Proceedings of the 2nd International Conference on Statistics: Theory and Applications (ICSTA'20)
~~*Prague, Czech Republic*~~ *Virtual Conference – August, 2020*
Paper No. ICSTA 105
DOI: 10.11159/icsta20.105

A MCEM based multistate model of interval-censored and correlated panel data for neurocysticercosis study

Hongbin Zhang[1], Elizabeth Kelvin[1], Arturo Carpio[2], Allen Hauser[3]
[1]Department of Epidemiology and Biostatistics/Institute of Implementation Science for Population Health/City University of New York
55 West 125th Street, New York, USA
hongbin.zhang@sph.cuny.edu; elizabeeth.kelvin@sph.cuny.edu
[2]School of Medicine/University of Cuenca/Ecuador
[3]Mailman School of Public Health/Columbia University/USA

Abstract - We propose a multistate model to analyse interval-censored event-history data subject to within-unit clustering. The model is motivated by a study of the neurocysticercosis evolution at cyst-level, considering the multiple cysts phases within the brain at pre-scheduled imaging time points. Of interest is the study of the intra-brain distribution of the process leading to cyst resolution, and whether this distribution varies with the anthelmintic treatment. We develop a likelihood-based method using Monto Carlo EM algorithm for the inference. The practical utility of the methods is illustrated using data from a longitudinal study on albendazole's therapeutic effect among patients from six hospitals in Ecuador.

Keywords**:** neurocysticercosis, multistate model, frailty survival model, correlated disease progressions, interval-censoring, MCEM algorithm.

1. Introduction

Neurocysticercosis (NC) is an infection of the central nervous system with the larval form of the pork tapeworm. When located in the human brain, the larval stage of T. solium appears to pass through three distinct stages of evolution before total disappearance [1]. In the first stage, the parasite is viable or alive. 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 sometimes remains in its place; this is termed the "calcified" phase.

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 trial conducted in Ecuador assessing the effectiveness of ALB for newly diagnosed NC patients [2]. Brian image on cyst type and location were obtained over five time points (at baseline and 1, 6, 12, and 24 months). We characterize the NC evolution with three transient states (active, degenerative, calcification) and one absorbing state (disappearance), as seen in Figure 1.

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, the exact times of cyst transition to a new phase are known only within an interval. Additionally, when multiple NC cysts evolutions were abstracted from the same patient, those evolutions were correlated. Moreover, our data also suffer the so-called left-censoring issue where the onset time of active or degenerative cysts is unknown when patients are recruited into the study. Several researchers have considered multistate models for interval-censored event-history data, see, for example, [3] and [4]. For intra-subject correlated data, [5] extended frailty survival model for interval-censored data and used approximated likelihood methods, e.g., Gaussian-quadrature, for the inference.

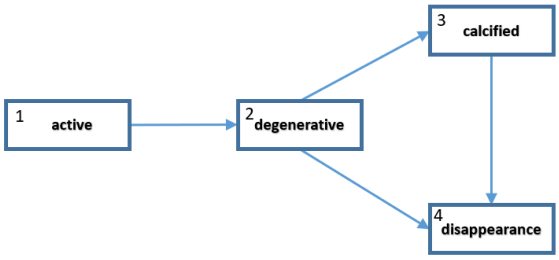

Fig. 1: Four states transition multistate model for NC.

We identified some common strategies in the literature to deal with the left-censoring problem. For example, [3] assume that the holding time of the initial state is exponentially distributed, rendering the time origin unnecessary due to the memory-less 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-individual correlation. Our data also present missing states and for simplicity, we assume the missing is at random, following the standard handling of missing data in literature. 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 likelihood-approximation methods in literature can be computationally very intensive and may even offer convergence problems. For the inference of the multistate model in this paper, we use the Monte Carlo EM (MCEM) algorithm, which is more stable and provides ``exact'' likelihood-based estimation [6]. MCEM algorithm involves multivariate Monte Carlo sampling with rejection, which is intrinsically challenging for interval-censored data. Incorporating the handling of correlated data lead to additional computational complexity for the implementation.

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

2. Statistical Methods

2.1. 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 time t, $t \geq 0$. The transition intensity from state r to state s at time t, defined as instantaneous probability, $lim_{\delta_t \to 0} P\left(y_{ij,t+\delta_t} = s | y_{ij,t} = r,\ x_i, \tau_i\right)$ is modelled by

$$q_{rs}(t|x_i, \tau_i) = q_{rs}^{(0)}(t)\exp(\beta_{rs}^T x_i + \tau_i) \tag{1}$$

where $q_{rs}^{(0)}(t)$ is the transition-specific baseline intensity function which is assumed to follow Exponential distribution, β_{rs}^T are the transition specific regression coefficients, x_i is the covariate vector (e.g., treatment, demographic variables), and τ_i is the individual random effect (aka frailty) which is introduced to account for the intral-brain clustering. The frailty can be flexibly specified for each type of transition and is assumed to follow a multivariate normal distribution.

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

$$L_{ij}^{c}(y_{ij}^{c}) = P(y_{ij,t_1}) \prod_{k=2}^{m} [P(y_{ij,t_k}|y_{ij,t_{k-1}}, x_i, \tau_i) P(r_{ij,t_k}|y_{ij,t_k}, x_i, \tau_i)] \quad (2)$$

where $[P(y_{ij,t_k}|y_{ij,t_{k-1}}, x_i, \tau_i)$ 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 $P(r_{ij,t_k}|y_{ij,t_k}, x_i, \tau_i)$ in (2) represents the selection-model approach for non-ignorable missingness [7]. When missing is assumed to be random, this quantity is constant [8]. Note that the same modelling 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, dissolved, 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 dispersion parameter for the frailty. Under the assumption that the cyst evolution processes are independent across subjects, the log-likelihood of the observed data $\boldsymbol{Y}$ is

$$l(\mathbf{Y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} \left[\sum_{\boldsymbol{y}_{ij}^{c} \in \Omega(\boldsymbol{y}_{ij})} L_{ij}^{c}(y_{ij}^{c}|x_i, \tau_i; \boldsymbol{\theta}) \right] \boldsymbol{f}(\boldsymbol{\tau_i}; \boldsymbol{\theta}) \boldsymbol{d\tau_i} \right) \quad (3)$$

where the integral can be multi-dimension, $\boldsymbol{y_{ij}}$ 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 4-state survival model, only patterns with monotone increase are possible. For example, if $\mathbf{y_{ij}} = (1, 1, 3, \bullet, \bullet)$ where the • represents missed state, i.e., 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

For our data, the key implementation difficulty is that the integral in the likelihood (3) which is typically quite intractable, due to the interval-censoring and non-ignorable missing induced structural complexity. When the dimension of the random effects is not low, numerical methods such as Gaussian quadrature can be computationally very intensive and may offer non-convergence. We therefore offer 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 extremely complicated, Monte Carlo methods can be used to approximate the expectation, leading to a MCEM algorithm [9]. As Monte Carlo sample size is increasing, the approximation can have high accuracy to the true expectation [10]. In our case, by treating the random effects $\boldsymbol{\tau_i}$ as additional ``missing data", we have ``complete data" (y_i, x_i, τ_i)--- where y_i represents the observed cyst profile for individual i --- and the ``complete-data" log-likelihood function for individual *i* can be expressed as

$$l_c^i(\boldsymbol{\theta}) = \log f(y_i|x_i, \tau_i; \boldsymbol{\theta}) + log\ f(\tau_i; \boldsymbol{\theta}) \quad (4)$$

where $f(y_i|x_i, \tau_i; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} \left[\sum_{\boldsymbol{y}_{ij}^{c} \in \Omega(\boldsymbol{y}_{ij})} L_{ij}^{c}(y_{ij}^{c}|x_i, \tau_i; \boldsymbol{\theta}) \right]$.

The EM algorithm iterates between an E-step which calculates the likelihood of missing data given the observed data and a M-step that maximizes the likelihood until convergence. To obtain the asymptotic variance-covariance matrix of the MLE parameters, we consider the formula in [11] which is a by-produce of Monte Carlo sampling.

3. Data Analysis

Between 2001 and 2005, researchers at Columbia University, together with physicians at six hospitals in Ecuador, conducted a multi-site randomized clinical trial where a total of 178 patients with newly diagnosed NC were recruited [2]. 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 CT or MRI image of the brain. Patients were randomly assigned to receive either albendazole (ALB) or placebo, given twice a day orally for eight days. At enrolment, patients were interviewed to collect information about demographics, symptoms, and risk factors for NC. A brain scan was taken at baseline, 1, 6, and 12 months of follow-up.

We recently acquired the 24-month image data, digitized, and integrated it with the trial data. We then dis-aggregated data from the patient to the cyst level and generated a cyst-level dataset with a total of 210 cysts from 112 patients. Among

the sample, 59 (52\%) individuals were treated with ALB. The average number of cysts per patient is 1.87, ranging from 1 to 8. Over the study period, five patients were loss to follow-up between 6-month and 12-month, and 25 patients had 12-months imaging done but did not contribute to the 24-months imaging. We excluded patients who deceased in the study to allow plausible data augment to the latent states. The objective of our data analysis is to assess the therapeutic effect of ALB on the NC cysts evolution, incorporating the handling of data issues.

Table 1: Frequencies of observed transitions for NC data.

	ALB					Placebo			
	act	deg	cal	dis		act	deg	cal	dis
act	36	9	2	47	act	58	8	0	24
deg	0	32	3	30	deg	0	52	5	32
cal	0	0	42	20	cal	0	0	28	14

Table 1 presents frequencies of observed transitions between the states in the NC data stratified by the treatment group. Note that the cases of ``reverse transitions", which are mostly related to imaging reading errors or due to new infection at the same location, were removed from the data. There are 10 intermittent missing cases and 30 patients were lost to follow-up.

There are a variety of reasons patients did not have their imaging done. If an individual misses imaging for unrelated reasons, then that is not a violation of model assumptions. Our preliminary analysis shows no association between the missing status and the treatment assignment, we therefore, fit the following model for the NC data.

$$\log(q_{rs}(x_i, \tau_i)) = \beta_{0,rs} + \beta_{1,rs} trt_i + \tau_i, \quad (4)$$

where the quantities $q_{rs}(x_i, \tau_i)$ is defined in model (1), trt_i is the treatment assignment for individual with a value of ``1" for ALB and ``0" for placebo. Note that the random effects τ_i are included in all the models to account for the within-unit clustering.

We apply our method on NC data. Besides our model, denoted as MCEM, we also fit another model, denoted as IND where we assume no within-individual correlation. We apply the algorithm described in Section 2 (implemented in R) for the MCEM model while for IND, R's package *msm* is used [12]. For the random effects, we start with an identity variance-covariance matrix and allow the algorithm to estimate the full structured variance-covariance parameters.

Table 2: Results of multistate model fitting, e.g., Est(se), for NC data.

	$\boldsymbol{\beta_{0,12}}$	$\boldsymbol{\beta_{0,23}}$	$\boldsymbol{\beta_{0,24}}$	$\boldsymbol{\beta_{0,34}}$	$\boldsymbol{\beta_{1,12}}$	$\boldsymbol{\beta_{1,23}}$	$\boldsymbol{\beta_{1,24}}$	$\boldsymbol{\beta_{1,34}}$
IND	-2.42(0.11)	-4.44(0.26)	-2.03(0.10)	-3.13(0.15)	0.96(0.21)	0.12(0.62)	0.64(0.21)	-0.03(0.30)
MCEM	-3.27(0.27)	-4.13(0.48)	-2.51(0.31)	-2.89(0.39)	1.21(0.32)	-0.15(0.73)	0.78(0.26)	-0.52(0.41)

Table 2 presents the results. The MCEM model tends to report larger standard error which is likely attributed to the incorporation of the clustering feature of the data. Both models report significant distributional parameters. Also, significant treatment effect are detected on transition from ``active" to ``degeneration," $\beta_{1,12}$ and on the transition from ``degenerative" to ``disappearance," $\beta_{1,24}$. However, the estimates from the `MCEM model shows larger differential magnitude than the one assuming independent data. In particular, the treatment effect on the transition from the ``active" stage to the ``degenerative" stage is estimated to be 1.21 ($\beta_{1,12}$) with standard error (se) 0.32 under MCEM while the estimate from IND is smaller, 0.96, se = 0.21. On the other hand, MCEM's estimate for effect on the transition from ``degenerative" to ``disappearance" is 0.78 ($\beta_{1,24}$), with standard error 0.26 while the estimate and standard error for this effect is 0.64 and 0.21, respectively, under model IND.

4. Conclusion

We have proposed a 4-state Markov joint model for cyst-level NC life course data that is measured over pre-scheduled imaging time points and subject to within-brain clustering. By considering several states and related evolution 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. When applying our methods to the NC data from Ecuador, we found that the treatment significantly accelerated the evolution from the active phase to the degenerative phase and the evolution from the degenerative phase to disappearance. While it is well-known that ALB kills viable cysts, our finding suggests that ALB also may have effects on degenerative cysts and that warrants further research. With regards to the impact of ALB on calcification, we found that ALB hastened the move of active cysts to the degenerative stage and from there most cysts transited into disappearance, not calcification. Therefore, we infer that ALB does not lead to calcification. These findings on cyst-level transitions and corresponding estimates are useful for treatment planning for better care of NC patients.

Our statistical methods are generally applicable to cohort studies and medical records involving chronic disease in which the disease progressions are assessed at intermittent time points from various locations of the body. 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. As convention, we assume missing at random for the loss to follow-up. Our next step is to investigate the approach to incorporate non-ignorable missingness where the mixture modelling approach [13] is promising. We will also on seek possible solution to relax the time-homogeneous assumption and the Markov assumption.

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 Centre, 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.

References

[1] A. Carpio and M. Placencia and F. Santillan and A. Escobar, "A proposal for classification of neurocysticercosis," Can *J Neurol Sci.*, vol. 21, pp. 43-47, 1994.

[2] A. Carpio and E. A. Kelvin and E. Bagiella, "Effects of albendazole treatment on neurocysticercosis: a randomised controlled trial," *J Neurosurg Psychiatry*, vol. 79, no. 9, pp. 1050-1055, 2008.

[3] J. Kalbfleisch and J. F. Lawless, "The analysis of panel data under a Markov assumption," *J. Am. Statist. Ass*., vol. 80, pp. 863-871, 1985.

[4] C. H. Jackson and L. D. Sharples and S. G. Thompson, "Multistate Markov models for disease progression with classification error," *Statistician*, vol. 52, pp. 193-209, 2003.

[5] D. Pak and C. Li and D. Todem, "A multistate model for correlated interval-censored life history data in caries research," *J Royal Stat Soc:-Ser C (Appl Stat)*, vol. 66, pp. 413-423, 2017.

[6] L. Wu, *Mixed Effects Models for Complex Data*, Chapman and Hall, London, 2010.

[7] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd Edition, Wiley, New York, 2002.

[8] A. Hout and F. E. Matthews, "Estimating stroke-free and total life expectancy in the presence of non-ignorable missing values," *J Royal Stat Soc:-Ser A*, vol. 173, no. 2, pp. 331-349, 2010.

[9] G. C. Wei and M. A. Tanner, "A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm,", *J. Am. Statist. Ass*., vol. 85, pp. 699-704, 1990.

[10] L. Wu, "Exact and approximate inferences for nonlinear mixed-effects models with missing covariates," *J. Am. Statist. Ass*., vol. 99, pp. 700-709, 2004.

[11] G. J. McKachlan and T. Krishnan, *The EM-Algorithm and Extension*, Wiley, New York, 1997.

[12] C. H. Jackson, "Multi-State Models for Panel Data: The msm Package for R," *Journal of Statistical Software*, Vol. 38, no. 8, 1-28.

[13] M. G. Larson and G. E. Dines, “A Mixture Model for the Regression Analysis of Competing Risks Data,” *J Royal Stat Soc:-Ser C (Appl Stat),* Vol. 34, no. 3, pp. 201-211,1985.