L_i(\mathbf{u}_i, \theta)$ denotes the posterior density of \mathbf{u}_i . Setting the above score functions to zero leads to the following estimates of σ_k^2 and σ_z^2 at the current iteration:

$$\sigma_b^{*2} = \left(\sum_{i=1}^n n_i\right)^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ (b_{ij0}^*)^2 + \mathbf{v}_{b,ij0}^* \right\} \quad \text{and} \quad \sigma_{\xi}^{*2} = n^{-1} \sum_{i=1}^n \left\{ (\xi_{i0}^*)^2 + \mathbf{v}_{\xi,i0}^* \right\}. \tag{10}$$

The likelihood-based SEs for $\hat{\sigma}_b^2$ and $\hat{\sigma}_{\xi}^2$ are equal to the diagonal elements of $(\sum_{i=1}^n \mathbf{V}_i \mathbf{V}_i^T)^{-1}$, where $\mathbf{V}_i = \{V_i(\hat{\sigma}_b^2), V_i(\hat{\sigma}_{\xi}^2)\}^T$.

The closed form solutions for the rest of the parameters, namely, θ_t and θ_y , cannot be obtained by maximizing the incomplete log-likelihood directly. Therefore, we employ a Newton-Raphson algorithm to maximize the approximated expected log-likelihood given in (9) and obtain estimates for the parameters $\theta^{,\sigma} = (\theta_t, \theta_y)^T$,

$$\theta^{\circ\sigma(it+1)} = \theta^{\circ\sigma(it)} - \{\mathcal{H}^{\circ\sigma(it)}\}^{-1} \mathbf{V}^{\circ\sigma(it)},\tag{11}$$

where *it* is the current iteration of the Newton-Raphson algorithm and $\mathbf{V}^{\sigma(it)}$ and $\mathcal{H}^{\gamma\sigma(it)}$ are the score function and the hessian of the approximated expected log-likelihood (9) with respect to $\theta^{\gamma\sigma}$, respectively, evaluated at the current estimates $\theta^{\sigma(it)}$. The likelihood-based SEs for $\theta^{\gamma\sigma} = (\theta_t, \theta_y)^T$ are equal to the diagonal elements of $(\sum_{i=1}^{n} \mathbf{V}_i^{\sigma} \mathbf{V}_i^{\gamma\sigma^T})^{-1}$, where $\mathbf{V}_i^{\gamma\sigma}$ contains the score values from the last iteration. Full definitions of the score function and the hessian are provided in the supporting information.

The likelihood-based SEs of the estimators in the EM algorithm are expected to be biased in estimating the true SEs because the variability in the estimation of the REs is not taken into account, similar to findings in previous works of Hsieh et al. (2006) and Kass and Steffey (1989) on joint/hierarchical modeling. Therefore, we examine the extent of this bias in the likelihood-based SEs via simulations in Section 3. Furthermore, we study the utility of the bootstrap estimates of SEs for MJM in simulation studies to provide guidance in practice.

In implementing the proposed approach, a suitable baseline hazard function $h_0(t)$ needs to be selected. We used the P-splines approach (Eilers & Marx, 1996), which provides a flexible specification for the baseline risk function. In particular, $\log\{h_0(t)\} = \sum_{m=1}^{M} \tau_m B_m(t)$, where τ_m denotes the coefficients for the baseline hazard and $B_m(t)$ is the *m*th basis function of a B-spline. Under this baseline hazard definition, the aforementioned parameter θ_{h_0} is $\boldsymbol{\tau} = (\tau_1, \dots, \tau_M)^T$, and it will be estimated as a part of the EM algorithm described above. The smoothness of the baseline hazard is achieved by a differencing penalty, which is subtracted from the log-likelihood defined in (4), $\ell^*(\mathbf{u}, \theta) = \ell(\mathbf{u}, \theta) - \lambda \mathbf{D}^T \mathbf{D}/2$ with λ as the penalty parameter and \mathbf{D} as a second-order difference matrix (Eilers, Marx, & Durbn, 2015). The above formulations and solutions stay the same under this formulation with the replacement of $\ell(\mathbf{u}, \theta)$ by $\ell^*(\mathbf{u}, \theta)$ and of $L(\mathbf{u}, \theta)$ by $L^*(\mathbf{u}, \theta) = \exp\{\ell^*(\mathbf{u}, \theta)\}$. For choosing the penalty parameter λ , we follow Eilers and Marx (2010) on studying the shape of the estimated log-likelihood as a function of λ . Note that other parametric and nonparametric forms for the baseline hazard function can also be accommodated in the proposed estimation and inference procedures.

3 | SIMULATION STUDIES

We conducted simulation studies to assess efficacy of the proposed model parameter estimates and the likelihood-based and bootstrap-based estimates of SEs. Performance of MJM was studied in two simulation scenarios with n = 200 and n = 500 facilities. Results reported for each case were based on 150 Monte Carlo datasets. The number of subjects within facilities was simulated from a discrete uniform distribution on the interval [50, 100] with 75 patients per facility on average, similar to the USRDS dialysis population. The maximum number of repeated measurements per subject was taken to be 20, mimicking the total number of longitudinal observations in the USRDS data. The longitudinal observations within an individual were equally spaced on the interval [0, 1] before censoring by survival. Among the subjects who were not censored, the number of observations varied from 1 to 17 with an average of 5.

The subject-level covariates, $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij})^{\mathsf{T}}$, were generated from normal distributions with means 0 and 1.5 and variances 1 and 0.5, respectively. Similarly, the facility-level covariates, $\mathbf{Z}_{i(j)} = \{Z_{1i(j)}, Z_{2i(j)}\}^{\mathsf{T}}$, were simulated from normal distributions with means –0.3 and 0 and variances 1 and 0.5, respectively. The parameters in (3) equaled $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathsf{T}} = (1.0, 0.2, -1.5)^{\mathsf{T}}, \boldsymbol{\psi} = (\psi_1, \psi_2)^{\mathsf{T}} = (2.5, 3.0)^{\mathsf{T}}, \boldsymbol{\gamma} = 0.5,$ $\boldsymbol{\eta} = (\eta_1, \eta_2)^{\mathsf{T}} = (0.30, 0.80)^{\mathsf{T}}, \boldsymbol{\zeta} = (\zeta_1, \zeta_2)^{\mathsf{T}} = (0.15, 0.10)^{\mathsf{T}}, \text{ and } \boldsymbol{\alpha} = 0.5$. The Weibull function with $\lambda = 1.1$ was used to generate the baseline hazard $h_0(t)$. The subject- and facility-level REs were independently simulated from a normal distribution with mean zero and variances 1 and 0.5, respectively.

The longitudinal outcome $Y_{ij}(\cdot)$ was generated via an underlying normal latent variable $Y_{ij}^*(\cdot)$, where $Y_{ij}(\cdot) = I\{Y_{ij}^*(\cdot) > 0\}$, and the mean of $Y_{ij}^*(\cdot)$ was determined by the longitudinal submodel. For the survival submodel, the true event times, T_{ij}^* , for subjects were generated using the inverse probability integral transformation with Weibull baseline hazard (Bender, Augustin, & Blettner, 2005). Following the descriptions in Section 2.1, the observed time and the event indicator were obtained as $T_{ij} = \min(C_{ij}, T_{ij}^*)$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$, respectively. The resulting overall hospitalization and censoring rates were approximately 32% and 67%, respectively, similar to the USRDS data application.

The baseline hazard function was estimated via the P-splines approach described in Section 2.2.2, with 18 and 28 equally spaced knots and $\lambda = 10$ and 6 for n = 200 and n = 500, respectively. Figure 1 displays the estimated baseline hazard functions for n = 200 and n = 500, along with bands representing the 2.5th and 97.5th pointwise percentiles. Estimates have smaller bias and variance as the number of facilities increases.



FIGURE 1 Estimated baseline hazard functions (dashed) in the simulation study overlaying the true functions (solid) along with along with the 2.5 and 97.5 percentiles based on 150 Monte Carlo runs (shaded) for (a) n = 200 and (b) n = 500

	True	Bias	SD	SE (SD _{SE})	$Boot_{SE}(Boot_{SD_{SE}})$	
		Number of facilities, $n = 200$				
Longitudinal						
β_0	1	-0.013	0.016	0.022 (2.11×10^{-5})	0.015 (0.002)	
β_1	0.2	-0.004	0.007	0.006 (1.47 × 10 ⁻⁶)	0.006 (0.001)	
β_2	-1.5	0.005	0.006	0.013 (6.39 × 10 ⁻⁶)	0.005 (0.001)	
ψ_1	2.5	-0.016	0.022	$0.013~(1.61 imes10^{-5})$	0.019 (0.004)	
ψ_2	3	-0.013	0.023	$0.018~(3.01 imes10^{-5})$	0.024 (0.004)	
γ	0.5	0.009	0.014	$0.013 (1.42 imes 10^{-5})$	0.014 (0.004)	
Survival						
ζ1	0.15	-0.006	0.012	0.011 ($4.05 imes 10^{-6}$)	0.011 (0.002)	
ζ2	-0.1	0.005	0.011	0.024 (3.21 × 10 ⁻⁵)	0.010 (0.001)	
η_1	0.3	0.004	0.016	0.023 (8.22×10^{-5})	0.013 (0.002)	
η_2	0.8	-0.009	0.021	$0.035~(1.48 imes10^{-4})$	0.022 (0.004)	
α	0.5	0.011	0.022	$0.048~(1.95 imes10^{-4})$	0.023 (0.003)	
Variance components						
σ_b^2	1	0.009	0.010	0.012 (3.33 × 10 ⁻⁶)	0.010 (0.001)	
σ_{ε}^2	0.5	0.008	0.022	$0.053~(5.24 imes10^{-4})$	0.028 (0.008)	
*		Number of	facilities, r	n = 500		
Longitudinal						
β_0	1	-0.007	0.009	0.014 (3.37 × 10 ⁻⁶)	0.009 (0.002)	
β_1	0.2	-0.001	0.003	$0.004 (2.06 \times 10^{-7})$	0.003 (0.001)	
β_2	-1.5	-0.001	0.002	0.008 (2.78 × 10 ⁻⁶)	0.002 (0.001)	
ψ_1	2.5	-0.010	0.015	0.008 (2.43 × 10 ⁻⁶)	0.016 (0.005)	
ψ_2	3	-0.009	0.012	$0.012 (5.85 \times 10^{-6})$	0.013 (0.003)	
γ	0.5	0.006	0.010	0.012 (1.84×10^{-6})	0.010 (0.003)	
Survival						
ζ1	0.15	0.001	0.006	$0.012 (1.65 imes 10^{-5})$	0.005 (0.001)	
ζ2	-0.1	0.002	0.005	$0.006 (2.35 \times 10^{-5})$	0.004 (0.001)	
η_1	0.3	0.002	0.007	0.012 (1.20×10^{-5})	0.004 (0.002)	
η_2	0.8	0.005	0.008	0.018 (2.13×10^{-5})	0.012 (0.003)	
α	0.5	0.009	0.013	$0.019 (1.76 imes 10^{-4})$	0.013 (0.003)	
Variance components						
σ_{h}^{2}	1	0.006	0.008	$0.007 (9.26 \times 10^{-7})$	0.006 (0.002)	
σ_{ϵ}^{2}	0.5	0.005	0.014	$0.033 (1.42 \times 10^{-4})$	0.018 (0.003)	

Note. Given are bias, standard deviation (SD), likelihood-based standard errors (SE), and bootstrap SE (Boot_{SE}). Given in parentheses (SD_{SE} and Boot_{SD_{SE}}) are standard deviations of the corresponding quantities.

TABLE 1Simulation results for 200 and500 facilities averaged over 150 datasets

The simulation results are presented in Table 1 for n = 200 and n = 500. The "true" standard deviations (SDs) of the proposed estimators, calculated based on 150 Monte Carlo runs, are denoted by SD. The sample average and the sample SD of 150 estimated likelihood-based SEs are denoted by SE and SD_{SE}, respectively. In addition, Boot_{SE} and Boot_{SD_{SE}} denote the bootstrap-based SEs and their sample SDs. The estimation bias of all the proposed estimators is relatively small and is less than the corresponding SD, indicating that our proposed estimation procedure performs well with respect to estimation of the model coefficients. However, with respect to SE estimates needed for inference, as explained in Section 2.2.2, the EM algorithm framework does not take into account the variability in estimation of the REs leading to potential bias in estimation of SEs. Indeed, this was verified for MJM and summarized in Table 1 where the likelihood-based SEs (Boot_{SE}) work reasonably well in estimating the true SE (the difference between Boot_{SE} and SD is smaller than twice Boot_{SD_{SE}}). These results hold across the two simulation scenarios, where the estimation bias, SD and SE estimates get smaller as the number of facilities increases, as expected. Thus, based on these simulation results, bootstrap estimates of SEs are more suitable for use in practice. We applied these findings to use bootstrap SEs to form confidence intervals for MJM estimates in the data analysis next.

We also compare our proposed MJM with a simplified two-level joint model that ignores the correlation at the highest level of the hierarchy (i.e., at the facility-level), which avoids integration over high-dimensional REs. The data for this simulation were generated using the setup described above under two different (increasing) facility-level REs variances: $\sigma_{\xi}^2 = 0.5$ and $\sigma_{\xi}^2 = 1$. As the simple joint model ignores the correlation at the facility-level, avoiding integration over high-dimensional REs, the Gauss-quadrature method was used in the integration of the subject-level REs. The results are presented in Table 2. These simulation studies indicate that ignoring the correlation at the highest level of the hierarchy (i.e., at the facility-level) leads to higher bias and lower overall efficiency (higher mean square error [MSE]) in estimation of model parameters. The highest bias and MSE are observed in estimation of the covariate effects at the facility level in the longitudinal model and the bias and MSE increase with increasing facility-level variance, as expected. Li et al. (2018) also report lower efficiency for ignoring the correlation at the facility-level in

TABLE 2Simulation results for 200facilities obtained using MJM and asimple joint model (SJM; with onlysubject-level random effects)averaged over 150 datasets

	мім				SIM		
	True	Bias	SD	MSE	Bias	SD	MSE
			$\sigma_{\xi}^2 = 0.$	5			
Longitudinal			2				
β_0	1	-0.013	0.016	4.3×10^{-4}	0.030	0.047	0.003
β_1	0.2	-0.004	0.007	6.5×10^{-5}	-0.003	0.016	2.6×10^{-4}
β_2	-1.5	0.005	0.006	6.1×10^{-5}	0.009	0.018	4.1×10^{-4}
ψ_1	2.5	-0.016	0.022	7.4×10^{-4}	-0.042	0.089	0.010
ψ_2	3	-0.013	0.023	$7.0 imes 10^{-4}$	-0.040	0.116	0.015
γ	0.5	0.009	0.014	$2.8 imes 10^{-4}$	0.012	0.036	0.001
Survival							
ζ1	0.15	-0.006	0.012	$1.8 imes 10^{-4}$	-0.004	0.017	3.1×10^{-4}
ζ2	-0.1	0.005	0.011	$1.5 imes 10^{-4}$	-0.002	0.017	$3.0 imes 10^{-4}$
η_1	0.3	0.004	0.016	2.7×10^{-4}	0.004	0.027	$7.5 imes 10^{-4}$
η_2	0.8	-0.009	0.021	5.2×10^{-4}	0.005	0.051	0.003
α	0.5	0.011	0.022	6.1×10^{-4}	0.014	0.040	0.002
Variance components							
σ_b^2	1	0.009	0.010	1.8×10^{-4}	0.272	0.033	0.075
σ_{ξ}^2	0.5	0.008	0.022	5.5×10^{-4}			
			$\sigma_{\xi}^2 = 1$				
Longitudinal							
β_0	1	-0.016	0.020	$6.6 imes 10^{-4}$	0.041	0.050	0.004
β_1	0.2	-0.008	0.010	$1.6 imes 10^{-4}$	-0.005	0.020	4.3×10^{-4}
β_2	-1.5	0.003	0.007	5.8×10^{-5}	0.010	0.028	$8.8 imes10^{-4}$
ψ_1	2.5	-0.015	0.025	8.5×10^{-4}	-0.083	0.117	0.021
ψ_2	3	-0.014	0.026	8.7×10^{-4}	-0.095	0.149	0.031
γ	0.5	0.010	0.020	5.0×10^{-4}	0.021	0.057	0.004
Survival							
ζ1	0.15	-0.005	0.014	2.2×10^{-4}	-0.003	0.022	$4.9 imes 10^{-4}$
ζ2	-0.1	0.007	0.012	$1.9 imes 10^{-4}$	-0.008	0.029	9.1×10^{-4}
η_1	0.3	0.005	0.018	3.5×10^{-4}	0.003	0.033	0.001
η_2	0.8	-0.008	0.022	5.5×10^{-4}	0.004	0.069	0.005
α	0.5	0.010	0.024	$6.8 imes 10^{-4}$	0.019	0.062	0.004
Variance components							
σ_b^2	1	0.006	0.011	$1.6 imes 10^{-4}$	0.515	0.077	0.271
σ_{ϵ}^2	1	0.012	0.030	0.001			

Note. Given are bias, standard deviation (SD), and mean-squared error (MSE).

estimation of facility-level covariate effects for multilevel varying coefficient models. More specifically, the simplified joint model yields higher MSEs for all estimates compared to the proposed MJM, and the biases of the estimated coefficients of the facility-level covariates are most severe (relative to subject-level covariates) in the longitudinal model of the simplified joint model. Estimation of the variance of the subject-level REs is also severely biased yielding high MSE values. Additional simulation results, on parametric estimation of the baseline hazard function, are reported in the supporting information, Appendix D.

4 | JOINT MODELING OF HOSPITALIZATION AND SURVIVAL OUTCOMES

4.1 | Study cohort and patient- and facility-level risk factors

Data on U.S. patients with ESRD who are on dialysis are captured in the United States Renal Data System (USRDS), including patient outcomes of hospitalization events over time and mortality. We applied the proposed MJM to jointly model longitudinal hospitalization and survival outcomes using a cohort of incident ESRD patients age 18 or older, initiating dialysis between January 1, 2006 and December 31, 2008. Study patients were followed for a maximum of 5 years, with the last date of follow-up as December 31, 2013, where follow-up was truncated if a patient switched dialysis facilities. Basic inclusion criteria required that a patient survived the first 90 days, did not recover the kidney function, did not have a kidney transplant, and was covered by Medicare as primary payer on Day 91. Thus, the first day of study follow up began on Day 91, as recommended by the USRDS Researcher's Guide "90-day rule" to allow for completion of the Medicare eligibility application process and establishment of stable dialysis treatment modality, and, furthermore, because USRDS hospitalization data are incomplete for non-Medicare patients (Chen et al., 2019; United States Renal Data System, 2014).

To illustrate the proposed MJM, the analysis cohort included 292,672 observations over time on 34,030 patients in 520 dialysis facilities, where the number of patients per facility ranged from 50 to 162 (median 61, Q1–Q3 [first-third quartile]: 54–71). The longitudinal part of the MJM included time (months); patient-level covariates of age, sex, and baseline comorbidities of chronic obstructive pulmonary disease (COPD), coagulopathy, cardiorespiratory failure, septicemia, and other infectious diseases (and pneumonias); and psychiatric disorders. These common

TABLE 3

Variable	Mean/count	SD/percent
Age	65.01	15.08
Female	15,374	45.18
COPD	6,364	18.70
Coagulopathy	2,688	7.90
Cardiorespiratory failure	4,066	11.95
Other Infectious disease and pneumonias	7,851	23.07
Psychiatric comorbidity	3,811	11.20
Septicemia	3,462	10.17
Percent nurse-to-patients	7.60	3.20
Percent PCT-to-patients	9.39	2.86

factors

Summary of patient-level and dialysis facility-level risk

Note. COPD, chronic obstructive pulmonary disease; PCT, patient care technician.



FIGURE 2 (Left) Distribution of the mean lengths of patient follow-up among the 520 facilities. (Right) Follow-up times of 61 patients in a randomly selected dialysis facility with longitudinal hospitalizations marked by black circles

comorbidities in the dialysis patients were determined using International Classification of Disease, Ninth Revision (ICD-9) diagnosis codes from institutional claims data 12 months prior to the start of dialysis treatment. Facility-level risk factors included the percentages of nurse-to-patient and patient care technician (PCT)-to-patient. The same patient- and facility-level risk factors were included in the survival submodel of the MJM to assess their joint contribution to longitudinal hospitalization and survival outcomes. As described by (2), the survival component of the MJM also included the longitudinal outcome risk score history up to time *t*. We report Cls for multilevel risk factors using the bootstrap-based SEs.

4.2 | Results

4.2.1 | Background/descriptive analysis

The study cohort included patients with mean age of 65 years old (SD 15) where 45% were females. Common serious baseline comorbidities in the ESRD patients included chronic obstructive pulmonary disease (COPD; 18.7%), septicemia (10.2%), other infectious diseases (23.1%), cardiorespiratory failure (12%), coagulopathy (7.9%), and psychiatric conditions (11.2%). On average, the percent of nurse-to-patient and PCT-to-patient (facility-level covariates) were 7.6 (SD 3.2) and 9.4 (SD 2.9), respectively. See Table 3 for details.

The median length of patient follow-up among the 520 facilities is 24.3 months (Q1–Q3: 21.1 to 27.4 months; Figure 2). The mean number of hospitalizations per person-year is 1.8 (SD 2.2). Figure 2 (right) shows the longitudinal hospitalizations during the study follow-up periods for 61 patients for a randomly selected facility, illustrating the typical high frequency of hospitalization for ESRD patients. As mentioned earlier, the risk of mortality is also particularly high, where median marginal (unadjusted) survival is 46.5 months (3.9 years), and survival drops markedly for patients with serious baseline comorbidities, such as ESRD patients with chronic conditions, for example, COPD, and/or septicemia.

4.2.2 | MJM results

The MJM allows for explicit modeling of patient- and facility-level variation, and the results showed that there were both significant within-subject and among facility variation: $\hat{\sigma}_b^2 = 1.43$ (95% CI 1.22–1.64) and $\hat{\sigma}_{\xi}^2 = 0.30$ (95% CI 0.11–0.48). Thus, about 17% of the estimated total variation was observed at the facility level.

TABLE 4 Multilevel joint model estimates of patient- and facility-level effects on longitudinal hospitalizations and survival

ongitudinal hospitalization			
Variable	OR	Lower CL	Upper CL
Time (Months)	1.1915ª	1.1051	1.2846
Age	1.0351 ^b	0.9943	1.0775
Female	1.2664	1.2410	1.2923
COPD	1.6218	1.5715	1.6737
Coagulopathy	1.3495	1.2999	1.4009
Cardiorespiratory failure	1.2116	1.1780	1.2462
Other infectious disease and pneumonias	1.4846	1.4456	1.5246
Psychiatric comorbidity	1.4862	1.4318	1.5427
Septicemia	1.5659	1.5064	1.6278
Percent nurse-to-patients	1.0112	0.9793	1.0442
Percent PCT-to-patients	1.0005	0.9669	1.0353
Survival			
Variable	HR	Lower CL	Upper CL
Age	1.1688ª	1.1350	1.2037
Female	0.8277	0.7256	0.9441
COPD	1.2638	1.0261	1.5567
Coagulopathy	1.2089	1.1074	1.3196
Cardiorespiratory failure	1.4067	1.1858	1.6688
Other infectious disease and pneumonias	1.0983	0.9255	1.3034
Psychiatric comorbidity	1.1521	1.0037	1.3225
Septicemia	1.2834	1.2043	1.3678
Percent nurse-to-patients	0.8201	0.7778	0.8647
Percent PCT-to-patients	0.7038	0.6755	0.7332
Hospitalization risk score	1.0739 ^c	1.0551	1.0932

Note. Given are estimates of odds ratios (ORs) of hospitalization and hazard ratios (HRs) of death along with corresponding 95% lower and upper confidence limits (CLs).

Abbreviations: COPD, chronic obstructive pulmonary disease; PCT, patient care technician.

^a For 12-month effect.

^b For 5-year effect.

 $^{\rm c}$ For ~0.1 effect size (20% of standard deviation of hospitalization risk score—see the main text).



FIGURE 3 (a) Distribution of $\hat{m}_{ij}(t)$, the hospitalization risk score, across all 30,030 patients with 292,672 total observations and corresponding estimated probability of hospitalization as function of $\hat{m}_{ij}(t)$ (right vertical axis). (b) Estimate of the baseline hazard function for $\lambda = 10$ (shaded: bootstrap 95% confidence intervals). (c) Estimate of the baseline hazard function for λ from 0.5 to 8 and (d) for λ from 10 to 30 resulting in similar smoothness of $\hat{h}_0(t)$

The effects of patient- and facility-level risk factors on longitudinal hospitalizations and survival outcomes are summarized in Table 4. With respect to hospitalization, patients with COPD and septicemia in the year prior to starting dialysis had the odds ratio (OR) of hospitalization (COPD: OR 1.62, 95% CI 1.57–1.67; septicemia: OR 1.57, 95% CI 1.51–1.63). Presence of other comorbidities (coagulopathy, psychiatric conditions, cardiorespiratory failure/shock, other infectious diseases, and pneumonias) were associated with about 21% to 47% higher odds of hospitalization. Consistent with previous reports (e.g., United States Renal Data System, 2018), female sex was associated with higher odds of hospitalization. One year on dialysis was associated with 19% higher odds of hospitalization. In addition, Figure 3a displays the distribution of the estimated hospitalization risk scores, $\hat{m}_{ij}(t)$, for all patients and time points *t*, (mean –0.47, SD 0.49; Q1–Q3: –0.84 to –0.20) and the corresponding increasing hospitalization probability $g(\hat{m}_{ij})$ as a function of $\hat{m}_{ij}(\cdot)$.

The survival component (Table 4B) of the MJM similarly found significant increases in hazard ratio (HR) of death associated with nearly all comorbidities ranging from 41% to 15% higher hazard of death for cardiorespiratory failure (HR 1.41, 95% CI 1.19–1.67) and psychiatric comorbidity (HR 1.15, 95% CI > 1.00–1.32). Further, older age was associated with an incremental increase in HR, and female sex was associated with lower mortality (HR 0.82, 95% CI 0.73–0.94).

The survival component of the joint model (2) also included $\mathcal{M}_{ij}(t)$, the longitudinal hospitalization risk score history up to time *t*, in estimating survival. The longitudinal hospitalization history had a "moderate" effect (relative to severe cormorbidities) on the subsequent mortality risk as revealed by the effect size estimate associated with an increase in the longitudinal hospitalization risk score history, $\mathcal{M}_{ij}(t)$. For example, Figure 3a displays the distribution of estimated hospitalization risk scores ($\hat{m}_{ij}(t)$'s) for all 34,030 patients with a range of -1.50 to 2.24 and SD of 0.49. Thus, an increase of 20% of the SD (0.2 × 0.49) in $\hat{m}_{ij}(t)$ was associated with a 7.4% (95% CI 1.06–1.09) increase in the hazard of death (Table 4B).

Facility-level staffing covariates, specifically the percentages of nurse-to-patient and PCT-to-patient, were negatively associated with mortality; that is, higher percentage of staff (nurse or PCT) to patient was associated with reduced hazard of patient death (HRs 0.82 and 0.70, respectively); see Table 4B. The effect of facility staffing variables on hospitalization was not significant in the MJM (Table 4A).

Finally, we note that the baseline hazard, $h_0(t)$, was estimated similarly as was done in the simulation studies and displayed in Figure 3b. More specifically, $\lambda = 10$ was chosen, which provided a smoothed estimate of the baseline hazard and higher values resulted in similar estimates as illustrated in Figure 3b,c.

5 | DISCUSSION

We have considered a joint model of longitudinal hospitalizations and survival with respect to the U.S. dialysis population for three-level hierarchical data where longitudinal hospitalizations over time are nested within subjects and subjects are further nested within dialysis facilities through multilevel REs. Application of the novel MJM revealed the relative contribution of modifiable risk factors for hospitalization and mortality. In particular, the MJM quantified the burden of hospitalizations over time on subsequent mortality risk; thus, the results suggest that concerted strategies to reduce patient hospitalization, including aggressive management of chronic comorbid conditions as well as prevention of hospitalization risk (e.g., infection-related hospitalizations, which are common for dialysis patients; e.g., see Dalrymple et al., 2011; Mohammed, Senturk, Dalrymple, & Nguyen, 2012) may contribute to reduction of overall patient mortality risk. Interestingly, the effect of facility staffing on mortality, accounting for patients' hospitalization history, was relatively large. Thus, evidence-based strategies for "optimal" staffing (e.g., the minimal number of nurses and PCTs relative to patient volume) is an area worthy of exploration in an overall effort of dialysis facilities to reduce patient mortality.

In this work, a generalization of the standard joint modeling framework was required to accommodate the multilevel USRDS data structure of patients on dialysis. Specifically, our proposed MJM utilized multilevel REs (at both the subject- and dialysis facility-level) and multilevel risk factors in modeling the dependency between hospitalization and survival. Several technical advancements were achieved, including the (1) development of feasible estimation that addressed the challenge of high-dimensional integrations in the EM algorithm and (2) derivation of asymptotic SEs formulas for the model parameters that allowed for a systematic assessment of their biases, resulting in investigation of alternative inference based on bootstrap SEs. Finally, R codes and documentation for fitting the proposed MJM are made publicly available.

ACKNOWLEDGEMENTS

This study was supported by a grant from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK092232–D. S., D. V. N., K. K., and C. M. R.). The interpretation and reporting of the data presented here are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S government. We are grateful to the AE and two referees, whose constructive comments strengthened the manuscript.

DATA AVAILABILITY STATEMENT

The release of the data used in this paper is governed by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) through the USRDS Coordinating Center. The data can be requested from the USRDS through a data use agreement.

REFERENCES

- Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11), 1713–1723.
- Chen, Y., Rhee, C. M., Senturk, D., Kurum, E., Campos, L. F., Li, Y., Kalantar-Zadeh, K., & Nguyen, D. V. (2019). Association of U.S. dialysis facility staffing with profiling of hospital-wide 30-day unplanned readmission. *Kidney Diseases*, *5*(3), 153–162.
- Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 34(2), 187-220.
- Dalrymple, L. S., Mohammed, S. M., Mu, Y., Johansen, K. L., Chertow, G. M., Grimes, B., Kaysen, G. A., & Nguyen, D. V. (2011). The risk of cardiovascular-related events following infection-related hospitalizations in older patients on dialysis. *Clinical Journal of the American Society of Nephrology*, 6(7), 1708–1713.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society:* Series B, 39(1), 1–22.
- Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. Statistical Science, 11, 89-102.
- Eilers, P. H., & Marx, B. D. (2010). Splines, knots, and penalties. Wiley Interdisciplinary Reviews: Computational Statistics, 2(6), 637-653.
- Eilers, P. H., Marx, B. D., & Durbn, M. (2015). Twenty years of P-splines. SORT: Statistics and Operations Research Transactions, 39(2), 149–186.
- Estes, J. P., Nguyen, D. V., Dalrymple, L. S., Mu, Y., & Senturk, D. (2016). Time-varying effect modeling with longitudinal data truncated by death: Conditional models, interpretations and inference. *Statistics in Medicine*, *35*(11), 1834–1847.
- Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. Biostatistics, 1(4), 465-480.
- Hsieh, F., Tseng, Y., & Wang, J. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. Biometrics, 62(4), 1037-1043.
- Kass, R. E., & Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). Journal of the American Statistical Association, 84(407), 717–726.
- Li, Y., Nguyen, D. V., Chen, Y., Rhee, C. M., Kalantar-Zedeh, K., & Senturk, D. (2018). Modeling time-varying effects of multilevel risk factors of hospitalizations in patients on dialysis. Statistics in Medicine, 37(30), 4707–4720.

- Liu, L., Ma, J. Z., & O'Quigley, J. (2008). Joint analysis of multi-level repeated measures data and survival: An application to the end stage renal disease (ESRD) data. *Statistics in Medicine*, 27(27), 5679–5691.
- Mohammed, S. M., Senturk, D., Dalrymple, D. S., & Nguyen, D. V. (2012). Measurement error case series models with application to infection-cardiovascular risk in older patients on dialysis. *Journal of the American Statistical Association*, 107(500), 1310–1323.
- Njagi, E. N., Rizopoulos, D., Molenberghs, G., Dendale, P., & Willekens, K. (2013). A joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Statistical Modelling*, 13(3), 179–198.
- Rizopoulos, D., Molenberghs, G., & Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6), 1261–1276.
- Rizopoulos, D., Taylor, J. M., Van Rosmalen, J., Steyerberg, E. W., & Takkenberg, J. J. (2015). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1), 149–164.
- Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal* of the Royal Statistical Society: Series B (Statistical Methodology), 71(3), 637–654.
- Rizopoulos, D., Verbeke, G., & Molenberghs, G. (2008). Shared parameter models under random effects misspecification. Biometrika, 95(1), 63-74.
- Song, X., Davidian, M., & Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58(4), 742–753.
- Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. Journal of the American Statistical Association, 84(407), 710–716.
- Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2), 447–458.
- Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica, 14, 809-834.
- United States Renal Data System (2014). Researcher's guide to the USRDS database. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Disease. https://www.usrds.org/media/2235/usrds_researchers_guide-14.pdf
- United States Renal Data System (2018). USRDS annual data report: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, https://www.usrds.org/annual-data-report/previous-adrs/
- Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. Biometrics, 53, 330-339.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kurum E, Nguyen DV, Li Y, Rhee CM, Kalantar-Zadeh K, Senturk D. Multilevel joint modeling of hospitalization and survival in patients on dialysis. *Stat.* 2021;10:e356. https://doi.org/10.1002/sta4.356