Published in final edited form as:

Stat Med. 2021 September 20; 40(21): 4582–4596. doi:10.1002/sim.9081.

Joint modeling of longitudinal data with informative cluster size adjusted for zero-inflation and a dependent terminal event

**Biya Shen¹, Chixiang Chen², Danping Liu³, Somnath Datta⁴, Nasrollah Ghahramani⁵,
Vernon M. Chinchilli¹, Ming Wang¹**

¹Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, Pennsylvania

²Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania

³Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland

⁴Department of Biostatistics, University of Florida, Rockville, Florida

⁵Division of Nephrology, Penn State College of Medicine, Hershey, Pennsylvania

Abstract

Repeated measures are often collected in longitudinal follow-up from clinical trials and observational studies. In many situations, these measures are adherent to some specific event and are only available when it occurs; an example is serum creatinine from laboratory tests for hospitalized acute kidney injuries. The frequency of event recurrences is potentially correlated with overall health condition and hence may influence the distribution of the outcome measure of interest, leading to informative cluster size. In particular, there may be a large portion of subjects without any events, thus no longitudinal measures are available, which may be due to insusceptibility to such events or censoring before any events, and this zero-inflation nature of the data needs to be taken into account. On the other hand, there often exists a terminal event that may be correlated with the recurrent events. Previous work in this area suffered from the limitation that not all these issues were handled simultaneously. To address this deficiency, we propose a novel joint modeling approach for longitudinal data adjusting for zero-inflated and informative cluster size as well as a terminal event. A three-stage semiparametric likelihood-based approach is applied for parameter estimation and inference. Extensive simulations are conducted to evaluate the performance of our proposal. Finally, we utilize the Assessment, Serial Evaluation, and Subsequent Sequelae of Acute Kidney Injury (ASSESS-AKI) study for illustration.

Correspondence: Ming Wang, Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Pennsylvania State College of Medicine, 90 Hope Drive, Hershey, PA 17033, mwang@phs.psu.edu.

AUTHOR CONTRIBUTIONS

All authors have made important contributions and have approved this work.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

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

Keywords

competing risk; informative cluster size; joint modeling; longitudinal data analysis; zero-inflation

1 | INTRODUCTION

Prospective cohort studies are commonly encountered in medical research and clinical trials, which often collect longitudinal data on participant outcome and covariates. The observation time of the study participants can be prespecified in some studies, but in many situations, such as observational studies or real-time surveillance, the time at which data are collected cannot be predetermined, but depends on the occurrence of some specific event. The measurements are only available when such an event is observed, for instance, duration of hospitalization or urine or serum biomarker measures from request laboratory tests. Therefore, the frequency of event occurrences are likely to be correlated with the measurements related to disease condition. As a consequence, the cluster size defined by the number of repeated observations within subject/cluster could be informative. Examples include widely investigated periodontal disease.¹ In our motivation study on acute kidney injury (AKI), experiencing more often AKI recurrences could be likely associated with a more severe kidney function of the individual, and hence may influence the distribution of the outcome measure of interest, leading to an informative cluster size issue. More detailed information about AKI recurrences and the associated cluster sizes as well as the trajectories across time since baseline from several random subjects are provided in the Supplementary Material. Based on preliminary check, the observed recurrent times are irregular and tend to be random, thus, the number of events occurring during the follow-up is our focus here, characterizing by a discrete Poisson distribution that allows to depend on potential covariates. In addition, there could be a large portion of participants without any recurrent events, which may be due to insusceptibility to such events or censoring before any events occur. This manifests the zero-inflation nature of cluster size, which needs to be taken into account for inference. On the other hand, there often exist competing risks (eg, a terminal event of death) that may be correlated with recurrent events and longitudinal measures during the follow-up.

Most previous work focused primarily on modeling observed longitudinally measured responses and covariates given that a terminal event has not yet occurred.^{2–4} To take correlations within study participants into account, a widely adopted approach is to incorporate random effects using likelihood-based regression models (eg, mixed-effects models). However, in many cases, the terminal event is attained and cannot be ignored, owing to its potential correlations with repeated measurements. In such situations, longitudinal data are nonignorable missing due to this competing event. In literature, three main approaches have been adopted for such scenarios, depending on the particular research interest: the joint modeling approach,^{5–8} the marginal estimating equation approach,^{9,10} and the conditional modeling approach.¹¹ The first of these relies on random effects shared by both processes of longitudinal data and the terminal event to indirectly quantify their dependence; the second adopts the inverse probability weighting technique to adjust the censoring effect due to the terminal event for observations of the longitudinal process; the

third treats the terminal event as a covariate in order to investigate the explicit effect of the terminal event time on the longitudinal process.^{11,12} In our work, we focus on the last approach because of our primary interest in the trajectory of longitudinal measures and our secondary goal of exploring how the longitudinal process is affected near the time of the terminal event. For instance, previous studies showed that a rise in serum creatinine level is associated with a higher risk of in-hospital mortality.¹³ Another example is the increase in length and variability of menstrual cycles as women approach menopause.¹⁴

On the other hand, it is recognized in real applications that the cluster size (ie, the number of within-subject observations) may influence the distribution of disease outcome, leading to an informative cluster size; examples include periodontal disease studies,¹ healthcare costs, and duration of hospitalization.¹⁵ Both marginal modeling methods (eg, weighted generalized estimating equations)^{1,16,17} and joint modeling methods (eg, shared random effects models)^{18–20} have been proposed in such contexts. The joint modeling method is widely used to incorporate a common set of random effects shared by both models for the outcomes and cluster size, and this approach is more flexible in that it can be adapted for complicated contexts such as ours. In our situation, there is a substantial portion of participants with zero recurrent events of interest during the follow-up, resulting in no “marker” measures being observed. Thus, an excess of zeros is recognized for the cluster size, leading to zero-inflation, a commonly occurring issue in real practice. Several zero-inflated models have been proposed, such as zero-inflated Poisson or negative binomial models,^{21,22} and these have been extended to longitudinal data^{23,24} or recurrent events analysis.^{25,26} However, so far, there have been few studies focusing on such longitudinal data with zero-inflated cluster size obtained from a recurrent “marked” process. The existing approaches utilizing only available measures for longitudinal analysis may lead to inefficient inference, because those participants with zero recurrences to some extent provide information related to disease outcome or death. Some of these participants may be insusceptible (ie, structural zeros) and are totally risk-free with regard to the event of interest; some may be susceptible (ie, random zeros), and the event will eventually occur if the follow-up time is long enough provided that no terminal event occurs. Thus, such informative cluster size and nonignorable zero-inflation need to be taken into account for more efficient inference to be possible.

This paper was motivated by the Assessment, Serial Evaluation, and Subsequent Sequelae of Acute Kidney Injury (ASSESS-AKI) study, a prospective study to characterize the natural history of AKI, a common complication in hospitalized patients, with a secondary interest in studying how kidney function deteriorates shortly before death.^{27,28} Note that the event of AKI can be captured and quantified by serum creatinine, and can be recurrently experienced. Previous studies have shown that the incidence of AKI increases after a first hospitalized AKI and is associated with a high risk of worsening kidney function and short-term death.^{29–31} Moreover, a high percentage (67.3%) of participants who had no AKI events do not contribute any longitudinal measurements, and are likely to be healthier (lower risk of death). Thus, some participants are likely to have no AKI events if they are free of AKI and chronic kidney disease (CKD) at baseline. Therefore, the number of trackable kidney function measures (eg, serum creatinine) associated with hospitalized AKI recurrence varies across participants, which is correlated in some way with overall kidney function, resulting

in informative cluster size. Until now, it has not been clear how to better characterize the profile of serum creatinine from hospitalizations and mitigate future risk of terminal events (ie, death/end-stage renal disease). The ASSESS-AKI study provides a new platform with real-time information on hospital-acquired AKI and associated markers in a longitudinal follow-up, which motivates the development of statistical methods.

In our framework, longitudinal measures (ie, serum creatinine) are adherent to recurrent events (ie, AKI) with potentially informative and zero-inflated cluster size due to overall disease status or terminal events (ie, death/end-stage renal disease). However, existing methods found in the literature cannot be directly applied and have the limitation that they are unable to handle all of the above issues simultaneously. Ignorance of any one of these issues could make inference less efficient. To address this concern, we propose an innovative joint modeling approach for longitudinal data captured from recurrent events with informative and zero-inflated cluster size in the presence of the terminal event. In particular, we consider a mixed-effects model for repeated measurement analysis by including the time to terminal events as a fixed-effect covariate. Also, a zero-inflated Poisson model for cluster size is incorporated into the joint model framework to account for its correlation with both the longitudinal profile. We propose a three-stage semiparametric likelihood-based approach for parameter estimation and inference, and compare its results with those of alternative approaches (eg, a complete case analysis in which participants with censored terminal event times are dropped) in simulation studies and an analysis of the ASSESS-AKI study.

The rest of this article is organized as follows. In Section 2, we propose a joint modeling approach and a three-step estimation procedure. In Section 3, we present extensive simulation studies numerically investigating the performance of the proposed method. We apply the proposed method to the ASSESS-AKI study in Section 4. Finally, we conclude with a discussion and suggestions for future work in Section 5. Of note, the additional sensitivity analysis for evaluating the robustness of our method and detailed real data description are presented in the Supplementary Material.

2 | METHODOLOGY

For each subject i ($i = 1, 2, \dots, n$), let C_i denote the noninformative censoring time and S_i denote the time to a terminal event; thus, the observed follow-up time $V_i = \min\{S_i, C_i\}$, with the terminal indicator $\delta_i = I\{S_i \leq C_i\}$. The number of repeated measures (ie, cluster size) is denoted by N_i with realization n_i (ie, the observed number of events during follow-up), where $n_i \geq 0$. Thus, the repeated marker measurements (ie, serum creatinine) from the i th subject are obtained from recurrent events during the follow-up, noted by $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{i, n_i})^T$, where Y_{ij} is measured at the time of each recurrent event (ie, hospitalized AKI), t_{ij} that is determined by real-time monitoring or electoral medical records, and also $t_{in_i} \leq V_i$. The individual covariates are denoted by X_i , a $p \times 1$ vector, which can be time-independent or time-varying; in our data application, however, we only had baseline predictors. We assume that the censoring is noninformative in the sense that $C_i \perp (S_i, Y_i) | (X_i, N_i)$. In addition, p_i denotes the probability that a subject is insusceptible to the recurrent event, ie, a structure zero in the cluster size.

2.1 | Joint modeling approach

To incorporate informative cluster size with zero-inflation, we adopt the shared random effect approach to the joint model Y_i and N_i , taking the time of terminal event into account as a covariate for adjustment. Specifically, if $N_i > 0$, $Y_i = \{Y_{ij}\}_{j=1}^{N_i}$, otherwise Y_i is not observed. Thus, given observed S_i , the repeated measure Y_{ij} can be modeled by the following mixed-effects model:¹¹

$$Y_{ij} = \mathbf{X}_i^T \boldsymbol{\beta} + g(S_i - t_{ij}, \kappa) + \mathbf{Z}_i^T \boldsymbol{\gamma}_i + \epsilon_{ij}, \quad (1)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients associated with the covariates $\mathbf{X}_i$ (the intercept included), $\boldsymbol{\gamma}_i \sim N(0, \Sigma_\gamma)$ include random effects, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is the error term, with $\boldsymbol{\gamma}_i \perp \epsilon_{ij}$. Here, $g(S_i - t_{ij})$ is used to quantify the relationship between the longitudinally measured response variable Y_{ij} and the terminal event S_i . Specifically, when the event time t_{ij} is far from the terminal event time S_i , the model (1) reduces to a simpler relationship of interest between the longitudinally measured outcome Y_{ij} and covariates $\mathbf{X}_i$, but when t_{ij} is close to S_i , a dramatic change is expected to be detected. There are multiple choices for $g(S_i - t_{ij})$; further discussion can be found in Chang and Wang (2010)³² and also Kong et al (2018),¹¹ but here we take $g(S_i - t_{ij})$ to be a normal kernel, one of the commonly used kernels, where $g(S_i - t_{ij}, \kappa) = \kappa_1 \exp[-\kappa_2 \times (S_i - t_{ij} - \kappa_3)^2]$, $\kappa = (\kappa_1, \kappa_2, \kappa_3)^T$. However, we admit that the other competing choices such as an exponential kernel could also be considered. Note that this selection could depend on specific clinical application, and some smoothing techniques can also be helpful to determine an appropriate one.

For an informative cluster size (ie, a nonnegative number of observations for each subject), zero-inflated Poisson (ZIP) regression is adopted for model fitting. Let J_i be an unobserved latent variable, with the value of 1 indicating that N_i is from the structure zeros, which happens with probability p_i , and it takes the value of 0 when N_i is from a Poisson distribution with probability $1 - p_i$. In another word, we have $N_i = 0$ if $J_i = 1$ and $N_i \sim \text{Poisson}(\lambda_i)$ if $J_i = 0$, where λ_i is the parameter for the Poisson distribution. Here, we assume that the probability of structure zeros $p_i = \exp(\mathbf{W}_i^T \boldsymbol{\theta}) / [1 + \exp(\mathbf{W}_i^T \boldsymbol{\theta})]$, where the covariates $\mathbf{W}_i$ can be a subset of, or can overlap with the covariates $\mathbf{X}_i$ from the main model (1). Furthermore, we define the parameter for the Poisson distribution as

$$\lambda_i = \exp(\mathbf{R}_i^T \boldsymbol{\vartheta}_1 + \vartheta_2 I(V_i) + \boldsymbol{\gamma}_i^T \boldsymbol{\eta}), \quad (2)$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1^T, \vartheta_2)^T$ are the parameters for the covariates $\mathbf{R}_i$ that can be different from or can overlap with $\mathbf{X}_i$ and $\mathbf{W}_i$, and $\boldsymbol{\eta}$ include the parameters for $\boldsymbol{\gamma}_i$ which are random effect terms in (1). Here, $I(V_i)$ is a prespecified function of the follow-up time V_i reflecting the fact that, on average, the cluster size of the longitudinal data tends to be larger for subjects with a longer follow-up V_i or more severe disease situation reflected by higher marker measures. There are several options for the function I , depending on the given disease-specific mechanism; for instance, it could be defined as a cumulative hazard function at the time V_i in the case that a high risk of occurring recurrent events is associated with death.

Note that here we do not include the random effects γ_i in modeling p_i ; this is because of identifiability issues, since no information about Y_i can be observed for structural zeros.

We can now write a detailed expression for the informative cluster size N_i :

$$P(N_i = n_i | \gamma_i, V_i, \mathbf{W}_i, \mathbf{R}_i) = \begin{cases} p_i + (1 - p_i)\exp(-\lambda_i), & n_i = 0, \\ (1 - p_i)\frac{\exp(-\lambda_i)(\lambda_i)^{n_i}}{n_i!}, & n_i > 0. \end{cases} \quad (3)$$

For the time to the terminal event, the commonly used semiparametric model, a Cox proportional hazards (PH) regression, can be fitted. In particular, the hazard function of S_i given the covariates $\mathbf{K}_i$ is $h(s | \mathbf{K}_i) = h_0(s)\exp(\mathbf{K}_i^T \boldsymbol{\alpha})$, where the covariates $\mathbf{K}_i$ can be a subset of, or overlap with $\mathbf{X}_i$. $\boldsymbol{\alpha}$ is a vector of regression parameters with the true values $\boldsymbol{\alpha}_0$, and $h_0(s)$ is the baseline hazard function. Also, the cumulative hazard can be written as $H(s | \mathbf{K}_i) = \int_0^s h(v | \mathbf{K}_i)dv$, with the baseline cumulative hazard as $H_0(s)$. Let $f(S_i | \mathbf{X}_i)$ be the conditional density on survival S_i given $\mathbf{X}_i$ and the conditional cumulative distribution function is denoted by $F(s | \mathbf{K})$ which is given by $1 - \exp[-H_0(s)\exp(\mathbf{K}_i^T \boldsymbol{\alpha})]$.

For simplicity and to avoid notation confusion, we consider an random intercept only in (1), that is, $\gamma_i \sim N(0, \sigma_\gamma^2)$, and also assume $\mathbf{W}_i$, $\mathbf{R}_i$ and $\mathbf{K}_i$ as subsets of, or the same as the vector $\mathbf{X}_i$, thus $\mathbf{X}_i$ is used throughout. However, these conditions are not necessary and can be easily and straightforwardly relaxed. Also, it is noted that the time to the terminal event could be unobserved, thus $g(S_i - t_{ij}; \boldsymbol{\kappa})$ is also unknown. For such censored subjects, the unknown survival time is expected to be imputed or integrated out based on distribution functions. Under the noninformative assumption for the random censoring time C_i that is conditionally independent with the longitudinal data and also the terminal event time S_i (ie, $C_i \perp S_i | (\mathbf{X}_i, \mathbf{Y}_i, N_i)$), the conditional distribution of S_i can be modeled using the subjects with the observed terminal event. This procedure allows extrapolation beyond the censoring time C_i for censored subjects. Note that the magnitude of efficiency gain depends on how well the right tail of the cumulative distribution $F(S_i | \mathbf{X}_i)$ beyond C_i can be estimated.

We omit the subscript i in the subsequent presentation of the likelihood. Here, the nuisance parameters include $\boldsymbol{\alpha}$ and $\boldsymbol{\phi} = (\sigma_e, \sigma_\gamma)^T$, and the parameter vector of interest is denoted by $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \boldsymbol{\vartheta}^T, \boldsymbol{\kappa}^T, \eta)^T$, which is our primary focus for estimation. The likelihood function can be factorized by the product of the density functions of longitudinal data and the terminal event, and then reorganized using the Bayes' rule along with similar lines as Kong et al (2018).¹¹ Given the data $(V, \mathbf{Y}, N, \mathbf{X})$, the likelihood function is written by

$$L(\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\psi}; t, V, \Delta, \mathbf{Y}, N, \mathbf{X}) \propto [f(\mathbf{Y}, N | S, \mathbf{X})f(S | \mathbf{X})]^\Delta \left[\int_C^\infty f(\mathbf{Y}, N | s, \mathbf{X})dF(s | \mathbf{X}) \right]^{1-\Delta},$$

where $F(s | \mathbf{X}) = 1 - \exp[-H_0(s)\exp(\mathbf{X}^T \boldsymbol{\alpha})]$, and the joint density of $\mathbf{Y}$ and N for the death cases ($\Delta = 1$) could be expressed as:

$$\begin{aligned}
 f(Y, N | S, X) &= \int f(Y, N | \gamma, S, X) f(\gamma) d\gamma \\
 &= \int f(Y, N | \gamma, S, X) P(N | \gamma, S, X) [I(N > 0) + P(N | \gamma, S, X) I(N = 0)] f(\gamma) d\gamma,
 \end{aligned}$$

where I is an indicator function. Note that for the subject with $N=0$, there will be no longitudinal observations at all. On the other hand, for the censored cases ($\delta = 0$), we need to calculate the integral $\int_C^\infty f(Y, N | s, X) dF(s | X)$ in the likelihood function. The strategy by Lu et al (2010)³³ and Kong et al (2018)¹¹ can be adopted for this purpose, and more details will be provided next. Finally, the log-likelihood function is derived as

$$\begin{aligned}
 l(\alpha, \phi, \psi; V, \Delta, Y, N, X) &\propto \Delta \log \int f(Y, N | \gamma, S, X) f(\gamma) d\gamma + \Delta \log f(S | X) \\
 &\quad + (1 - \Delta) \log \int_C^\infty f(Y, N | \gamma, s, X) f(\gamma) dF(s | X) d\gamma. \quad (4)
 \end{aligned}$$

Note that the second part of the likelihood is the contribution from the censored participants, for whom the terminal event S was not observed. Therefore, the integration beyond the observed censoring time C is needed, and the amount of efficiency gain depends on how well we can estimate this right tail of the conditional distribution $F(s | X)$. Here, we have proposed a novel parameter estimation procedure that is computationally convenient and suitable for software implementation, which will be described next.

2.2 | Three-stage procedure for parameter estimation

We propose a three-stage procedure for parameter estimation with computational convenience. By utilizing all available subjects and incorporating extra information that is relevant to the outcome of interest, the efficiency of parameter estimates are anticipated, which will be exclusively evaluated in the following simulation studies. The nuisance parameters α and ϕ are estimated in stages 1 and 2, and the parameter of interest ψ is then estimated by maximizing (4) in stage 3 with the nuisance parameters plugged in. The procedure can be summarized as follows.

Stage 1: First, we estimate the nuisance parameter α in the Cox PH regression by maximizing the partial likelihood, and we estimate the cumulative baseline hazard function $H_0(\cdot)$ by the Breslow estimator.³⁴ The parameter estimates are denoted by $\hat{\alpha}$ and $\hat{H}_0(\cdot)$. Accordingly, the cumulative density function $F(s | X)$ can be estimated by $\hat{F}(s | X) = 1 - \exp[-\hat{H}_0(s) \exp(X^T \hat{\alpha})]$, which is asymptotically equivalent to the product integral expression.¹¹

Stage 2: In this stage, we can estimate the nuisance parameters $\phi = (\sigma_e, \sigma_\gamma)^T$ by utilizing complete death cases only (the data with $\delta = 1$). After plugging in the estimates $\hat{\alpha}$ and $\hat{H}_0(\cdot)$ from stage 1, the estimates $\hat{\phi}$ and $\hat{\psi}$ can be obtained by maximizing the log likelihood $\log \int f_{\psi, \phi, \hat{\alpha}}(Y, N | \gamma, S, X) f(\gamma) d\gamma$ based on the complete death cases only. The random effect

could be integrated out by Gauss–Hermite quadrature approximation³⁵ or other numerical approximation.³⁶

Stage 3: This stage has a similar procedure to stage 2, but now we use all data to finally estimate the primary parameter of interest $\boldsymbol{\psi}$. In particular, we can directly use the estimation result $\hat{\boldsymbol{\alpha}}$ from stage 1 and $\hat{\boldsymbol{\phi}}$ from stage 2, and set the estimates, $\hat{\boldsymbol{\psi}}$, from stage 2 as the initial values for the optimization algorithm to maximize the log likelihood here. Note that, we do not need to update the nuisance parameters in this stage. Given $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\alpha}}$, and $\hat{H}_0$, we have the following log pseudo-likelihood function:

$$l(\boldsymbol{\psi}; V, \Delta, Y, N, X) \propto \Delta \log \int f_{\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}}(Y, N | \gamma, S, X) f(\gamma) d\gamma \\ + (1 - \Delta) \log \iint_C f_{\boldsymbol{\psi}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}}(Y, N | \gamma, s, X) d\hat{F}(s | X) f(\gamma) d\gamma. \quad (5)$$

Note that the term $\log f(S|X)$ in (4) is dropped because it does not involve any information on $\boldsymbol{\psi}$. The estimates $\hat{\boldsymbol{\psi}}$ can be obtained by maximizing (5). $\hat{F}$ is obtained from an estimated step function, the integral can be written as a weighted average of the conditional density $f(Y, N | s, X)$ at all the jump points in $\mathcal{M}$ that is the set including all observed death times and also the last observation time if it is a censoring time for $\hat{F}(s | X)$. The weights will be the jump sizes of $\hat{F}(s | X)$ at those time points.³³ The random effects can be integrated using a prespecified numerical integration method following a similar procedure as before.

With regard to the main model alone under the context without informative and zero-inflated cluster size, Kong et al (2018)¹¹ have already proved the properties of consistency and asymptotic normality, which is based on empirical process theory. However, they still recommended the bootstrap-based variance estimator due to the complicated expression of their derived asymptotic variance. In our case, the joint modeling has even more complex expression because of the involvement of additional features of cluster size as well as plug-in estimators. Therefore, the bootstrap approach is adopted to obtain variance estimates. Next, we will perform extensive simulation studies to evaluate the performance of our inference procedure, and also the asymptotic property of parameter estimates are empirically investigated.

3 | SIMULATION STUDIES

We conduct extensive simulations to investigate the finite-sample performance of our proposed joint modeling approach. For each subject i ($i=1, \dots, n$), the terminal event time S_i and the censoring time C_i is predefined, and the observed follow-up time V_i is generated as $\min(S_i, C_i)$. We consider two settings with different censoring rates: (1) the event time S_i follows an exponential distribution with the rate parameter $\exp(a_0 + a_1 K_{i1} + a_2 K_{i2})$, where $K_{i1} \sim \text{Bernoulli}(0.5)$, $K_{i2} \sim \mathcal{N}(0, 1)$, and the parameters $a_0 = 0.5$, $a_1 = -0.5$, and $a_2 = 0.5$. The censoring time C_i follows an exponential distribution with the rate parameter of 0.5, leading to a censoring rate of about 50%; (2) the event time S_i remains the same, but the censoring

time C_j follows an exponential distribution with the rate parameter of 1.5, thus the censoring rate is about 75%.

The cluster size N_j is assumed to follow the ZIP distribution, with the probability of structure zeros $p_j = \exp(\theta_0 + \theta_1 W_j) / [1 + \exp(\theta_0 + \theta_1 W_j)]$ and the parameter for the Poisson distribution $\lambda_j = \exp[\vartheta_0 + \vartheta_1 W_j + \vartheta_2 \mathcal{K}(V_j) + \eta \gamma_j]$. The covariate W_j is generated from a normal distribution $\mathcal{N}(1.3, 0.5)$, and $\mathcal{K}(V_j)$ is defined as $\exp(a_0 + a_1 K_{j1} + a_2 K_{j2}) V_j$. Note that $\mathcal{K}(V_j)$ is defined based on the hazard function $\exp(a_0 + a_1 K_{j1} + a_2 K_{j2})$, and we want to utilize $\mathcal{K}(V_j)$ to describe the survival risk at observed time V_j . For the Poisson mean λ_j , we set $\vartheta = (\vartheta_0, \vartheta_1, \vartheta_2)^T = (0.5, -0.5, 1)^T$ and $\eta = 0.5$. In particular, we consider different setups for the parameters in p to attain different zero rates for the observed cluster size N_j and evaluate their effects on the efficiency of parameter estimation. Here, two scenarios are presented: (1) $\theta_0 = 1.5$, $\theta_1 = -1$, for which the observed zero rate is around 70%, and (2) $\theta_0 = -1$, $\theta_1 = -0.5$, for which the zero rate is around 40%.

After all the cluster sizes have been determined as above, we assign these events within the observed follow-up window with equally spaced time intervals. Noting that we also explore the scenarios with noninformative and unequally spaced time intervals following the strategy as Kong et al (2018),¹¹ and similar results are concluded. Thereafter, the associated measures are generated from the following nonlinear mixed effects model:

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \kappa_1 \exp\left[-\kappa_2 * (S_i - t_{ij} - \kappa_3)^2\right] + \gamma_i + \epsilon_{ij}, \quad (6)$$

where $(X_{i1}, X_{i2})^T \sim \mathcal{N}(0, \Sigma_X)$ with $\Sigma_X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\beta = (\beta_0, \beta_1, \beta_2)^T = (2, 1, 1)^T$, and $\kappa = (\kappa_1, \kappa_2, \kappa_3)^T = (1, 2, 1)^T$. Also, for subject i , S_i is the terminal time and t_{ij} is the j th recurrent event time. Furthermore, the random intercept and the residual term are assumed to be independent, which are generated from $\gamma_i \sim \mathcal{N}(0, 1)$ and $\epsilon_{ij} \sim \mathcal{N}(0, 0.5)$. Noting that for the sake of our clinical interest, the baseline covariate effects (ie, baseline AKI status CKD status are our main focus, thus, the model with a random intercept only is reasonable.

For each scenario, 500 Monte Carlo replicates each with sample size $n = 300, 600, 900$ are generated. We compare the results from three models with death cases only (death cases), all cases including all available and observed data (all cases), and full data assuming all deaths known and observed (full data). Note that the first two stages of our proposed estimation procedure are used to handle the model with known death cases only, and the three-stage procedure is for the model with both death cases and censoring cases. The summary statistics of the parameter estimates including Monte Carlo bias (Bias), Monte Carlo standard derivation (SD), and mean square error (MSE) are reported for all models. In particular for the model using all cases, we also report 95% coverage probability (CP) obtained from 250 bootstrap samples for further evaluation.

The results summary for all primary parameters of interest are shown in Tables 1 to 3. We can see that for all three models, almost all bias, SD, and MSE decrease as the sample size increases. In particular, for the parameters quantifying the effects of risk factors, β , in the longitudinal model, the biases from the all cases analysis are comparable with the other two

models even for a relatively small sample size with a high proportion of zeros, but it is noted that SDs are relatively smaller than the death cases analysis, indicating the benefit of our proposal in efficiency gain. Also, for the estimates of the kernel function, κ , characterizing the association between Y_j and S_j , our proposed model has substantial improvement with an increased sample size and a decreased rate of zero-inflation. These results show much better performance compared to the death cases analysis in terms of less bias and variability, and get similar to the full data analysis results particularly when the sample size is large enough. Overall, higher zero rates lead to better estimates for p , while lower zero rates produce better estimates with lower SDs for the parameter estimates in Poisson regressions and mixed-effects models. Tables 2 and 3 show results from the scenarios with different sample size and censoring rates, and we can find out that higher censoring rates lead to increased bias and standard deviation of κ , but will not affect substantially for the parameters β_1 and β_2 . Also, the robustness of our method regarding to the kernel function is empirically explored, and negligible effects (relative bias in percent $<5\%$) are recognized for primary parameter estimates. Further, the asymptotic normality of parameter estimates are also evaluated and verified from extensive simulation. Overall, the bootstrap method performs well in estimating the variance, yielding satisfactory CPs especially for a relatively large sample size (ie, $n = 900$).

4 | DATA APPLICATION TO THE ASSESS-AKI STUDY

The ASSESS-AKI study, a multicenter cohort study, includes 1603 hospitalized adult participants who do or do not have an episode of AKI at baseline stratified by CKD and follows them up to 84 months. Note that the work by Ikizler et al (2020)³⁷ utilized the subset of 1538 adults matched by baseline CKD/AKI due to the focus of primary outcome analysis. Here, we will leverage all available data from the baseline visit and all subsequent recurrent episodes of hospitalized AKI with inpatient serum creatinine measurements collected. In particular, the AKI episode is determined based on the definition of a 50% increase in serum creatinine from nadir to peak during hospital admission. Also, the peak serum creatinine during hospital admission will be utilized for profile evaluation of kidney function over time. In addition, renal outcomes, including development of end-stage renal disease (ESRD) and also mortality, will be monitored during follow-up. Here, we treat both ESRD and death as terminal events due to irreversible renal failure, and so the time to the terminal event will be the earlier of the times for these two events, and the censoring time is defined as the last contact or follow-up date.

From Figure 1, we present basic summary information of the ASSESS-AKI data. In particular, we find out that a larger variability in serum creatinine exists for the subjects with less AKI recurrences; however, the average levels are slightly different across cluster size. Also, from random subjects with death, the trajectory of serum creatinine shows an increasing trend, which tends to be largest particularly when the death event is approaching. Also, in Figure 1C, the proportions of participants by cluster size (ie, the number of AKI recurrences within the follow-up period) in the ASSESS-AKI study are provided. We can see a high percentage of participants with zero recurrence of hospitalized AKI, which is about 67.31% (1079 participants among 1603 in total), and also the maximum cluster size is up to 23 during the follow-up. If we further divide participants with regard to baseline

AKI and baseline CKD, then there will be four groups (baseline non-AKI and non-CKD, baseline non-AKI but CKD, baseline AKI but non-CKD, baseline AKI and CKD). In the group of baseline non-AKI and non-CKD, the percentage of participants with zero recurrent AKI is $439/522=84.1\%$, whereas in the group of baseline AKI and CKD, the percentage of participants with zero recurrent AKI is $150/309=48.5\%$, which is the lowest among the four groups. We find that baseline AKI and baseline CKD status have strong effects on future AKI events. More detailed information can be seen in the Supplementary Material and the primary clinical paper.^{27,28} By the end of the ASSESS-AKI study, there are 355 individuals who have died or have developed ESRD, among whom about 40.56% (144/355) have both baseline AKI and CKD and 14.93% (53/355) are free of AKI and CKD at baseline. The Kaplan-Meier curves in Figure 1D provides more information on the effect of AKI or CKD on risk of death/ESRD with the subgroup with both AKI and CKD at baseline having the highest risk, providing evidence for informative dropouts due to the terminal events that have potential correlations with progressive kidney function. Note that among those participants who have not experienced any AKI during the follow-up, some may be in good kidney condition completely free of such renal events, whereas some may be at some risk of AKI but have been censored by death/ESRD or last follow-up time. Thus, baseline status of AKI and CKD will be the key covariates for analysis. Besides these, we also include other available and important risk factors (ie, age, gender, diabetes, hypertension) for survival regression depending on clinical knowledge and prior studies.^{28,38} The kidney function, measured by serum creatinine (after log transformation), is modeled by the following random effect model, conditioning on the terminal event S_i ,

$$\begin{aligned} \log(Y_{ij}) = & \beta_0 + \beta_{\text{CKD}} \times \text{CKD} + \beta_{\text{AKI}} \times \text{AKI} + \beta_{\text{Female}} \times \text{Female} + \beta_{\text{Black}} \times \text{Black} \\ & + \beta_{\text{Age}} \times \text{Age} + \beta_{\text{Diabetes}} \times \text{Diabetes} \\ & + \kappa_1 \exp\left[-\kappa_2(S_i - t_{ij} - \kappa_3)^2\right] + \gamma_i + \epsilon_{ij}. \end{aligned} \quad (7)$$

Also, we assume the ZIP distribution for the cluster size N_i with the probability of structure zeros modeled by

$$p_i = \frac{\exp(\theta_0 + \theta_{\text{CKD}} \times \text{CKD} + \theta_{\text{AKI}} \times \text{AKI})}{1 + \exp(\theta_0 + \theta_{\text{CKD}} \times \text{CKD} + \theta_{\text{AKI}} \times \text{AKI})},$$

and the parameter for the Poisson distribution

$$\lambda_i = \exp\left[\theta_0 + \theta_{\text{CKD}} \times \text{CKD} + \theta_{\text{AKI}} \times \text{AKI} + \theta_{l(V_i)} \times l(V_i) + \eta \times \gamma_i\right],$$

where $l(V_i) = \exp(a_0 + a_{\text{AKI}} \times \text{AKI} + a_{\text{CKD}} \times \text{CKD}) \times V_i$ demonstrates the survival risk at the observation time V_i . For this real data analysis, we consider three candidate models for the comparison of results. In particular, we generate three models to compare the estimation effects. In the first model (event cases), the data with at least one record of serum creatinine is utilized, and the conditional longitudinal mixed-effects model with a random intercept implemented in Equation (7) is used for parameter estimation. Under such context, the informative cluster size is ignored but the terminal event is still taken into account for analysis. In the second model (death cases), the longitudinal data with observed death cases

only are considered, and the first two stages of our proposed estimation procedure are used. In the third model (all cases), all observed data will be incorporated for analysis, and our proposed joint modeling method with the three-stage estimation procedure is adopted.

From Table 4, we see that participants with baseline CKD and AKI are more likely to have more recurrence events of AKI, and hence larger cluster size. For example, in all cases, compared with the participants without CKD at baseline, the participants had a $\exp(0.44) - 1 = 55\%$ increase in the number of future subsequent AKI recurrences with CKD at baseline, but no significant effect for death group. In addition, the participants with CKD and AKI at baseline had higher serum creatinine values, since a high serum creatinine level usually indicates that the kidney is not functioning well. Women have lower creatinine levels than men, in agreement with previous studies, because the amount of creatinine in the blood increases with muscle mass. The serum creatinine level increase somewhat when the observation time is very closer to death/ESRD. Previous studies have demonstrated the incidence of AKI is associated with higher mortality among hospitalized patients. Diabetes is a significant effect when we use simpler mixed effect model in the first model (event cases), but it is no longer significant based on our joint modeling approach. Also, we could see the random effect parameter η has a positive effect but is not significant in the joint models (death cases and all cases), indicating that cluster size may not be highly associated with serum creatinine (refer to Figure 1). Although the death case model (yielded similar point estimates of the parameters to those obtained from the all cases model, we note that the standard errors are generally smaller in this model, and this improved precision was due to the inclusion of participants without terminal events.

5 | DISCUSSION AND CONCLUSION

In this work, we propose an innovative joint modeling method with a three-stage estimation procedure for longitudinal data from a marked recurrent process by considering informative cluster size with zero-inflation and a terminal event. To address the deficiencies of existing approaches that only handle the situation with cluster size of at least one, we incorporate the large portion of participants with zero recurrence of events into the modeling by utilizing a ZIP distribution, which results in improved estimation of the longitudinal trajectory. Specifically, we consider conditional modeling of the longitudinal data given the terminal event, using a flexible kernel function, to explicitly account for the informative dropout due to the terminal event. Extensive simulation studies have been conducted for systematic evaluation of the performance of the proposed method. In particular, our proposed joint modeling with a three-stage estimation procedure is sufficiently flexible to allow its extension to more complicated scenarios, since we are incorporating more information compared with traditional methods. Note that the bootstrap technique is applied to obtain standard errors, leading to satisfactory coverage probabilities for the resulting confidence intervals. In addition, our method is robust to the specification of the kernel function, as shown by our simulation studies.

Furthermore, we gain more information from our proposed method by borrowing a variety of relevant information for inference compared with simple mixed-effects models. Our new estimation procedure is easy to implement in programming. The simulation studies also

show that our estimators are more efficient compared to those based on the complete case analysis (death only). However, with regard to real applications, there are several issues to which further attention needs to be paid. First, the improvement in efficiency could be less when the proportion of zero recurrences is relatively low (ie, <50%), which is anticipated owing to the difficulty in distinguishing between structure zeros and random zeros with limited data resources. Also, how to choose an appropriate set of starting values for the optimization function to maximize the likelihood is crucial for reducing computing burden, particularly under such a context with model complexity resulting from the need for a double integration. In addition, there are several lines of work that could be pursued in future studies. One of these is the possibility of generalization of the longitudinal model to non-Gaussian outcomes, for example, logistic or Poisson regressions with nonidentity link functions if the model is identifiable and regularity conditions are suitably modified. Also, we consider the model for the number of observations without the consideration of specific nonhomogeneous temporal pattern because of our motivation data, but, in some scenarios, this assumption that the observation times are not informative given the cluster size may be violated. Thus, directly modeling a “marked” recurrent event process could be more valid and efficient. Lastly, several other smoothing techniques (eg, cubic B-splines) could be considered for the function g , and ways in which an optimal functional form could be determined are of interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This project was supported in part by research grant U01 DK082183 from the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (4 U01 DK082183-09), Pennsylvania Department of Health and Human Services and by the National Center for Advancing Translational Sciences of the National Institutes of Health (KL2 TR000126 and TR002015). The work of Dr. Danping Liu is supported by the Intramural Research Program of the National Cancer Institute (NCI), the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding information

National Center for Advancing Translational Sciences, Grant/Award Number: KL2 TR000126 and TR002015; the National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: 4 U01 DK082183-09

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available upon reasonable request which should be directed to the Data Coordinating Center of the Assessment, Serial Evaluation, and Subsequent Sequelae in Acute Kidney Injury (ASSESS-AKI) at Penn State and by approval from the Ancillary study Committee of the ASSESS-AKI Study.

REFERENCES

1. Wang M, Kong M, Datta S. Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Stat Methods Med Res*. 2011;20(4):347–367. [PubMed: 20223781]

2. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sin.* 2004;14(3):809–834.
3. Hsieh F, Tseng YK, Wang JL. Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics.* 2006;62(4):1037–1043. [PubMed: 17156277]
4. Albert PS, Shih JH. An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *Ann Appl Stat.* 2010;4(3):1517. [PubMed: 21938267]
5. Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics.* 2004;60(3):747–756. [PubMed: 15339298]
6. Liu L, Huang X, Yaroshinsky A, Cormier JN. Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics.* 2016;72(1):204–214. [PubMed: 26295794]
7. Zeng D, Lin D. Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics.* 2009;65(3):746–752. [PubMed: 18945267]
8. Li Z, Chinchilli VM, Wang M. A Bayesian joint model of recurrent events and a terminal event. *Biom J.* 2019;61(1):187–202. [PubMed: 30479030]
9. Ghosh D, Lin DY. Marginal regression models for recurrent and terminal events. *Stat Sin.* 2002;12(3):663–688.
10. Ye Y, Kalbfleisch JD, Schaubel DE. Semiparametric analysis of correlated recurrent and terminal events. *Biometrics.* 2007;63(1):78–87. [PubMed: 17447932]
11. Kong S, Nan B, Kalbfleisch JD, Saran R, Hirth R. Conditional modeling of longitudinal data with terminal event. *J Am Stat Assoc.* 2018;113(521):357–368. [PubMed: 30853735]
12. Deng S, Zhao X. Covariate-adjusted regression for distorted longitudinal data with informative observation times. *J Am Stat Assoc.* 2019;114(527):1241–1250.
13. Moriyama N, Ishihara M, Noguchi T, et al. Early development of acute kidney injury is an independent predictor of in-hospital mortality in patients with acute myocardial infarction. *J Cardiol.* 2017;69(1):79–83. [PubMed: 26917196]
14. Harlow SD, Mitchell ES, Crawford S, et al. The ReSTAGE collaboration: defining optimal bleeding criteria for onset of early menopausal transition. *Fertil Steril.* 2008;89(1):129–140. [PubMed: 17681300]
15. Picone G, Mark Wilson R, Chou SY. Analysis of hospital length of stay and discharge destination using hazard functions with unmeasured heterogeneity. *Health Econ.* 2003;12(12):1021–1034. [PubMed: 14673811]
16. Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics.* 2003;59(1):36–42. [PubMed: 12762439]
17. Seaman S, Copas A. Doubly robust generalized estimating equations for longitudinal data. *Stat Med.* 2009;28(6):937–955. [PubMed: 19153970]
18. Wang MC, Qin J, Chiang CT. Analyzing recurrent event data with informative censoring. *J Am Stat Assoc.* 2001;96(455):1057–1065.
19. Zhang B, Chen Z, Albert PS. Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. *Biostatistics.* 2012;13(1):74–88. [PubMed: 21908867]
20. Zhang B, Liu W, Zhang Z, Qu Y, Chen Z, Albert PS. Modeling of correlated data with informative cluster sizes: an evaluation of joint modeling and within-cluster resampling approaches. *Stat Methods Med Res.* 2017;26(4):1881–1895. [PubMed: 26113386]
21. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics.* 1992;34(1):1–14.
22. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics.* 2000;56(4):1030–1039. [PubMed: 11129458]
23. Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Stat Methods Med Res.* 2002;11(4): 341–355. [PubMed: 12197301]
24. Xie H, McHugo G, Sengupta A, Clark R, Drake R. A method for analyzing longitudinal outcomes with many zeros. *Ment Health Serv Res.* 2004;6(4):239–246. [PubMed: 15588034]
25. Hsu CH. A weighted zero-inflated Poisson model for estimation of recurrence of adenomas. *Stat Methods Med Res.* 2007;16(2): 155–166. [PubMed: 17484298]

26. Xu C, Chinchilli VM, Wang M. Joint modeling of recurrent events and a terminal event adjusted for zero inflation and a matched design. *Stat Med*. 2018;37(18):2771–2786. [PubMed: 29682772]
27. Go AS, Parikh CR, Ikizler TA, et al. The assessment, serial evaluation, and subsequent sequelae of acute kidney injury (ASSESS-AKI) study: design and methods. *BMC Nephrol*. 2010;11(1):22. [PubMed: 20799966]
28. Hsu C, Chinchilli VM, Coca S, et al. Post-acute kidney injury proteinuria and subsequent kidney disease progression: the assessment, serial evaluation, and subsequent sequelae in acute kidney injury (ASSESS-AKI) study. *JAMA Intern Med*. 2020;180(3):402–410. [PubMed: 31985750]
29. Lo LJ, Go AS, Chertow GM, et al. Dialysis-requiring acute renal failure increases the risk of progressive chronic kidney disease. *Kidney Int*. 2009;76(8):893–899. [PubMed: 19641480]
30. Ishani A, Xue JL, Himmelfarb J, et al. Acute kidney injury increases risk of ESRD among elderly. *J Am Soc Nephrol*. 2009;20(1):223–228. [PubMed: 19020007]
31. Wald R, Quinn RR, Luo J, et al. Chronic dialysis and death among survivors of acute kidney injury requiring dialysis. *JAMA*. 2009;302(11):1179–1185. [PubMed: 19755696]
32. Chan KCG, Wang MC. Backward estimation of stochastic processes with failure events as time origins. *Ann Appl Stat*. 2010;4(3):1602. [PubMed: 21359167]
33. Lu X, Nan B, Song P, Sowers M. Longitudinal data analysis with event time as a covariate. *Stat Biosci*. 2010;2(1):65–80.
34. Breslow NE. Discussion of professor Cox's paper. *J Royal Stat Soc B*. 1972;34:216–217.
35. Pan J, Thompson R. Gauss-Hermite quadrature approximation for estimation in generalised linear mixed models. *Comput Stat*. 2003;18(1):57–78.
36. Pinheiro JC, Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J Comput Graph Stat*. 2006;15(1):58–81.
37. Ikizler TA, Parikh CR, Himmelfarb J, et al. Acute kidney injury and kidney outcomes, cardiovascular events and death: a prospective cohort study. *Kidney Int*. 2021; In press.
38. Wang HE, Muntner P, Chertow GM, Warnock DG. Acute kidney injury and mortality in hospitalized patients. *Am J Nephrol*. 2012;35(4):349–355. [PubMed: 22473149]

FIGURE 1.

A, the scatter plots of (log) serum creatinine by cluster size (ie, the number of AKI recurrences) for the subjects with death; B, the scatter plots of (log) serum creatinine by cluster size (ie, the number of AKI recurrences) for the subjects without death; C, the proportion of participants by the number of AKI recurrences in the ASSESS-AKI study; D, the Kaplan-Meier curves stratified by the baseline status of AKI and CKD (the P -value < .05 based on the log-rank test) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

TABLE 1

Result summary of the parameter estimates for the scenario with the zero rate of 40% and the censoring rate of 50%

Sample size	Parameter	Full data			Death cases			All cases			
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	CP (%)
n = 300	$\theta_0 = -1$	0.32	0.59	0.45	0.52	0.69	0.75	0.45	0.61	0.57	88
	$\theta_1 = -0.5$	-0.21	0.52	0.31	-0.47	0.63	0.62	-0.28	0.54	0.37	91
	$\vartheta_0 = 0.5$	0.27	0.50	0.32	0.44	0.66	0.62	0.29	0.52	0.35	91
	$\vartheta_1 = -0.5$	-0.18	0.38	0.18	-0.38	0.55	0.44	-0.28	0.44	0.27	92
	$\vartheta_2 = 1$	-0.19	0.47	0.26	-0.34	0.69	0.59	-0.26	0.51	0.33	91
	$\eta = 0.5$	0.43	0.39	0.34	0.51	0.65	0.68	0.47	0.49	0.46	90
	$\beta_0 = 2$	-0.07	0.13	0.02	-0.15	0.18	0.05	-0.08	0.15	0.03	90
	$\beta_1 = 1$	0.00	0.08	0.01	-0.02	0.11	0.01	0.00	0.08	0.01	93
	$\beta_2 = 1$	0.01	0.08	0.01	0.02	0.11	0.01	0.01	0.08	0.01	93
	$\kappa_1 = 1$	-0.15	0.19	0.06	-0.33	0.24	0.17	-0.23	0.20	0.09	92
	$\kappa_2 = 2$	0.19	0.35	0.16	0.49	0.52	0.51	0.31	0.43	0.28	90
	$\kappa_3 = 1$	-0.09	0.06	0.01	-0.25	0.13	0.08	-0.17	0.08	0.04	90
n = 600	$\theta_0 = -1$	0.25	0.53	0.34	0.41	0.55	0.47	0.29	0.56	0.40	89
	$\theta_1 = -0.5$	-0.17	0.39	0.18	-0.39	0.51	0.41	-0.19	0.43	0.22	91
	$\vartheta_0 = 0.5$	0.18	0.38	0.18	0.33	0.57	0.43	0.21	0.44	0.24	90
	$\vartheta_1 = -0.5$	-0.15	0.30	0.11	-0.30	0.39	0.24	-0.24	0.27	0.13	92
	$\vartheta_2 = 1$	-0.14	0.27	0.09	-0.37	0.43	0.32	-0.22	0.32	0.15	94
	$\eta = 0.5$	0.31	0.24	0.15	0.44	0.54	0.49	0.36	0.36	0.26	92
	$\beta_0 = 2$	-0.04	0.10	0.01	-0.09	0.15	0.03	-0.06	0.13	0.02	92
	$\beta_1 = 1$	0.01	0.06	0.00	0.01	0.09	0.01	0.01	0.07	0.01	94
	$\beta_2 = 1$	0.00	0.06	0.00	-0.01	0.08	0.01	0.00	0.06	0.00	93
	$\kappa_1 = 1$	-0.11	0.11	0.02	-0.25	0.18	0.09	-0.14	0.16	0.05	95
	$\kappa_2 = 2$	0.12	0.23	0.07	0.44	0.38	0.34	0.23	0.34	0.17	92
	$\kappa_3 = 1$	-0.05	0.04	0.00	-0.17	0.07	0.04	-0.12	0.05	0.02	93
n = 900	$\theta_0 = -1$	0.19	0.45	0.24	0.33	0.41	0.28	0.22	0.48	0.29	92
	$\theta_1 = -0.5$	-0.12	0.27	0.09	-0.36	0.45	0.33	-0.14	0.35	0.14	93
	$\vartheta_0 = 0.5$	0.16	0.38	0.17	0.26	0.43	0.25	0.17	0.32	0.13	93
	$\vartheta_1 = -0.5$	-0.12	0.17	0.04	-0.22	0.27	0.12	-0.21	0.19	0.08	92
	$\vartheta_2 = 1$	-0.11	0.16	0.04	-0.26	0.31	0.16	-0.19	0.24	0.09	94
	$\eta = 0.5$	0.24	0.17	0.09	0.32	0.41	0.27	0.28	0.29	0.16	93
	$\beta_0 = 2$	-0.01	0.06	0.00	-0.07	0.12	0.02	-0.05	0.09	0.01	93
	$\beta_1 = 1$	0.00	0.04	0.00	0.01	0.07	0.01	0.00	0.04	0.00	94
	$\beta_2 = 1$	0.00	0.04	0.00	0.00	0.07	0.00	0.00	0.04	0.00	95
	$\kappa_1 = 1$	-0.07	0.07	0.01	-0.19	0.11	0.05	-0.11	0.09	0.02	95
	$\kappa_2 = 2$	0.10	0.14	0.03	0.32	0.22	0.15	0.16	0.19	0.06	93

Sample size	Parameter	Full data			Death cases			All cases			
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	CP (%)
	$\kappa_3 = 1$	-0.03	0.01	0.00	-0.13	0.03	0.02	-0.09	0.02	0.01	94

Note: Full data: the estimates are based on all data with underlying true death time; death cases: the estimates are based on the data with observed death cases only; all cases: the estimates are based on all observed data using our proposed joint modeling. Bias: Monte Carlo bias; SD; Monte Carlo standard derivation; MSE: mean square error; CP (%): coverage probability of the 95% Bootstrap confidence interval.

TABLE 2

Result summary of the parameter estimates for the scenario with the zero rate of 70% and the censoring rate of 50%

Sample size	Parameter	Full data			Death cases			All cases			
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	CP(%)
n = 300	$\theta_0 = 1.5$	-0.26	0.50	0.32	-0.33	0.61	0.48	-0.29	0.55	0.39	89
	$\theta_1 = -1$	0.18	0.46	0.24	0.25	0.62	0.44	0.19	0.54	0.33	93
	$\vartheta_0 = 0.5$	0.25	0.49	0.30	0.32	0.64	0.51	0.27	0.51	0.33	93
	$\vartheta_1 = -0.5$	-0.14	0.30	0.11	-0.21	0.42	0.22	-0.16	0.32	0.13	91
	$\vartheta_2 = 1$	-0.16	0.33	0.13	-0.19	0.43	0.22	-0.15	0.37	0.16	90
	$\eta = 0.5$	0.54	0.45	0.49	0.76	0.72	1.10	0.67	0.58	0.79	89
	$\beta_0 = 2$	0.14	0.17	0.05	0.21	0.23	0.10	0.16	0.19	0.06	89
	$\beta_1 = 1$	0.02	0.11	0.01	0.02	0.16	0.03	0.03	0.12	0.01	93
	$\beta_2 = 1$	0.02	0.11	0.01	0.01	0.16	0.03	0.03	0.12	0.01	92
	$\kappa_1 = 1$	-0.21	0.18	0.08	-0.54	0.28	0.37	-0.37	0.23	0.19	93
	$\kappa_2 = 2$	0.44	0.46	0.41	0.65	0.74	0.97	0.52	0.56	0.58	90
	$\kappa_3 = 1$	-0.15	0.09	0.03	-0.30	0.14	0.06	-0.24	0.11	0.07	91
n = 600	$\theta_0 = 1.5$	-0.19	0.45	0.24	-0.24	0.47	0.28	-0.22	0.40	0.21	93
	$\theta_1 = -1$	0.12	0.31	0.11	0.19	0.41	0.20	0.14	0.37	0.16	94
	$\vartheta_0 = 0.5$	0.17	0.35	0.15	0.25	0.49	0.30	0.22	0.38	0.19	90
	$\vartheta_1 = -0.5$	-0.09	0.22	0.05	-0.19	0.31	0.13	-0.13	0.19	0.05	92
	$\vartheta_2 = 1$	-0.08	0.21	0.05	-0.18	0.34	0.15	-0.11	0.22	0.06	94
	$\eta = 0.5$	0.39	0.32	0.25	0.65	0.60	0.78	0.42	0.47	0.40	91
	$\beta_0 = 2$	0.09	0.13	0.03	0.16	0.18	0.06	0.11	0.15	0.03	92
	$\beta_1 = 1$	0.01	0.09	0.01	0.02	0.12	0.01	0.02	0.09	0.01	93
	$\beta_2 = 1$	0.01	0.09	0.01	0.01	0.12	0.01	0.01	0.10	0.01	92
	$\kappa_1 = 1$	-0.15	0.15	0.05	-0.33	0.22	0.16	-0.27	0.19	0.11	94
	$\kappa_2 = 2$	0.29	0.32	0.19	0.55	0.62	0.75	0.48	0.44	0.42	92
	$\kappa_3 = 1$	-0.10	0.06	0.01	-0.20	0.11	0.05	-0.16	0.09	0.03	92
n = 900	$\theta_0 = 1.5$	-0.13	0.32	0.12	-0.19	0.33	0.15	-0.16	0.36	0.16	92
	$\theta_1 = -1$	0.09	0.29	0.09	0.14	0.37	0.16	0.11	0.33	0.12	94
	$\vartheta_0 = 0.5$	0.11	0.20	0.05	0.16	0.34	0.14	0.12	0.24	0.07	92
	$\vartheta_1 = -0.5$	-0.06	0.15	0.03	-0.15	0.24	0.08	-0.10	0.18	0.04	93
	$\vartheta_2 = 1$	-0.05	0.14	0.02	-0.14	0.26	0.09	-0.09	0.19	0.04	93
	$\eta = 0.5$	0.31	0.26	0.16	0.53	0.53	0.56	0.33	0.32	0.21	92
	$\beta_0 = 2$	0.04	0.10	0.01	0.12	0.16	0.04	0.07	0.12	0.02	93
	$\beta_1 = 1$	0.01	0.06	0.00	0.01	0.09	0.01	0.01	0.06	0.00	93
	$\beta_2 = 1$	0.01	0.06	0.00	-0.01	0.09	0.01	0.01	0.06	0.00	95
	$\kappa_1 = 1$	-0.11	0.12	0.03	-0.28	0.18	0.11	-0.18	0.14	0.05	94
	$\kappa_2 = 2$	0.17	0.23	0.08	0.47	0.48	0.56	0.36	0.32	0.23	93

Sample size	Parameter	Full data			Death cases			All cases			
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	CP(%)
	$\kappa_3 = 1$	-0.06	0.04	0.01	-0.15	0.07	0.03	-0.11	0.06	0.02	93

Note: Full data: the estimates are based on all data with underlying true death time; death cases: the estimates are based on the data with observed death cases only; all cases: the estimates are based on all observed data using our proposed joint modeling. Bias: Monte Carlo bias; SD; Monte Carlo standard derivation; MSE: mean square error; CP (%): coverage probability of the 95% Bootstrap confidence interval.

TABLE 3

Result summary of the parameter estimates for the scenario with the zero rate of 70% and the censoring rate of 75%

Sample size	Parameter	Full data			Death cases			All cases			
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	CP (%)
n = 300	$\theta_0 = 1.5$	-0.26	0.50	0.32	-0.37	0.69	0.61	-0.31	0.57	0.42	90
	$\theta_1 = -1$	0.18	0.46	0.24	0.29	0.67	0.53	0.23	0.51	0.31	93
	$\vartheta_0 = 0.5$	0.25	0.49	0.30	0.38	0.78	0.75	0.30	0.55	0.39	92
	$\vartheta_1 = -0.5$	-0.14	0.30	0.11	-0.29	0.64	0.49	-0.21	0.39	0.20	91
	$\vartheta_2 = 1$	-0.16	0.33	0.13	-0.26	0.59	0.42	-0.20	0.32	0.14	90
	$\eta = 0.5$	0.54	0.45	0.49	0.83	0.77	1.28	0.71	0.59	0.85	88
	$\beta_0 = 2$	0.14	0.17	0.05	0.25	0.26	0.13	0.21	0.22	0.09	90
	$\beta_1 = 1$	0.02	0.11	0.01	0.05	0.21	0.05	0.03	0.13	0.02	93
	$\beta_2 = 1$	0.02	0.11	0.01	0.04	0.23	0.05	0.02	0.12	0.01	93
	$\kappa_1 = 1$	-0.21	0.18	0.08	-0.60	0.32	0.46	-0.42	0.26	0.24	91
	$\kappa_2 = 2$	0.44	0.46	0.41	0.81	0.84	1.36	0.62	0.69	0.86	89
	$\kappa_3 = 1$	-0.15	0.09	0.03	-0.44	0.17	0.22	-0.34	0.13	0.13	88
n = 600	$\theta_0 = 1.5$	-0.19	0.45	0.24	-0.29	0.63	0.48	-0.23	0.43	0.24	92
	$\theta_1 = -1$	0.12	0.31	0.11	0.24	0.59	0.41	0.17	0.39	0.18	93
	$\vartheta_0 = 0.5$	0.17	0.35	0.15	0.31	0.64	0.51	0.22	0.36	0.18	90
	$\vartheta_1 = -0.5$	-0.09	0.22	0.05	-0.22	0.42	0.22	-0.16	0.26	0.09	93
	$\vartheta_2 = 1$	-0.08	0.21	0.05	-0.21	0.51	0.30	-0.15	0.28	0.10	94
	$\eta = 0.5$	0.39	0.32	0.25	0.72	0.62	0.90	0.57	0.49	0.57	90
	$\beta_0 = 2$	0.09	0.13	0.03	0.19	0.21	0.08	0.15	0.19	0.06	93
	$\beta_1 = 1$	0.01	0.09	0.01	0.03	0.17	0.03	0.02	0.09	0.01	93
	$\beta_2 = 1$	0.01	0.09	0.01	0.03	0.16	0.03	0.01	0.09	0.01	94
	$\kappa_1 = 1$	-0.15	0.15	0.05	-0.45	0.27	0.28	-0.32	0.23	0.15	92
	$\kappa_2 = 2$	0.29	0.32	0.19	0.72	0.71	1.02	0.53	0.53	0.56	91
	$\kappa_3 = 1$	-0.10	0.06	0.01	-0.31	0.64	0.51	-0.21	0.10	0.05	91
n = 900	$\theta_0 = 1.5$	-0.13	0.32	0.12	-0.22	0.58	0.38	-0.19	0.54	0.33	91
	$\theta_1 = -1$	0.09	0.29	0.09	0.20	0.50	0.29	0.13	0.38	0.16	94
	$\vartheta_0 = 0.5$	0.11	0.20	0.05	0.21	0.55	0.35	0.17	0.29	0.11	93
	$\vartheta_1 = -0.5$	-0.06	0.15	0.03	-0.17	0.29	0.11	-0.11	0.20	0.05	94
	$\vartheta_2 = 1$	-0.05	0.14	0.02	-0.14	0.34	0.14	-0.11	0.21	0.06	92
	$\eta = 0.5$	0.31	0.26	0.16	0.59	0.59	0.70	0.36	0.38	0.27	91
	$\beta_0 = 2$	0.04	0.10	0.01	0.14	0.24	0.08	0.09	0.15	0.03	93
	$\beta_1 = 1$	0.01	0.06	0.00	0.02	0.12	0.01	0.01	0.07	0.01	94
	$\beta_2 = 1$	0.01	0.06	0.00	-0.01	0.13	0.02	0.01	0.07	0.01	95
	$\kappa_1 = 1$	-0.11	0.12	0.03	-0.40	0.21	0.20	-0.27	0.19	0.11	93
	$\kappa_2 = 2$	0.17	0.23	0.08	0.63	0.58	0.73	0.48	0.41	0.40	92

Sample size	Parameter	Full data			Death cases			All cases			
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	CP (%)
	$\kappa_3 = 1$	-0.06	0.04	0.01	-0.25	0.12	0.08	-0.18	0.08	0.04	93

Note: Full data: the estimates are based on all data with underlying true death time; death cases: the estimates are based on the data with observed death cases only; all cases: the estimates are based on all observed data using our proposed joint modeling. Bias: Monte Carlo bias; SD; Monte Carlo standard derivation; MSE: mean square error; CP (%): coverage probability of the 95% Bootstrap confidence interval.

TABLE 4

Results summary of parameter estimates with bootstrap standard errors for the ASSESS-AKI study

Parameter	Event cases (n = 524)	Death cases (n = 355)	All cases (n = 1603)
θ_{CKD}	—	-1.49 (0.51) *	-1.58 (0.41) *
θ_{AKI}	—	-1.37 (0.48) *	-1.12 (0.38) *
ϑ_{CKD}	—	0.45 (0.29)	0.44 (0.18) *
ϑ_{AKI}	—	0.56 (0.27) *	0.41 (0.19) *
ϑ_{All}	—	0.16 (0.11)	0.05 (0.08)
η	—	1.17 (1.11)	0.89 (0.52)
β_{CKD}	0.48 (0.04) *	0.31 (0.06) *	0.43 (0.04) *
β_{AKI}	0.16 (0.04) *	0.19 (0.07) *	0.14 (0.05) *
$\beta_{Diabetes}$	0.07 (0.04) *	0.02 (0.06)	0.01 (0.04)
β_{Female}	-0.14 (0.05) *	-0.15 (0.06) *	-0.13 (0.04) *
β_{Black}	0.15 (0.05) *	0.12 (0.07)	0.07 (0.05) *
β_{Age}	-0.01 (0.00) *	-0.01 (0.00) *	-0.01 (0.00) *
κ_1	0.39 (0.16) *	0.63 (0.19) *	0.42 (0.14) *
κ_2	1.87 (0.62) *	2.02 (0.59) *	2.08 (0.47) *
κ_3	0.07 (0.19)	0.03 (0.15)	0.05 (0.10)

Note: Event cases (n = 524): the estimates are based on the data with at least one record of serum creatinine and a simpler mixed effect model is fitted without considering the death events; death cases (n = 355): the estimates are based on the data with observed death cases only; all cases (n = 1603): the estimates are based on all observed data using our proposed joint modeling. Note that standard errors and *P*-values are calculated from 250 bootstrap samples.

* denotes the *P*-value < .05.