Received: 16 July 2019 | Revised: 15 January 2020 | Accepted: 19 May 2020

DOI: 10.1002/sim.8663

RESEARCH ARTICLE

Statistics in Medicine WILEY

A multistate joint model for interval-censored event-history data subject to within-unit clustering and informative missingness, with application to neurocysticercosis research

Hongbin Zhang[1] | **Elizabeth A. Kelvin**[1] | **Arturo Carpio**[2] | **W. Allen Hauser**[3]

[1]Department of Epidemiology and Biostatistics, Graduate School of Public Health and Health Policy, Institute for Implementation Science in Population Health, City University of New York, New York, New York

[2]School of Medicine, University of Cuenca, Cuenca, Ecuador

[3]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York

Correspondence
Hongbin Zhang, School of Public Health, City University of New York, 55 West 125th Street, New York, NY 10027.
Email: hongbin.zhang@sph.cuny.edu

We propose a multistate joint model to analyze interval-censored event-history data subject to within-unit clustering and nonignorable missing data. The model is motivated by a study of the neurocysticercosis (NC) cyst evolution at the cyst-level, taking into account the multiple cysts phases with intermittent missing data and loss to follow-up, as well as the intra-brain clustering of observations made on a predefined data collection schedule. Of particular interest in this study is the description of the process leading to cyst resolution, and whether this process varies by antiparasitic treatment. The model uses shared random effects to account for within-brain correlation and to explain the hidden heterogeneity governing the missing data mechanism. We developed a likelihood-based method using a Monte Carlo EM algorithm for the inference. The practical utility of the methods is illustrated using data from a randomized controlled trial on the effect of antiparasitic treatment with albendazole on NC cysts among patients from six hospitals in Ecuador. Simulation results demonstrate that the proposed methods perform well in the finite sample and misspecified models that ignore the data complexities could lead to substantial biases.

KEYWORDS

neurocysticercosis, multistate joint model, interval-censoring, frailty survival model, nonignorable missingness

1 | INTRODUCTION

Neurocysticercosis (NC) is an infection of the central nervous system (CNS) with the larval stage of Taenia solium,[1] the pork tapeworm. NC is the most common parasitic disease of the CNS and a major cause of seizures and other neurological symptoms in endemic countries.[2] The World Health Organization has estimated that there are 50 million new NC cases and 50 000 deaths related to NC worldwide each year.[3] When located in the human brain, the larval stage of T. solium appears to pass through three distinct stages of evolution before total disappearance (eg, References 4,5). In the first stage, the parasite is viable or alive, and able to avoid detection by the host's immune system by secreting immunomodulatory molecules. These cysts are classified as being in the active phase. In the second phase, the parasite is degenerating (colloidal and granular-nodular forms) and is targeted by the host's immune system. This stage is called the transitional or degenerative phase and is most frequently associated with symptomatic disease. After the parasite dies, a calcified nodule

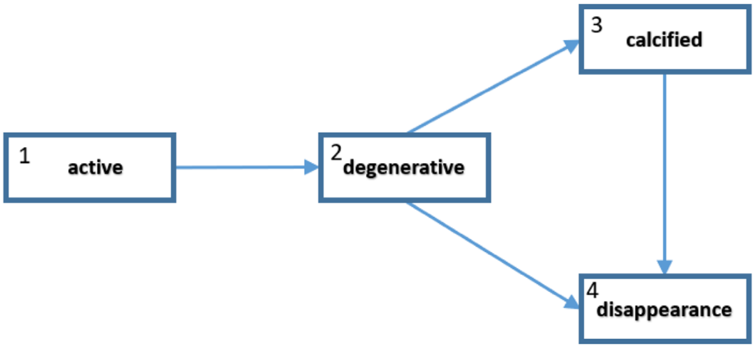

FIGURE 1 An irreversible disease progression evolution with four states for neurocysticercosis [Color figure can be viewed at wileyonlinelibrary.com]

sometimes remains in its place; this is termed the calcified phase, while in other cases the cysts disappear, referred to as complete resolution.

Existing NC research has primarily focused on aggregated patient-level measures such as the presence of any NC cysts and number or percent of a certain type of cyst.[6-8] Studying cysts at the patient level may help identify factors associated with differences across patients, but has limitations in understanding the natural history of cysts in the brain such as the intensity of transition and the length of time cysts take to progress through each phase. For example, it appears that antihelmintic drugs, such as albendazole (ALB), decrease the burden of active cysts in the brain parenchyma; however, studies on the impact of ALB on degenerative cysts and cysts located in extraparenchymal brain locations are inconclusive.[9,10] Also, to what extent and length cysts stay in each particular phase, with or without the treatment, are largely unknown.

In this paper, we are interested in formulating a flexible multistate model to describe the evolution of NC cysts and estimate the treatment effect on evolution. The data motivating the proposed research come from a randomized clinical trials conducted in Ecuador assessing the effectiveness of ALB for newly diagnosed NC patients.[8] The trial has been deemed one of only two high-quality trials on the impact of ALB for treatment of NC.[11] Brain image on cyst type and location were obtained over five time points (at baseline and 1, 6, 12, and 24 months) from 2001 to 2005. These cyst-level data provide a unique opportunity to understand the intrabrain distribution of the life course of NC cysts to guide future treatment (see Reference 12). We characterize the NC evolution with three transient states (active, degenerative, calcification) and one absorbing state (disappearance), as seen in Figure 1. For NC, only these transitions from state r to state s, $(r, s) \in \{(1, 2), (2, 3), (2, 4), (3, 4)\}$, are biological meaningful, that is, a cyst transits from active phase to degenerative phase and then to either calcification or be dissolved. Although rare, calcified cysts can also be dissolved occasionally.[13]

The multiple cyst phases of the NC data over time construct a multivariate event-history data. The data are interval-censored due to the fixed-schedule for imaging, meaning that we have data on the cysts phases only at pre-specified time points and the exact timing of a cyst transition to a new phase is known only within an interval. Additionally, when multiple NC cysts evolutions were abstracted from the brain of the same patient, those evolutions were correlated (intrabrain correlation). Moreover, we have missing data on our outcome due to missed clinic visits, loss to follow-up, and death, which might be nonignorable (or informative) in the sense that the missingness may be related to the unobserved values.[14] Several researchers have considered multistate models for interval-censored multivariate event-history data, see, for example, Kalbfleisch and Lawless,[15] Satten and Longini[16] and Jackson et al.[17] For nonignorable missing data, Hout and Matthews[18] took a selection model based approach for independent disease progressions. For intrasubject correlated data, Pak et al[19,20] extended frailty survival model for interval-censored data and used approximated likelihood methods, for example, Gaussian-quadrature and the so-called h-likelihood[21] for the inference.

Our data are also left-censored because the date of infection and, therefore, the onset time for the active or degenerative cysts identified at baseline is unknown. We have identified some common strategies in the literature to deal with the left-censoring problem. Satten and Sternberg[22] assume that the time elapsed before the first observation follows a given distribution and is independent of the time to the next transition from the first observation. Satten and Longini[16] develop a procedure to estimate Markov model parameters that conditions on the initiation time in order to remove dependence on this time. Kalbfleisch and Lawless[15] simply assume that the holding time of the initial state is exponentially distributed, rendering the time origin unnecessary due to the memoryless property of the exponential distribution. In this study, we use this strategy to simplify our model and focus on interval-censored event-history data subject to within-brain correlation and nonignorable missingness. Methods for similar data have not been studied in the literature, to the best of our knowledge. We propose a joint model framework where shared random effects (ie, frailty) are introduced to account for with-individual clustering and the unobserved individual characteristics that influence the missingness. The analysis is nontrivial since interval censoring causes theoretical difficulty for the use of counting process techniques, hence prohibiting the use of martingale theory. Numerical methods are often used instead, which are often complex, inducing difficulties for the implementation. The methods previously used in the literature, such as likelihood-approximation, can be computationally very intensive and may have convergence problems. For the inference of the multistate joint model in this paper, we use the Monte Carlo EM (MCEM) algorithm, which is more stable and provides "exact" likelihood-based estimation.[23] MCEM algorithm involves multivariate Monte Carlo sampling with rejection, which is intrinsically challenging for interval-censored data. Incorporating the handling of missing data and correlated data leads to additional computational complexity.

The rest of the paper is organized as follows. Section 2.1 introduces the joint model and the likelihood. Section 2.2 discusses the MCEM method and inference. In Section 3, we apply the method to the NC data, and in Section 4, we conduct a simulation study. Section 5 gives some concluding remarks.

2 | STATISTICAL METHODS

2.1 | Joint model and likelihood

For an individual i with cyst j, $i = 1, \dots, n, j = 1, \dots, n_i$, we consider the cyst evolution process $y_{ij}(t) \equiv y_{ij,t}$ for continuous time t $(t \geq 0)$. The transition intensity from state r to state s at time t, defined as instantaneous probability, $\lim_{\delta t \to 0} P(y_{ij,t+\delta t} = s | y_{ij,t} = r, \boldsymbol{x}_i, \omega_{i,rs})$, is modeled by

$$q_{rs}(t|\boldsymbol{x}_{ij}, \omega_{i,rs}) = q_{rs}^{(0)}(t) \exp(\boldsymbol{\beta}_{rs}^T \boldsymbol{x}_i + \omega_{i,rs}), \tag{1}$$

where $q_{rs}^{(0)}(t)$ is the transition-specific baseline intensity function which is assumed to follow an Exponential distribution, $\boldsymbol{\beta}_{rs}$ are the transition-specific regression coefficients, $\boldsymbol{x}_i$ is the covariate vector (eg, treatment, demographic variables), and $\omega_{i,rs}$ is the individual random effects for the transition. Defining $\boldsymbol{\omega}_i = (\omega_{i,12}, \omega_{i,23}, \omega_{i,24}, \omega_{i,34})^T$, we assume

$$\boldsymbol{\omega}_i \sim N(\boldsymbol{0}, D), \quad \text{where} \quad D = \begin{pmatrix} \sigma_{12}^2 & \sigma_{12}\sigma_{23}\rho_{12,23} & \sigma_{12}\sigma_{24}\rho_{12,24} & \sigma_{12}\sigma_{34}\rho_{12,34} \\ & \sigma_{23}^2 & \sigma_{23}\sigma_{24}\rho_{23,24} & \sigma_{23}\sigma_{34}\rho_{23,34} \\ & & \sigma_{24}^2 & \sigma_{24}\sigma_{34}\rho_{24,34} \\ & & & \sigma_{34}^2 \end{pmatrix}.$$

Here the random effects are introduced to account for the intrabrain association of the cysts progressions among different locations, but also represent the heterogeneity in cysts transition intensities between subjects that is not captured by the observed covariates. In addition, the correlation among the transition specific random effects reflects the association among the within-brain transitions. For example, a positive $\rho_{12,24}$ implies that the two intensities evolve similarly (for example, if the transition from 1 to 2 is fast and so is 2 to 4).

For the missing state problem, let $\boldsymbol{y}_{ij}^c$ be the complete-data trajectory of the cyst over the prescheduled imaging visits indexed by $k, k = 1, \dots, m$, thus, $\boldsymbol{y}_{ij}^c = (y_{ij,t_1}, \dots, y_{ij,t_m})$. Denote $r_{ij,k}$ as the observation indicator of visit k at time t_k such that $r_{ij,k} = 1$ if state y_{i,j,t_k} is observed and $r_{ij,k} = 0$ otherwise. We assume that the baseline state is always observed; that is $r_{ij,1} = 1$. Using the conditional Markov assumption, the contribution of the cyst to the likelihood of $\boldsymbol{y}_{ij}^c$ given the covariates and the random effects can be written as

$$L_{ij}^c(\boldsymbol{y}_{ij}^c) = P(y_{ij,t_1}) \prod_{k=2}^{m} \left[P(y_{ij,t_k} | y_{ij,t_{k-1}}, \boldsymbol{x}_i, \boldsymbol{\omega}_i) P(r_{ij,k} | y_{ij,t_k}, \boldsymbol{x}_i, \boldsymbol{\omega}_i) \right], \tag{2}$$

where $P(y_{ij,t_k} | y_{ij,t_{k-1}}, \boldsymbol{x}_i, \boldsymbol{\omega}_i)$ is the transition probability for a cyst to move from state $y_{ij,t_{k-1}}$ at visit $k-1$ to state y_{ij,t_k} at visit k. The quantity $P(r_{ij,k} | y_{ij,t_k}, \boldsymbol{x}_i, \boldsymbol{\omega}_i)$ in (2) represents the selection model approach for non-ignorable missingness where "missing is not at random" (MNAR) is assumed for the missing data mechanism.[14] Other mechanism, for example, missing at random (MAR) can also be specified (see Data Analysis Section) for sensitivity analysis. For the NC data, we assume that the probability of observing a state $y_{ij,t_k} = y_s, y_s \in \{1, 2, 3, 4\}$, can be described by a logistic model

$$p_{y_s}(r_{ij,k} = 1 | y_{ij,t_k} = y_s, \boldsymbol{x}_i, \boldsymbol{\omega}_i) = \frac{\exp(\boldsymbol{\alpha}_{y_s}^T \boldsymbol{x}_i + \boldsymbol{\xi}_{y_s}^T \boldsymbol{\omega}_i)}{1 + \exp(\boldsymbol{\alpha}_{y_s}^T \boldsymbol{x}_i + \boldsymbol{\xi}_{y_s}^T \boldsymbol{\omega}_i)}, \tag{3}$$

where $\boldsymbol{\alpha}_{y_s}$ are the coefficients for the covariates, $\boldsymbol{\xi}_{y_s}$ are the coefficients for the random effects. For ease of notation, we use the same covariate vector as in model (1) while other choices are available as will be illustrated in the Data Analysis section. We include the random effects to capture the possible association of latent individual characteristics and the missingness. In other words, the probability of missing data is related to the random effects $\boldsymbol{\omega}_i$ for unobserved covariates, which appears to be reasonable for the NC data. Therefore, the random effects are used to account for within-brain correlation and to account for variability of the missing data probabilities in our model setting, under the shared-parameters modeling framework.[24-26] Note that the same modeling framework apply to the problem of intermittent missing states and the missing due to loss to follow-up. We use the 24-month imaging time as the end of study. For each cyst in the study, we can classify its progression stage as either active, degenerative, calcified, resolved or missing, therefore, right censoring is not an issue.

Let $\boldsymbol{\theta}$ be the collection of baseline hazard parameters, regression coefficients for the transition intensity model (1) and the missing data logistic regression model (3), as well as the dispersion parameters in D for the frailties. Also, let $f(\cdot)$ denote a generic density function. Under the assumption that the cyst evolution processes are independent across subjects, the joint log-likelihood of the observed data Y is

$$l(Y|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \left[\sum_{\boldsymbol{y}_{ij}^c \in \Omega(\boldsymbol{y}_{ij})} L_{ij}^c(\boldsymbol{y}_{ij}^c | \boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}) \right] f(\boldsymbol{\omega}_i; \boldsymbol{\theta}) d\boldsymbol{\omega}_i \right), \tag{4}$$

where the integral can be multidimension, $\boldsymbol{y}_{ij}$ is the observed profile for the cyst, and $\Omega(\boldsymbol{y}_{ij})$ is the set with all the trajectories where missing states are replaced by feasible latent states. For our four-state survival model, only patterns with monotone increase are possible. For example, if $\boldsymbol{y}_{ij} = (1, 1, 3, \bullet, \bullet)$ where the $\bullet$ represents missed state, that is, patient loss to follow-up after the first three visits, then we have $\Omega(\boldsymbol{y}_{ij}) = \{(1, 1, 3, 3, 3), (1, 1, 3, 3, 4), (1, 1, 3, 4)\}$.

2.2 | Inferences method— A MCEM algorithm

As described in Section 1, the transition times between cyst phases in the NC study are only known up to an interval between two consecutive imaging visits. The within-brain correlation and missing data further complicate the estimation of the multistate model for NC evolution. When transition onset time can be exactly measured, transition-specific partial likelihood and Breshlow's estimator (see section 8.3 of Reference 27) can be used, even after conditioning on random effects. Without clustering, for nonignorable missing data, the transition intensities and the parameters for state observation probabilities can be estimated separately estimated as shown in Reference 18, assuming distinct parameters. For our data, full-likelihood-based joint modeling is proposed in Section 2.1. The key implementation difficulty is that the integral in the likelihood (4) which is typically quite intractable due to the interval-censoring and nonignorable missing induced structural complexity. When the dimension of the random effect, that is, $\dim(\boldsymbol{\omega}_i)$, is not low, numerical methods such as Gaussian quadrature can be computationally very intensive and may offer non-convergence. We, therefore, propose a Monte Carlo EM algorithm.

The EM algorithm is a standard approach for likelihood estimation in the presence of missing data. When the E-step is highly complicated, Monte Carlo methods can be used to approximate the expectation, leading to a MCEM algorithm. In our case, by treating the random effects $\boldsymbol{\omega}_i$ as additional "missing data", we have "complete data" $\{\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\omega}_i\}$ — where $\boldsymbol{y}_i$ represent the observed cyst profiles for individual i — and the "complete-data" log-likelihood function for individual i can be expressed as

$$l_c^i(\boldsymbol{\theta}) = \log f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}) + \log f(\boldsymbol{\omega}_i; \boldsymbol{\theta}), \tag{5}$$

where $f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} \left[\sum_{\boldsymbol{y}_{ij}^c \in \Omega(\boldsymbol{y}_{ij})} L_{ij}^c(\boldsymbol{y}_{ij}^c | \boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}) \right]$. Note that we denote $f(X|Y)$ a conditional density of X given Y.

The EM algorithm iterates between an E-step and a M-step until convergence. Let $\boldsymbol{\theta}^{(v)}$ be the parameter estimates from the vth EM iteration. The E-step for individual i at the $(v+1)$th EM iteration can be expressed as

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(v)}) = \int [\log f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}^{(v)}) + \log f(\boldsymbol{\omega}_i; \boldsymbol{\theta}^{(v)})] f(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)}) d\boldsymbol{\omega}_i. \tag{6}$$

The above E-step again involves an intractable integration. However, because expression (6) is an expectation with respect to $f(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)})$, it can be evaluated using the MCEM algorithm.[28,29]

Specifically, for individual i, let $\{\tilde{\boldsymbol{\omega}}_i^{(1)}, \ldots, \tilde{\boldsymbol{\omega}}_i^{(h_v)}\}$ denote a random sample of size h_v generated from $[\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)}]$ by multivariate rejection sampling (see section 3.2 of Reference 30) using the result $f(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)}) \propto f(\boldsymbol{\omega}_i; \boldsymbol{\theta}^{(v)}) \cdot f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}^{(v)})$ (see also Appendix A1). Then we can approximate the conditional expectation $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(v)})$ in the E-step by its empirical mean, with missing data replaced by simulated values, as follows

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(v)}) &\approx \frac{1}{h_v} \sum_{v=1}^{h_v} l_c^i(\boldsymbol{\theta}^{(v)}; \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\omega}_i^{(v)}) \\ &= \frac{1}{h_v} \sum_{v=1}^{h_v} \log f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\omega}_i^{(v)}; \boldsymbol{\theta}^{(v)}) + \frac{1}{h_v} \sum_{v=1}^{h_v} \log f(\boldsymbol{\omega}_i^{(v)}; \boldsymbol{\theta}^{(v)}). \end{aligned}$$

We may choose a reasonable h_0 and at the vth iterations, and set $h_v = h_{v-1} + h_{v-1}/c$, $v = 1, 2, 3, \ldots$, for some positive constant c. This way, h_v will increase with each EM iteration, which may speed up the EM convergence.[31] The M-step then maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(v)}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(v)})$ to produce an updated estimate $\hat{\boldsymbol{\theta}}^{(v+1)}$ for the $(v+1)$th iteration. Note that for both E-step and M-step, the evaluation of the probability transition functions in Equation (2) is required (see Appendix A2 for details).

To obtain the asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$, we can use the formula of Reference 32, which involves evaluating the second-order derivative of the complete-data log-likelihood function. Alternatively, we may consider the following approximate formula.[33] Let $S_c^i = \partial l_c^i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Then an approximate formula for the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \left[\sum_{i=1}^{n} \mathrm{E}(S_c^i|\boldsymbol{y}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}) \mathrm{E}(S_c^i|\boldsymbol{y}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}})^T \right]^{-1},$$

where the expectation can be approximated by MC empirical means, as above.

3 | DATA ANALYSIS

Between 2001 and 2005, researchers at Columbia University and the University of Cuenca, Ecuador conducted a multisite randomized clinical trial.[8] A total of 178 patients with newly diagnosed NC were recruited from six hospitals in Ecuador. Patients were eligible to participate if they had experienced a new onset of symptoms associated with NC within two months before recruitment and had active or transitional NC cysts identified on Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) image of the brain. Patients with calcified lesions were eligible to participate if they also had active or transitional cysts. Patients were randomly assigned to receive either ALB or placebo, given twice a day orally for 8 days. At enrollment, patients were interviewed to collect information about demographics, symptoms, and risk factors for NC. A brain CT or MRI with and without contrast was taken at baseline, 1, 6, 12, and 24 months of follow-up.

Radiologists read the CT/MRI images and completed a form documenting the number of cysts of each phase within each region of the brain from which patient-level summaries were evaluated.[8] In order to conduct cyst-level analysis, we disaggregated the data from the patient to the cyst level, generating a cyst-level dataset with a total of 210 cysts from 112 patients. Due to the difficulty in following the evolution of each individual cyst when there are multiple cysts in a single brain location, we excluded such cysts. Among the patients included in this sample, 59 (52%) individuals were treated with ALB. The average number of cysts per patient was 1.87, ranging from 1 to 8. Over the study period, five patients were lost to follow-up between 6- and 12-month, and 25 patients had 12-month imaging but were missing 24-month imaging data.

TABLE 1 Frequencies of observed transitions between states (act: active; deg: degenerative; cal: calcified; dis: dissolved; mis: missing)

ALB						**Placebo**					
	To						To				
From	act	deg	cal	dis	mis	From	act	deg	cal	dis	mis
act	36	9	2	47	7	act	58	8	0	24	6
deg	0	32	3	30	2	deg	0	52	5	32	7
cal	0	0	42	20	8	cal	0	0	28	14	10
mis	1	1	0	2	1	mis	0	1	2	3	0

We excluded patients who died in the study (n=7) to allow plausible data augment to the latent states. The objective of our data analysis was to assess the therapeutic effect of ALB on the individual NC cysts evolution, incorporating statistical methods to handle the complexities inherent in the data.

Table 1 presents frequencies of observed transitions between the states in the NC data stratified by the treatment group. Note that we removed the cases (n=13) of “reverse transitions," which are most likely related to challenges in comparison of images taken at different times where the height or the level of cut in the images is difficult to align for a patient/location. Therefore, there are no backward transitions from late-stage to an earlier stage in our data for these analyses. Transitions initiated from a missing state represent the intermittent missing, and there are 10 such cases. In addition, a total of 30 patients were lost to follow-up in this dataset. Among the 10 intermittent missing cases, six has nonconstant pre- and postmissing states. Although it is possible to treat the intermittent missing as data generated from wider interval, to seek missing data related insights, we model such missingness together with the lost to follow-up case in the same framework described in Section 2.

There are a variety of possible reasons why patients might have missing data. If an individual is missing imaging data for reasons unrelated to the outcome of cyst evolution, then the missing data does not violate model assumptions. However, if an individual is missing imaging because s/he recently had experienced a symptom, such as seizure, that prevented her/him from making the clinic appointment and is related to the outcome of NC cyst phase, then the missing state is nonignorable.[14] Our preliminary analysis indicates that there is no association between missing status and the treatment assignment, but there was an associations between missing status and the age, gender and imaging time variables. We therefore, fit the following joint model for the NC data.

$$\log(q_{rs}(\boldsymbol{x}_i, \omega_{i,rs})) = \boldsymbol{\beta}_{0,rs} + \boldsymbol{\beta}_{1,rs}\mathrm{trt}_i + \omega_{i,rs}, \tag{7}$$

$$\mathrm{logit}(p_{y_s}(r_{ij,k} = 1 | y_{ij,t_k} = y_s, \boldsymbol{x}_i, \boldsymbol{\omega}_i)) = \boldsymbol{\alpha}_{0,y_s} + \boldsymbol{\alpha}_{1,y_s} t_k + \boldsymbol{\alpha}_{2,y_s}\mathrm{age}_i + \boldsymbol{\alpha}_{3,y_s}\mathrm{sex}_i + \mathbb{1}^T \boldsymbol{\omega}_i, \tag{8}$$

where the quantities $q_{rs}(\cdot)$ and $p_{y_s}(\cdot)$ are defined in model (1) and (3), $\beta_{0,rs}$ is the log-transformed baseline hazard function parameter. The quantity trt_i is the treatment assignment for individual i with a value of “1" for ALB and “0" for placebo, t_k is the imaging time at visit k, age_i is a dichotomized baseline age with “0" for patients younger than 40 years old and “1" otherwise, sex_i is coded with “0" for female and “1" for male. Note that the random effects are included in the models to account for the within-unit clustering, also to partially explain the variability of the probability observing a state y_s where we use an “offset" type of modeling for the random effects due to the relative small sample size.

The joint model (7) resembles the MNAR (missing not at random) missing data mechanism where a logistic regression model is used for the probability of observing a state. In our study, we have a reasonable sense that those who skipped the imaging or dropped out of the study are not random, although such assumption is not verifiable with the current data.[14] Following Reference 18, we implemented the models corresponding to mechanism of MAR and MCAR (missing completely at random) as a sensitivity analysis. Both MNAR and MAR mechanisms augment the data to include the latent missing states but only MNAR incorporates the missing data probability models. The implementation for MCAR is an observed data analysis where the missing states are ignored and in such case, the likelihood is degenerated to interval-censored event-history data subject to within-unit clustering.

We apply our method on NC data under the three model assumptions. In addition, a so-called IND (for "independent") model is fitted where we assume no within-individual correlation; also, for the missing data problem, an approach similar to that in that in Reference 18 is applied. For example, the transition density parameters, $\boldsymbol{\beta}_{*,rs}$, are obtained by fitting standard multistate model with R's *msm* package assuming independent data and for the state observation probability parameters, $\boldsymbol{\alpha}_{*,y_s}$, we fit four separate generalized linear mixed-effects models corresponding to $y_s = \{1, 2, 3, 4\}$, using R's *lme*4 package. Although biases are expected, the IND model provides reasonable initial values for other more advanced models. When fitting the MNAR, MAR, and MCAR models using R, for the random effects, we start with an identity variance-covariance matrix and allow the MCEM algorithm to estimate the full structured variance-covariance parameters. We start the MCEM estimation procedure with $h_0 = 100$ Monte-Carlo samples and increase the Monte-Carlo sample size as the number of iteration v increases: $h_{v+1} = h_v + h_v/c$ with $c = 5$ (see Reference 31). The convergence of the EM algorithm was considered to achieve when the maximum percentage change of all estimates was less than 0.01 in two consecutive iterations. We observe that it takes about 30 iterations to achieve convergence, so the final Monte Carlo sample size for the standard error calculation is about 20 000.

Table 2 presents the results. There is not much difference between the MCAR and the MAR models in the point estimates, while the MAR missing data model tend to report larger standard error as the incorporation of missing data leads to higher degree of uncertainty. All models report significant treatment effect on transition from "active" to "degeneration" ($\beta_{1,12}$) and on the transition from "degenerative" to "disappearance" ($\beta_{1,24}$). The estimates from the MNAR model shows larger differential magnitude than other models. In particular, the treatment effect on the transition from the "active" stage to the "degenerative" stage is estimated to be ($\beta_{1,12} = 1.24, \text{se} = 0.40$) under MNAR while the estimates from other models are all smaller, ($\beta_{1,12} = 0.96, \text{se} = 0.21$), ($\beta_{1,12} = 1.04, \text{se} = 0.27$) and ($\beta_{1,12} = 1.21, \text{se} = 0.32$) for the IND, MCAR and MAR model, respectively. On the other hand, MNAR model's estimate for effect on the transition from "degenerative" to "disappearance" is $\beta_{1,24} = 0.87$, with SE 0.37 while the estimates and SEs for this effect are estimated to be ($\beta_{1,24} = 0.64, \text{SE} = 0.21$), ($\beta_{1,24} = 0.79, \text{SE} = 0.26$), ($\beta_{1,24} = 0.78, \text{SE} = 0.26$) under IND, MCAR, and MAR, respectively. It is worth noticing is that, negative point estimates were produced from all three missing data models for the treatment effect on transition from the state "degenerative" to the state "calcification", although nonsignificant.

Regarding the estimation of missing data model parameters, both the IND model and the MNAR model detect significant intercept and time effects on the probability of observing a state for the "active phase," $\alpha_{*,1}$, and the "calcified phase," $\alpha_{*,3}$. The effect of age is only significant for the "active phase" while the effect of gender is only significant for the "disappearance phase." Similar estimations of the variance-covariance matrix D for the random effects are obtained from the three missing data models, as seen in Table 2. Notice that the random effect for the transition from state "degenerative" to "calcification" is negatively correlated with the other three random effects, for example, the one for the transition from state "active" to "degenerative," the one for the transition from state "degenerative" to "disappearance" and the one for the transition from state "calcified" to "disappearance" while among these three random effects, positive correlations are obtained among those random effects.

4 | A SIMULATION STUDY

In this section, we evaluate the proposed model and method when there are different amounts of missing data and different degree of within-unit correlation, and we also assess model misspecifications through simulation. We generate the data based on the MNAR joint model for interval-censored event-history data subject to clustering and missing data, and then we conduct data analysis using the MNAR joint model and the misspecified IND model where both within-unit clustering and missing data are ignored and MCAR model where the missing data are ignored. We omit MAR in the simulation as its implementation incorporates the within-unit clustering and missing data handling, although differing from MNAR in that it dose not model the missing state probability. The true parameter values in the simulations are chosen to be the same (or similar) as those obtained from fitting the same models in the real data analysis. This setting allows us to validate the analysis results, in addition to the evaluation of model performance.

Specifically, for each sample γ, the transition intensity and state observation probability can be calculated from the multistate joint model as

$$\begin{aligned} q_{rs}^{(\gamma)} &= \exp(\hat{\boldsymbol{\beta}}_{0,rs} + \hat{\boldsymbol{\beta}}_{1,rs}\text{trt}_i^{(\gamma)} + \omega_i^{(\gamma)}), \\ p_{y_s}^{(\gamma)} &= 1/(1 + \exp[(\hat{\boldsymbol{\alpha}}_{0,y_s} + \hat{\boldsymbol{\alpha}}_{1,y_s} t_k + \omega_i^{(\gamma)})^{-1}]), \end{aligned}$$

TABLE 2 Results of multistate model under different assumptions, $est_{(se)}$, for neurocysticercosis data

θ	IND	MCAR	MAR	MNAR
Transition model				
$\beta_{0,12}$	$-2.42_{(0.11)}$	$-3.22_{(0.24)}$	$-3.27_{(0.27)}$	$-3.10_{(0.21)}$
$\beta_{0,23}$	$-4.44_{(0.26)}$	$-4.15_{(0.47)}$	$-4.13_{(0.48)}$	$-3.98_{(0.75)}$
$\beta_{0,24}$	$-2.03_{(0.10)}$	$-2.51_{(0.21)}$	$-2.51_{(0.31)}$	$-2.52_{(0.21)}$
$\beta_{0,34}$	$-3.13_{(0.15)}$	$-2.94_{(0.30)}$	$-2.89_{(0.39)}$	$-2.91_{(0.41)}$
$\beta_{1,12}$	$0.96_{(0.21)}$	$1.04_{(0.27)}$	$1.21_{(0.32)}$	$1.24_{(0.40)}$
$\beta_{1,23}$	$0.12_{(0.62)}$	$-0.15_{(0.72)}$	$-0.15_{(0.73)}$	$-0.21_{(1.01)}$
$\beta_{1,24}$	$0.64_{(0.21)}$	$0.79_{(0.26)}$	$0.78_{(0.26)}$	$0.87_{(0.37)}$
$\beta_{1,34}$	$-0.03_{(0.30)}$	$-0.53_{(0.45)}$	$-0.52_{(0.41)}$	$-0.62_{(0.47)}$

Missing data model

	IND				MNAR			
	int	t_k	age	sex	int	t_k	age	sex
$\alpha_{*,1}$	$6.79_{(2.91)}$	$-6.32_{(1.94)}$	$-5.57_{(1.88)}$	$0.81_{(1.46)}$	$5.50_{(2.61)}$	$-5.09_{(1.66)}$	$-5.27_{(1.42)}$	$0.72_{(1.94)}$
$\alpha_{*,2}$	$4.43_{(3.31)}$	$-5.02_{(3.26)}$	$0.76_{(1.11)}$	$-1.43_{(1.07)}$	$7.72_{(4.56)}$	$-4.09_{(3.53)}$	$0.07_{(2.13)}$	$-2.67_{(3.22)}$
$\alpha_{*,3}$	$4.54_{(2.06)}$	$-5.17_{(1.23)}$	$1.78_{(1.88)}$	$1.33_{(1.49)}$	$4.84_{(1.25)}$	$-5.32_{(1.25)}$	$2.65_{(3.76)}$	$0.44_{(1.72)}$
$\alpha_{*,4}$	$5.97_{(3.11)}$	$-5.34_{(3.03)}$	$0.43_{(0.67)}$	$-1.33_{(0.74)}$	$4.01_{(3.18)}$	$-6.16_{(4.16)}$	$0.75_{(1.56)}$	$-3.42_{(1.46)}$

estimates of **the variance-covariance matrix parameters of random effects D**

MCAR MAR

$$\begin{pmatrix} 10.22_{(4.21)} & -0.89_{(0.43)} & 3.34_{(1.27)} & 0.67_{(0.40)} \\ & 1.22_{(0.45)} & -1.23_{(0.56)} & -0.60_{(0.41)} \\ & & 6.32_{(3.11)} & 1.34_{(0.78)} \\ & & & 1.72_{(0.66)} \end{pmatrix} \begin{pmatrix} 9.56_{(3.35)} & -0.71_{(0.42)} & 3.24_{(2.10)} & 0.65_{(0.39)} \\ & 1.44_{(0.46)} & -1.02_{(0.71)} & -0.61_{(0.43)} \\ & & 5.82_{(3.33)} & 1.00_{(0.71)} \\ & & & 1.44_{(0.43)} \end{pmatrix}$$

MNAR

$$\begin{pmatrix} 10.26_{(4.11)} & -0.66_{(0.32)} & 3.45_{(1.71)} & 1.01_{(0.38)} \\ & 1.11_{(0.41)} & -1.44_{(0.82)} & -0.81_{(0.44)} \\ & & 5.80_{(3.12)} & 1.31_{(0.87)} \\ & & & 1.48_{(0.45)} \end{pmatrix}$$

Abbreviations: IND, independent; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

where we suppress $(\boldsymbol{x}_i, \omega_i)$ in q_{rs} and p_{y_s} to ease the notation hereafter. The treatment assignment $\text{trt}_i^{(\gamma)}$ is obtained from binomial distribution with probability 0.5. For computational resource consideration, we generate random effects from univariate normal distribution, $\omega_i^{(\gamma)} \sim N(0, \sigma^2)$ which is shared by the four transitions; also, we simplify the missing data model to include only the intercept and the time variable.

We then generate the continuous-time transitions within the study period, from baseline to 24 months as below.

- From state 1 at $T_1 = 0$. A draw from the exponential distribution with rate $\lambda = q_{12}^{(\gamma)}$ yields a time to event T_2. If $T_2 < 24$, set state 2 at T_2; otherwise, set $T_2 = 24$ and corresponding state as state 1.
- From state 2 at T_g (where $g = 1$ for cyst started with degenerative phase or $g = 2$ for cyst progressed from active phase). A draw from the exponential distribution with rate $\lambda = q_{23}^{(\gamma)} + q_{24}^{(\gamma)}$ yields a time to event T_{g+1}. If $T_{g+1} < 24$, then at T_{g+1}, choose state 3 with probability $q_{23}^{(\gamma)}\lambda^{-1}$ and state 4 with probability $q_{24}^{(\gamma)}\lambda^{-1}$; If $T_{g+1} >= 24$, set $T_{g+1} = 24$ and corresponding state as state 2.
- From state 3 at T_g (where g = 1 for cyst started with calcified phase or $g = 2$ for cyst evolved from degenerative phase or $g = 3$ for a cyst started with active phase then evolved to degenerative phase and then progressed to degenerative

phase). A draw from the exponential distribution with rate $\lambda = q_{34}^{(\gamma)}$ yields a time to event T_{g+1}. If $T_{g+1} < 24$, choose state 4 at T_{g+1}, otherwise, set $T_{g+1} = 24$ and corresponding state as state 3.

Note that for the transitions from state 1 to state 2 and from state 3 to state 4, we employ the commonly used inversion method for random number generating. For the transitions from state 2, the scheme is a Gillespie algorithm[34] —it is a stochastic approach to the differential equation that describes the relation between transition intensities and the transition probabilities.[35] Ideally, we would like to be able to observe exact transition times. However, in practice it is not always possible to monitor the process continually and the states are observed only at a limited number of times and within the limited time window of the study. We simulate this scenario by imposing our original RCT's study design on the simulated continuous time transitions, for example, imaging take place at 0, 1, 6, 12, and 24 months. Also, at the imaging time beyond the baseline, we simulate the missing state status by generating a binary state observation indicator using the missing data model generated probability $p_{y_s}^{(\gamma)}$ defined above. We assume that 50% of the cysts start in state 1 (active phase), 30% in state 2 (degenerative phase), and 20% in state 3 (calcified phase). This is to mimic the distribution in the RCT. These baseline states are assigned randomly.

We simulated our original RCT with the same level of missing data and degree of correlation by using the estimated parameters as the true value. Then we adjusted the value of α_{0,y_s}, α_{1,y_s}, and σ^2 to simulate another scenario with a higher amount of missing data and a higher degree of within-unit correlation. We use the sample size of 112 as in our RCT data. For each individual, we generate the nonzero number of cysts by using a Poisson distribution with a mean value of 1.87, where the value was obtained from the real data. For each scenario, we produce 100 datasets and then fit the data with the IND, MCAR, and MNAR models for the methods described in the last section.

Table 3 summarizes the simulation results for the transition parameters. We present the average estimates and average percent biases (Bias%) (percentage on the difference between the estimated parameter and the true value relative to the true value) as well as the coverage rate (CR) of the 95% confidence intervals of the parameters. We see that under both simulation scenario, the MNAR joint model produces the most accurate estimate and near the nominal coverage for both the distributional parameters $\beta_{0,rs}$ and the treatment effect parameters $\beta_{1,rs}$. Ignoring the data complexity leads to less accurate results, as seen with the IND model where within-unit clustering is ignored and MCAR model where the missing data is ignored. For example, with 20% missing data and larger magnitude of frailty, $\sigma^2 = 1.6^2$, the IND approach results in more than 10% percent biases for half of the parameters while the coverage can be as low as 27%.

5 | CONCLUSIONS AND DISCUSSION

We have proposed a continuous-time four-state Markov joint model for cyst-level NC life course data that was measured at prescheduled imaging time points and subject to within-brain clustering and informative missingness. By considering several states and related evolutions simultaneously under the multistate model framework and by incorporating the handling of data complexities, we may better understand the evolutionary pathway of NC cysts and how treatment impacts that evolution. Applying our methods to the NC data from the Ecuador RCT, we found that ALB treatment significantly accelerated the evolution from the active phase to the degenerative phase, which had been identified in the patient-level analysis,[8] as well as the evolution from the degenerative phase to disappearance, which had not been definitively determined through patient-level analysis.[9,10] Thus our findings suggest that ALB also has an effect on degenerative cysts, which warrants further research using these cysts-level analytic techniques. We also found that ALB hastened the movement of active cysts to the degenerative stage, and from there the majority of cysts transited to complete resolution, not calcification. Therefore, we infer that ALB does not lead to an increase in calcified cysts. These findings on cyst-level transitions and corresponding estimates are useful for treatment planning for better care of NC patients.

Our joint model can be viewed as an extension of the frailty model in References 19,20 and the selection model in Reference 18 for the multivariate interval-censored event-history data to account for the within-unit clustering and informative missingness. Our results indicate that when incorporating the handling of both data complexities, a larger ALB's treatment effect on the active to degenerative transition and the degenerative to disappearance transition was obtained. The simulation further validated the joint model's performance under finite sample and models that ignore data complexity can lead to severe biases.

TABLE 3 Simulation results

		$\beta_{0,12}$	$\beta_{0,23}$	$\beta_{0,24}$	$\beta_{0,34}$	$\beta_{1,12}$	$\beta_{1,23}$	$\beta_{1,24}$	$\beta_{1,34}$
Model	**True**	**−3.10**	**−4.00**	**−2.50**	**−2.90**	**1.20**	**−0.20**	**0.90**	**−0.60**
		$\sigma^2 = 0.8^2$(10% missing)							
IND	Est	−2.47	−4.20	−2.03	−3.13	0.87	−0.39	0.82	−0.52
	Bias%	−4.20	6.05	−1.19	2.08	−7.18	10.97	−3.18	−5.13
	CR	0.79	0.98	0.64	0.66	0.62	0.99	0.75	0.90
MCAR	Est	−3.34	−4.19	−2.19	−3.10	1.00	−0.31	0.78	−0.55
	Bias%	3.94	5.33	−1.10	2.96	−5.56	6.79	−2.53	−3.49
	CR	0.85	0.92	0.80	0.79	0.83	0.93	0.86	0.92
MNAR	Est	−3.18	−4.12	−2.50	−2.93	1.14	−0.16	0.89	−0.58
	Bias%	2.53	2.99	−0.07	0.95	−3.07	−7.52	−1.41	−2.85
	CR	0.91	0.96	0.89	0.93	0.92	0.96	0.90	0.94
		$\sigma^2 = 1.6^2$ (20% missing)							
IND	Est	−2.61	−4.48	−2.12	−3.09	0.76	−0.34	0.57	−0.40
	Bias%	−3.98	15.86	−1.93	4.96	−14.55	15.62	−10.81	−7.35
	CR	0.57	0.73	0.49	0.63	0.27	0.95	0.42	0.83
MCAR	Est	−3.94	−4.31	−2.15	−3.11	0.94	−0.35	0.63	−0.43
	Bias%	5.67	10.33	−1.93	5.03	−6.47	7.62	−7.77	−6.57
	CR	0.73	0.82	0.77	0.88	0.93	0.89	0.84	0.85
MNAR	Est	−3.21	−4.16	−2.53	−2.97	1.15	−0.10	1.02	−0.60
	Bias%	3.66	3.98	1.05	2.12	−4.42	−10.87	1.90	−0.66
	CR	0.84	0.94	0.82	0.91	0.85	0.94	0.82	0.90

Abbreviations: CR, coverage rate; IND, independent; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

In the model, we have introduced multidimensional frailty that is shared by the transition model and the missing data model. Our results show that the frailty captures the strength of the intra-brain association of cysts progressions but also likely captures the heterogeneity among transitional intensities and among missing data characteristics between participants that are not captured by observed covariates. As majority transitions are evolving toward a final resolution, the degenerative to calcification represents a divergence manifested as a negative correlation for such transition with others in the estimated variance-covariance matrix of the frailty. Inference wise, for the frailty induced integral, which is intractable and has high dimension, we developed a Monte Carlo EM algorithm for the joint likelihood. Information matrix estimation was generated as the by-product, avoiding the hessian-matrix-based approach used by numerical methods such as likelihood approximation.

Our statistical methods are applicable to the study of other diseases that impact various regions of the body using longitudinal designs to assess multiple disease progression endpoints measured at intermittent time points and when there is missing data. Both intermittent missing and missing due to lost to follow-up can be addressed in an integrated approach where different missing data mechanisms can be assessed. When constant hazard is warranted by the data as in our application, an exponential distribution is a straightforward assumption for the baseline hazard to address the left-censoring issue and it is easy to implement. For the cases where time-homogeneous assumption or the Markov assumption are questionable, extension of our model is obviously needed. In that regarding, the more robust piecewise constant hazard methods[20] and semi-Markov models are promising. In this paper, we follow the convention in the literature to assume a multivariate normal distribution for the frailties. The impact of this assumption to the estimates for the joint model will be investigated as the next step. In addition, competing risks models[36] can be explored to address missing data as an alternative.

ACKNOWLEDGEMENTS

This original RCT was supported by the National Institute of Neurological Disorders and Stroke at the National Institutes of Health (R01-NS39403). The analyses presented here were supported by the City University of New York (PSC-CUNY ENHC 47). This work was also partially supported by the CUNY (City University of New York) High Performance Computing Center, College of Staten Island, funded in part by the City and State of New York, CUNY Research Foundation, and National Science Foundation Grants CNS-0958379, CNS-0855217 , and ACI-1126113.

ORCID

Hongbin Zhang https://orcid.org/0000-0002-2156-1005

REFERENCES

1. Gripper LB, Welburn SC. The causal relationship between neurocysticercosis infection and development of epilepsy - a systematic review. *Infect Dis Poverty*. 2017;6:31.
2. Carabin H, Ndimubanzi PC, Budke CM, et al. Clinical manifestations associated with neurocysticercosis: a systematic review. *PLoS Negl Trop Dis*. 2011;5:e1152.
3. WHO. *Assembling a Framework for Intensified Control of Taeniasis and Neurocysticercosis Caused by Taenia Solium*. Geneva, Switzerland: WHO Headquarters; 2014.
4. Escibar A. *The pathology of neurocysticercosis*. In: Palacios E, Rodriquez-Carbajal J, Taveras JM, eds. *Cysticercosis of the Central Nervous System*. Springfield: Thomas, IL; 1983.
5. Carpio A, Placencia M, Santillan F, Escobar A. Neurocysticercosis: an update. *Can J Neurol Sci*. 1994;21:43-47.
6. Garcia HH, Pretell EJ, Gilman RH, et al. A trial of antiparasitic treatment to reduce the rate of seizures due to cerebral cysticercosis. *N Engl J Med*. 2004;350:249-258.
7. Sharma SR, Agarwal A, Kar AM, et al. Evaluation of role of steroid along and with albendazole in patients with epilepsy with single-small enhancing computerized tomography lesions. *Ann Indian Acad Neurol*. 2007;10:39-43.
8. Carpio A, Kelvin EA, Bagiella E. Effects of albendazole treatment on neurocysticercosis: a randomised controlled trial. *J Neurosurg Psych*. 2008;79(9):1050-1055.
9. Singh G, Rajshekhar V, Muethy J, et al. A diagnostic and therapeutic schema for a solitary cysticercus granuloma. *Neurology*. 2010;75:2236-2245.
10. Carpio A, Fleury A, Romo ML, Abraham R. Neurocysticercosis: the good, the bad and the missing. *Expert Rev Neurother*. 2018;18:289-301.
11. White ACJ, Coyle CM, Rajshekhar V. Diagnosis and treatment of neurocysticercosis: 2017 clinical practice guidelines by the Infectious Disease Society of America (IDSA) and American Society of Tropical Medicine and Hygiene (ASTMH). *Clin Infect Dis*. 2018;66:1159-1163.
12. Montgomery MA, Ramos M, Kelvin EA, et al. A longitudinal analysis of albendazole treatment effect on neurocysticercosis cyst evolution using multistate models. *R Soc Tropic Med Hyg*. 2019;113(12):781-788.
13. Meneses LJP, Gonzales I, Pretell EJ. Occasional resolution of multiple parenchymal brain calcifications in patients with neurocysticercosis. *Neurol Clin Pract*. 2015;5:531-533.
14. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. New York, NY: Wiley; 2002.
15. Kalbfleisch J, Lawless JF. The analysis of panel data under a Markov assumption. *J Am Stat Assoc*. 1985;80:863-871.
16. Satten GA, Longini JM. Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease (with discussion). *Appl Stat*. 1996;45:275-309.
17. Jackson CH, Sharples LD, Thompson SG, et al. Multistate Markov models for disease progression with classification error. *Statistician*. 2003;52:193-209.
18. Hout A, Matthews FE. Estimating stroke-free and total life expectancy in the presence of non-ignorable missing values. *J R Stat Soc Ser A*. 2010;173(2):331-349.
19. Pak D, Li C, Todem D. A multistate model for correlated interval-censored life history data in caries research. *J R Stat Soc Ser C (Appl Stat)*. 2017;66:413-423.
20. Pak D, Li C, Todem D. Semiparametric analysis of correlated and interval-censored event-history data. *Stat Methods Med Res*. 2018;0(0):1-14.
21. Lee Y, Nelder JA. Hierarchical generalized linear models. *J R Stat Soc Ser B*. 1996;58:619-678.
22. Satten GA, Sternberg M. Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics*. 1999;55(2):507-513.
23. Wu L. *Mixed Effects Models for Complex Data*. London, UK: Chapman & Hall; 2009.
24. Follmann D, Wu M. An approximiate generalized linear model with random effects for informative missing data. *Biometrics*. 1995;51:15-168.
25. Ten Have TR, Pulkstenis E, Kunselman A, Landis JR. Mixed effects logistics regression models for longitudinal binary response data with informative dropout. *Biometrics*. 1998;54:367-383.
26. Tsiatis AA, Davidian M. An overview of joint modeling of longitudinal and time-to-event data. *Stat Sin*. 2004;14:793-818.
27. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. Hoboken, NJ: John Wiley & Sons; 2002.

28. Wei GC, Tanner MA. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J Am Stat Assoc*. 1990;85:699-704.
29. Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J R Stat Soc Ser B*. 1999;61:173-190.
30. Geweke J. *Handbook of Computational Economics, Ch. 15*. Amsterdam, the Netherlands: North-Holland; 1996.
31. Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J R Stat Soc Ser B*. 1999;61:265-285.
32. Louis TA. Finding the observed information matrix when using the EM algorithm. *J R Stat Soc Ser B*. 1982;44:226-233.
33. McLachlan GJ, Krishnan T. *The EM-Algorithm and Extension*. New York, NY: Wiley; 1997.
34. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 1977;25:2340-2361.
35. Norris JR. *Markov chains*. Cambridge, UK: Cambridge University Press; 1997.
36. Larson MG, Dinse GE. A mixture model for the regression analysis of competing risks data. *J R Stat Soc Ser C Appl Stat*. 1985;34(3):201-211.
37. Gilks WA, Wild P. Adaptive rejection sampling for Gibbs sampling. *Appl Stat*. 1992;41:337-348.
38. Hout A. *Multi-state Survival Models for Interval-Censored Data*. Boca Raton, FL: CRC Press; 2017.

How to cite this article: Zhang H, Kelvin EA, Carpio A, Allen Hauser W. A multistate joint model for interval-censored event-history data subject to within-unit clustering and informative missingness, with application to neurocysticercosis research. *Statistics in Medicine*. 2020;1–12. https://doi.org/10.1002/sim.8663

APPENDIX A.

Multivariate rejection algorithm

Sampling from the conditional distribution of the random effects can be accomplished by a multivariate rejection algorithm. If the density functions are log-concave in the appropriate parameters, the adaptive rejection algorithm of Reference 37 may be used, as in Reference 29. However, for the multistate joint model, some densities may not be log-concave. In such cases, the multivariate rejection sampling method[30] may be used to obtain the desired samples[31] discussed such a method in the context of complete data generalized linear mixed models, which can be extended to our models as follows.

Consider sampling from $f(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)})$. Let $f^*(\boldsymbol{\omega}_i) = f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta}^{(v)})$, and $\xi = \sup_{\mathbf{u}}\{f^*(\mathbf{u})\}$. A random sample from $f(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)})$ can be obtained as follows:

Step 1: sample $\boldsymbol{\omega}_i^*$ from $f(\boldsymbol{\omega}_i; \boldsymbol{\theta}^{(v)})$, and independently, sample w from the uniform(0,1) distribution;

Step 2: if $w \leq f^*(\boldsymbol{\omega}_i^*)/\xi$, then accept $\boldsymbol{\omega}_i^*$ as a sample point from $f(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(v)})$, otherwise, go back to step 1 and continue.

Transition probabilities

The transition probabilities $P(y_{ij,t_k}|y_{ij,t_{k-1}}, \boldsymbol{x}_i, \boldsymbol{\omega}_i)$ involved in $L^c_{ij}(\boldsymbol{y}^c_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\omega}_i)$ from Equation (2) can be written as a function $p_{y_{ij,t_{k-1}}\, y_{ij,t_k}}(t_{k-1}, t_k)$, which can be computed using $q_{rs}(t|\boldsymbol{x}_i, \boldsymbol{\omega}_i)$ as follows (suppressing $\boldsymbol{x}_i, \boldsymbol{\omega}_i$) for the irreversible disease progression model in Figure 1 (see also Reference 38):

$$
\begin{aligned}
p_{11}(t_1, t_2) &= \exp(-H_1(t_1, t_2)) \\
p_{12}(t_1, t_2) &= \int_{t_1}^{t_2} \exp(-H_1(t_1, t))q_{12}(t)\exp(-H_2(t, t_2))dt \\
p_{13}(t_1, t_2) &= \int_{t_1}^{t_2}\int_{u}^{t_2} \exp(-H_1(t_1, u))q_{12}(u)\exp(-H_2(u, v))q_{23}(v)\exp(-H_3(v, t_2))dvdu \\
p_{14}(t_1, t_2) &= 1 - p_{11}(t_1, t_2) - p_{12}(t_1, t_2) - p_{13}(t_1, t_2) \\
p_{22}(t_1, t_2) &= \exp(-H_2(t_1, t_2)) \\
p_{23}(t_1, t_2) &= \int_{t_1}^{t_2} \exp(-H_2(t_1, t))q_{23}(t)\exp(-H_3(t, t_2))dt \\
p_{24}(t_1, t_2) &= 1 - p_{22}(t_1, t_2) - p_{23}(t_1, t_2) \\
p_{33}(t_1, t_2) &= \exp(-H_3(t_1, t_2)) \\
p_{34}(t_1, t_2) &= 1 - p_{33}(t_1, t_2)
\end{aligned}
$$

where $H_1(t_1, t_2) \equiv \int_{t_1}^{t_2} q_{12}(t)dt$, $H_2(t_1, t_2) \equiv \int_{t_1}^{t_2} q_{23}(t) + q_{24}(t)dt$, and $H_3(t_1, t_2) \equiv \int_{t_1}^{t_2} q_{34}(t)dt$ are the cumulative hazard functions for leaving state 1, 2 and 3, respectively.