A two-part mixed-effects model for analyzing clustered time-to-event data with clumping at zero

Abstract:

Background and Objective: In longitudinal epidemiological studies consisting of a baseline stage and a follow-up stage, observations at the baseline stage may contain a countable proportion of negative responses. The time-to-event outcomes of those observations corresponding to negative responses at baseline can be denoted as zeros, which are excluded from standard survival analysis. Consequently, some important information on these subjects is therefore lost in the analysis. Furthermore, subjects are often clustered within hospitals, communities or health service centers, resulting in correlated observations. The framework of the two-part model has been developed and utilized widely to analyze semi-continuous data or count data with excess zeros, but its application to clustered time-to-event data with clumping at zero remains sparse.

Methods: A two-part mixed-effects modeling approach was proposed. A logistic mixed-effects regression model was used in the first part to determine factors associated with the prevalence of the baseline event of interest. Parametric frailty models (including Weibull, exponential, log-logistic and log-normal) were used in the second part to assess associations between exposures and time-to-event outcomes. Correlated random effects were incorporated within the two regression models to accommodate the inherent correlation within each clustering unit and the correlation between the two parts. As an illustrative example, the method was applied to exclusive breastfeeding data from a community-based prospective cohort study in Nepal.

Results: A significantly positive correlation between the baseline prevalence of exclusive breastfeeding and exclusive breastfeeding duration was confirmed ($\rho = 0.67$, $P<0.001$). The

correlated two-part model outperformed the independent two-part model (likelihood ratio test statistic=8.6, df=1, *P*=0.003).

Conclusions: The proposed approach makes full use of all available information at baseline and during the follow-up, compared to the conventional survival analysis. In addition to breastfeeding studies, the method can be applied to other research areas where clustered time-to-event data with clumping at zero arise.

Keywords: Clumping at zero; Frailty model; Mixed-effects; Time-to-event data; Two-part model

1. Introduction

In longitudinal epidemiological studies consisting of a baseline stage and a follow-up stage, observations at the baseline stage may contain a countable proportion of negative responses (i.e., "No" or "False" to a research question regarding a baseline event of interest). In the follow-up stage, only subjects with positive responses (i.e., "Yes" or "True" to the same research question regarding the same baseline event of interest) at baseline are continuously followed up to measure the time (duration) up to the occurrence of a failure event of interest, as a consequence the time to the failure event occurrence for those unfollowed-up subjects is denoted as zero. A motivating example is the longitudinal data arising from breastfeeding studies, which normally have a baseline stage for measuring the prevalence of breastfeeding at discharge and a followed-up stage for measuring the duration of breastfeeding. Breastfeeding, especially exclusive breastfeeding (EBF), is beneficial to both infants and mothers.[1-4] The World Health Organization has recommended EBF for at least 6 months.[5] However, the prevalence of EBF at hospital discharge varies globally and remains low in many countries, for example, 75.6% in Australia,[6] 50.3% in China[7] and 68.6% in Spain.[8] In other words, there exist a high proportion of non-exclusively breastfed infants, corresponding to those "no" responses to the status of EBF, at discharge (baseline). Their EBF duration (i.e., time to cessation of EBF) would be noted as zero.

These two-stage studies produce two processes of outcome data, namely, a binary outcome to describe the prevalence of the event of interest at baseline and a time-to-event outcome measuring the duration up to the failure event occurring in the follow-up. Statistical analyses, including logistic regression and survival analysis, for identifying factors associated with the prevalence of the baseline event, and with the time to the

failure event occurrence, are widely performed in literature. It is known that standard survival analysis considers only positive time-to-event outcomes so that subjects with a negative response at baseline are excluded from the risk set for estimating the survival probability (morbidity or mortality in some instances). Consequently, the information from these subjects, which may be important to the failure event of interest, is abandoned in the analysis. Given the two-stage data structure, in the literature, a framework of two-part (or two-stage) model has been introduced to analyze outcomes with a two-component structure, such as semi-continuous data[9,10] and count data with excess zeros[11,12] with applications in physical activity,[13] healthcare cost,[14,15] alcohol use,[16] and household debt.[17] However, to our best knowledge, its application to time-to-event data with clumping at zero is limited to the lapse of insurance and duration analysis in political science.[18,19] Furthermore, in epidemiological settings, subjects are often clustered within hospitals, communities, or health service centers. It is expected that observations exhibit intra-cluster correlation due to similar socio-economic, environmental or health conditions for individuals in the same cluster. To account for the heterogeneity between clustering units, an adjustment for the underlying correlation structure via random effects in a two-part model becomes necessary.

In addition, in some cases it is reasonable to assume the binary data and time-to-event data generated from the two processes are related to each other, indicating an inherent correlation between the two stages. In breastfeeding studies, for example, mothers in one community which has a higher prevalence of EBF at hospital discharge are more likely to have longer EBF durations compared to those in other communities. Ignoring such correlation may introduce bias in statistical inferences. Therefore, we proposed a two-part mixed-effects (fixed and random effects) modeling approach to analyze

clustered time-to-event data with clumping at zero, with application to EBF data as an illustrative example. Correlated random effects accounting for possible correlation between baseline and follow-up stages were considered in this approach.

2. Methods

2.1. Two-part mixed-effects model

This paper considers that longitudinal observations can be partitioned structurally into two parts. The first part is a baseline part (Part 1), where subjects' statuses regarding a 'baseline event' are observed. Examples of positive responses to the 'baseline event' include "EBF at discharge" or "hold a health insurance at baseline". The second part is a follow-up part (Part 2), where positive time to a 'failure event' of interest, conditional on the positive baseline response in the first part, is observed. Examples of the 'failure event' could be "EBF cessation" or "health insurance lapse".

Suppose the observations from a longitudinal epidemiological study are given by $(y_{ij}, \delta_{ij}),\ j = 1, \cdots, n_i, i = 1, \cdots, n,\ \mathrm{N} = \sum_{i=1}^{n} n_i,$ where y_{ij} is the observed time to event for the j th subject within the i th cluster, and δ_{ij} is the censoring indicator for the event. For each subject $j = 1, 2, \ldots, n_i$ within a cluster $i = 1, 2, \ldots, n$, let a binary variable $Z_{ij} = 1$ corresponding to the positive response to the baseline event and $Z_{ij} = 0$ otherwise in Part 1. The likelihood function is then given by

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}) &= \prod_{(i,j):Z_{ij}=1} p_{ij} f(y_{ij}, \xi_{ij}, \boldsymbol{\beta}) \prod_{(i,j):Z_{ij}=0} (1 - p_{ij}) \\ &= [\prod_{(i,j):Z_{ij}=1} p_{ij} \prod_{(i,j):Z_{ij}=0} (1 - p_{ij})][\prod_{(i,j):Z_{ij}=1} f(y_{ij}, \xi_{ij}, \boldsymbol{\beta})], \end{aligned} \tag{1}$$

where $p_{ij} = p(\eta_{ij}, \boldsymbol{\gamma}) = p(Z_{ij} = 1 | \eta_{ij})$ is the probability of the baseline event occurrence for the j th subject within the i th cluster for given covariates x_{ij} via $\eta_{ij} = x_{ij}'\boldsymbol{\gamma} + u_i$, with u_i being the random effects for adjusting the subject-level correlation at baseline and the parameter vector $\boldsymbol{\gamma}$ represents the effects of x_{ij} on the binary outcome; $f(y_{ij}, \xi_{ij}, \boldsymbol{\beta})$ denotes the probability density of the observed time to the failure event y_{ij} for covariates w_{ij} via $\xi_{ij} = w_{ij}'\boldsymbol{\beta} + v_i$, with v_i being the random effects for adjusting the subject-level correlation in the follow-up and $\boldsymbol{\beta}$ is parameter vector associated with w_{ij}. The likelihood function $L(\boldsymbol{\gamma}, \boldsymbol{\beta})$ can be factorized into two parts as

$$L_1(\boldsymbol{\gamma}) = \prod_{(i,j):Z_{ij}=1} p_{ij} \prod_{(i,j):Z_{ij}=0} (1 - p_{ij}) \tag{2}$$

and

$$L_2(\boldsymbol{\beta}) = \prod_{(i,j):Z_{ij}=1} f(y_{ij}, \xi_{ij}, \boldsymbol{\beta}) \tag{3}$$

Here, this two-part mixed-effects model aims to: (i) determine factors associated with the prevalence of the baseline event of interest; (ii) assess associations between exposures and time-to-event outcomes; (iii) capture/ account for the possible correlation between the prevalence part and the time-to-event part.

2.1.1. Part 1: logistic mixed-effects regression model

For the first part $L_1(\boldsymbol{\gamma})$, a logistic mixed-effects regression model can be applied to achieve the first objective via:

$$\log(p_{ij}/(1-p_{ij})) = \eta_{ij} = x'_{ij}\boldsymbol{\gamma} + u_i \tag{4}$$

2.1.2. Part 2: parametric frailty model

For the second part $L_2(\boldsymbol{\beta})$, either a conditional semi-parametric Cox proportional hazards model with random effects (i.e., Cox frailty model) or a class of conditional parametric survival models incorporating random effects (i.e., parametric frailty models) can be used. However, compared to semi-parametric proportional hazard model (i.e., Cox proportional hazard model), parametric survival models provide more efficient and informative estimations when the baseline hazard function could be specified in advance.[20,21] This paper focuses on parametric frailty models in Part 2 modeling. The resulting log-likelihood function is given by

$$\begin{aligned} l_2(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{(1-\delta_{ij})\log S(y_{ij};\xi_{ij},\boldsymbol{\beta}) + \delta_{ij}\log f(y_{ij};\xi_{ij},\boldsymbol{\beta})\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{\log S(y_{ij};\xi_{ij},\boldsymbol{\beta}) + \delta_{ij}\log h(y_{ij};\xi_{ij},\boldsymbol{\beta})\}, \end{aligned} \tag{5}$$

where $S(y_{ij};\xi_{ij},\boldsymbol{\beta})$ is the survival function, and $h(y_{ij};\xi_{ij},\boldsymbol{\beta})$ is the hazard function, $\xi_{ij} = w'_{ij}\boldsymbol{\beta} + v_i$. The hazard function can be one of the various forms depending on different distributions of the time-to-event outcome y_{ij}, including Weibull, exponential, log-logistic and log-normal. In practice, gamma and lognormal are commonly used distributions for the frailty term (exponential transformation of the random component) in the above frailty model.[22]

2.1.3. Assumptions on random effects u_i and v_i

For the relationship between random effects u_i and v_i in Part 1and Part 2, the study considers the following two assumptions:

1) If u_i and v_i are assumed to be independent, following a normal distribution $N(0,\sigma_u^2)$ and $N(0,\sigma_v^2)$, respectively, the approach is an independent two-part mixed-effects modeling.

2) If u_i and v_i are assumed to be correlated, following a bivariate normal distribution with a variance-covariance matrix $\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}\right)$, the approach is a correlated two-part mixed-effects modeling.

Practically the likelihood ratio test can be used to assess and compare the goodness-of-fit between the independent and the correlated two-part models.

2.1.4. Parameter estimation

When the random effects u_i and v_i are independent and note that the likelihood function $L(\boldsymbol{\gamma},\boldsymbol{\beta})$ allows for different sets of covariates in both parts, it is computationally feasible and fully efficient to fit the two regression models separately. When the random effects u_i and v_i are assumed to be correlated, the Laplace approximation[9] or adaptive Gaussian quadrature,[10,15] which has been utilized to handle correlated two-part random effects for modeling semi-continuous data, can be adapted. Under either assumption, model fitting and parameters estimation can be conveniently implemented by using the adaptive Gaussian quadrature technique available in Proc

NLMIXED of SAS (SAS Institute Inc., Cary, NC, USA). Appendix provides the SAS codes used in our illustrative example.

Figure 1 presents the two-part mixed-effects modeling process.

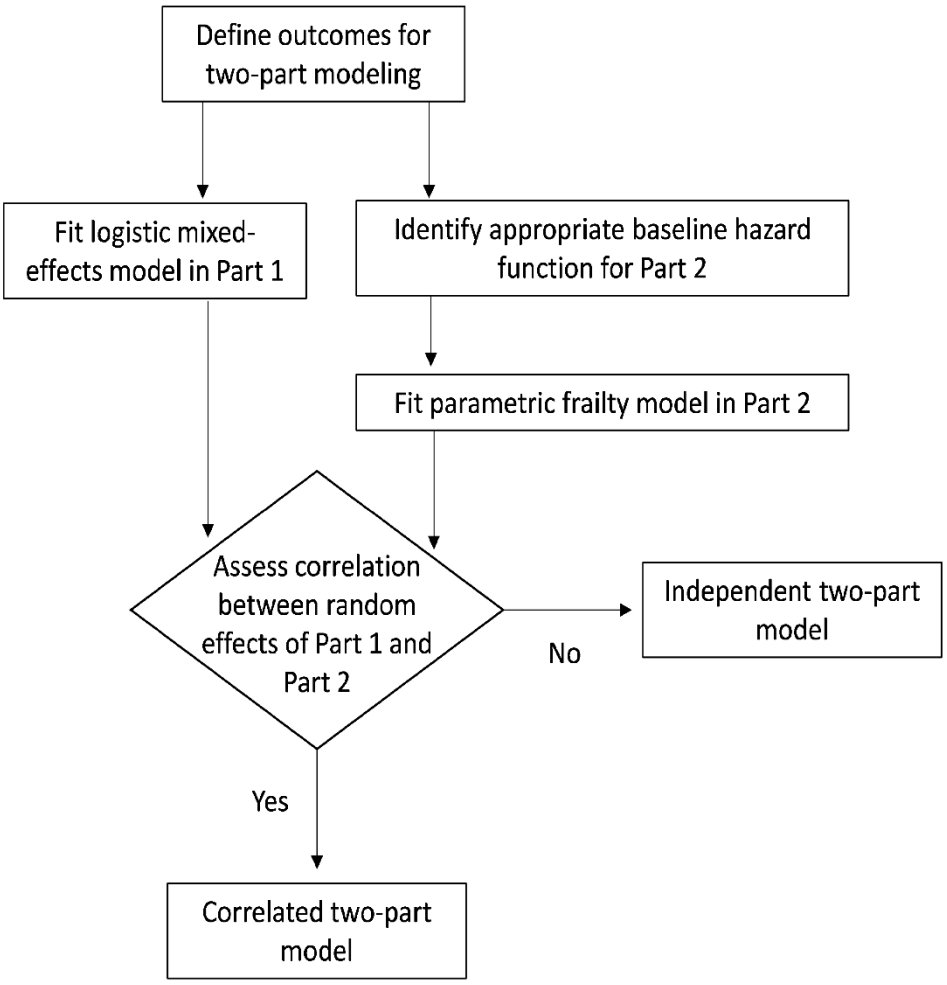

Figure 1. Flow chart of the two-part mixed-effects modeling process

2.2. Illustrative example

2.2.1. Breastfeeding data

A community-based prospective cohort study was conducted in Nepal between January and October 2014 to determine factors associated with breastfeeding duration, involving 27 randomly selected communities (15 village development committees and 12 wards of two municipalities). Details of the study design, setting and sampling had been reported previously.[23] Briefly, a total of 735 mothers were recruited by a proportionate sampling scheme from the target population (i.e., infants<1-year old) and interviewed shortly after giving birth, and those mothers who confirmed they exclusively breastfed their infants at discharge (baseline) were followed up for measuring the EBF duration by six months postpartum. Information about the EBF duration (i.e., time to EBF cessation) was collected for the mothers who were exclusively breastfeeding their babies at baseline. The final sample (N=649) excluded 86 mothers who delivered at home.

2.2.2. Two-part modeling procedure

2.2.2.1. Outcome variables

The status of EBF at baseline was coded as a binary outcome, with '1' for EBF and '0' for non-EBF and used as the outcome variable in Part 1 logistic mixed-effects regression model. For those mothers who exclusively breastfed their infants at the baseline interview, the EBF duration (days) up to 6 months postpartum obtained in the follow-up was used as the outcome variable in Part 2 parametric frailty model. Three covariates, namely, birth mode (caesarean section vs natural birth (reference group)), grandmother feeding preference (breastfeeding vs other feeding (reference group)), and mother-child bonding (yes vs no (reference group)), were included in both parts of the model.

2.2.2.2. Identifying an appropriate baseline hazard function

For Part 2 parametric frailty model, an appropriate form of the baseline hazard function is needed to be specified. With no other suggestive information available from the data, we fitted the data with four commonly used event time distributions, namely, Weibull, exponential, log-logistic and log-normal, without any covariates, then compared their Akaike information criterion (AIC), Bayesian information criterion (BIC) and -2log-likelihood values. The one with the smallest value for these model selection criteria was chosen as the baseline hazard function. Cumulative hazard was estimated with Fleming-Harrington method from EBF duration data and visually compared with fitted hazards from alternative parametric models.

2.2.2.3. Assessing correlation between two stages of random effects

The breastfeeding data were fitted with a logistic mixed-effects model and a Weibull accelerated failure time (AFT) frailty model initially assuming independent random effect u_i and v_i. Empirical Bayes estimates of the random effects were plotted against each other to examine the degree of correlation between the two random effects. The Pearson's correlation coefficient between the estimated u_i and v_i $(i = 1, 2, \ldots, 27)$ was calculated as well.

2.2.2.4. Fitting data with a two-part mixed-effects model incorporating correlated random intercepts

The adaptive Gaussian quadrature technique was used for parameter estimation via the SAS NLMIXED procedure. Perhaps due to data scaling issues,[24] the estimation procedure failed to converge with this dataset. We also suspected that unequal numbers

of observations used in the correlated two-part modeling may cause some problems related to a non-invertible variance-covariance matrix during parameter estimation. As a computational solution, to ensure the two-part regression models having the same number of observations, each zero EBF duration observation was transformed by adding a very small positive value (e.g., 1E-6), and was then included as a censored observation in Part 2 parametric frailty modeling.

The likelihood ratio test was used to assess whether the correlated two-part model provided a better fit to the data. Finally, sensitivity analyses were performed to evaluate the robustness of the results after transformation of zero observations in Part 2 parametric frailty modeling. The parameter estimation and goodness-of-fit were compared between models excluding zero observations and including transformed zero observations.

3. Results

Among those 649 mothers, 434 mothers (66.9%) exclusively breastfed their infants at baseline, leading to 33.1% 'zero' EBF duration observations. Conditional on those mothers who practiced EBF at baseline, only 434 mothers were followed up to six months postpartum for measuring the EBF duration. In Part 2, as shown in Table 1, the Weibull distribution, which had the smallest AIC, BIC and -2log-likelihood values compared with the exponential, log-logistic and log-normal distributions, was chosen as the most appropriate distribution for the EBF duration data. The same model selection decision was evidenced by the visual inspection in Figure 2. As shown in Figure 3, a positive correlation (r=0.55, P=0.003) was found between the random effects u_i and v_i,

suggesting an inherent correlation between the occurrence of EBF for mothers after giving birth and time to EBF cessation by 6 months postpartum.

Table 1. Information criteria values from fitting alternative parametric models

	Parametric models			
	Weibull	Exponential	Log-logistic	Log-normal
AIC	283.97	926.02	354.71	526.94
BIC	292.11	930.10	362.85	535.09
-2log-likelihood	279.97	924.02	350.71	522.94

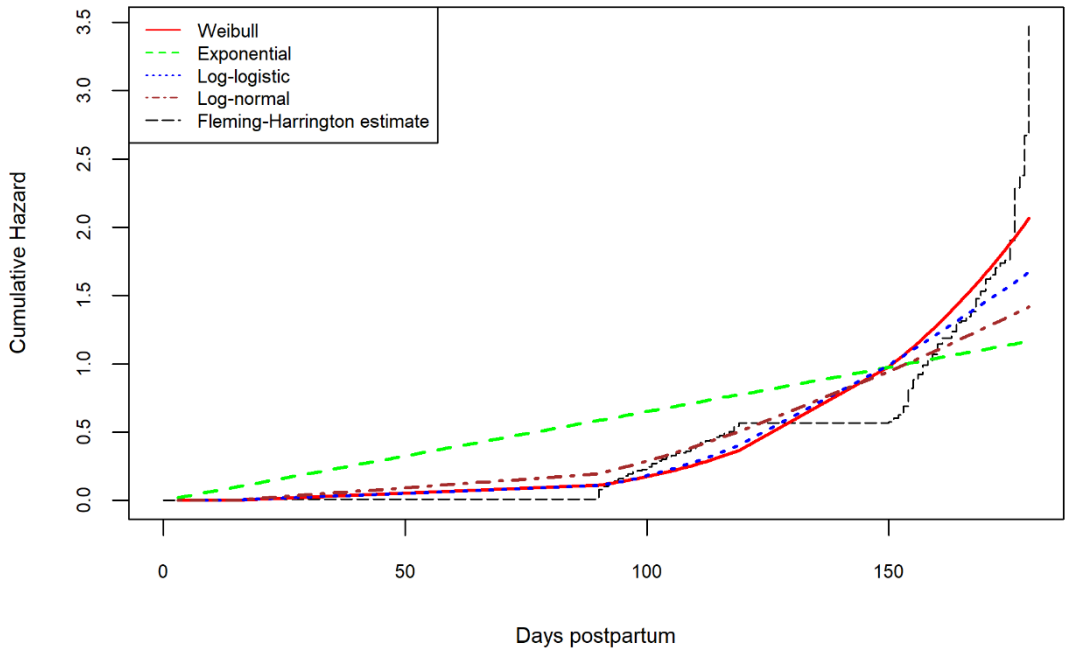

Figure 2. Cumulative hazard estimates from EBF duration data and fitted hazards from alternative parametric models

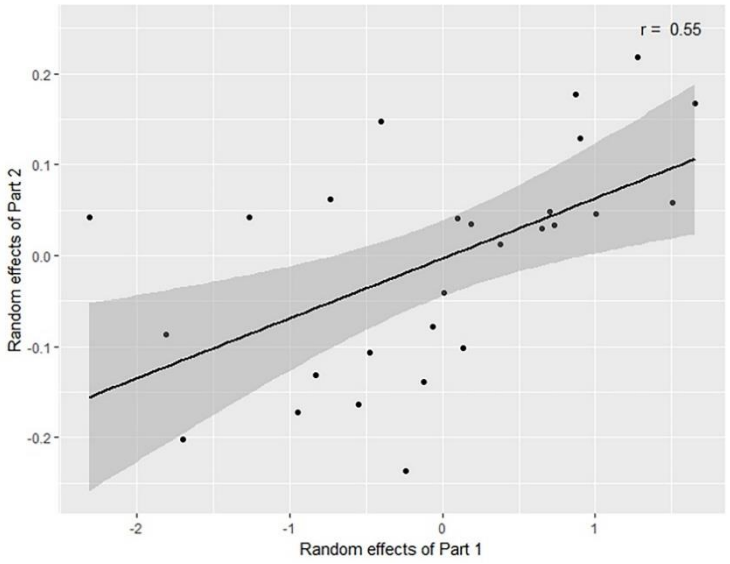

Figure 3. Correlation between random effects estimated from two independent models (the logistic mixed-effects model and the Weibull AFT frailty model).

A correlated two-part mixed-effects model was consequently fitted to the data, while a two-part mixed-effects model assuming independent random effects was also fitted for comparison purpose. As shown in Table 2, a significantly positive correlation ($\rho = 0.67$, $P<0.001$) between the two parts was confirmed, indicating that the independent random effects assumption was inappropriate. A positive correlation was evident between the baseline prevalence of EBF and EBF duration, i.e., mothers in the community with a higher prevalence of EBF at hospital discharge tended to have longer EBF durations. There were only slight differences in the estimated regression coefficients between the independent and correlated two-part mixed-effects models; however, variance components estimated by the independent two-part model were slightly smaller than those by the correlated two-part model, the latter also fitted the data better with a significantly smaller -2log-likelihood (likelihood ratio test statistic being 8.6, df=1, P=0.003).

Table 2. Parameter estimates in the two-part model with different assumptions for breastfeeding data

Parameters	Independent two-part model			Correlated two-part model		
	Estimate	SE	*P*	Estimate	SE	*P*
Part 1 logistic mixed-effects model						
Intercept	0.8361	0.4564	0.0780	0.8381	0.4584	0.0800
Birth mode	-1.8801	0.2777	<.0001	-1.9039	0.2785	<.0001
Grandmother feeding preference	0.4480	0.3485	0.2100	0.4691	0.3472	0.1900
Mother-child bonding	0.1443	0.2916	0.6200	0.1407	0.2894	0.6300
Part 2 Weibull AFT frailty model						
Gamma	4.9438	0.2146	<.0001	4.9465	0.2146	<.0001
Intercept	4.8690	0.0518	<.0001	4.8689	0.0510	<.0001
Birth mode	-0.0462	0.0399	0.2600	-0.0491	0.0397	0.2300
Grandmother feeding preference	0.0772	0.0413	0.0730	0.0781	0.0410	0.0680
Mother-child bonding	0.0798	0.0360	0.0350	0.0736	0.0355	0.0490
σ_u^2	1.4218	0.5515	0.0160	1.4928	0.5702	0.0150
$\rho\sigma_u\sigma_v$				0.1111	0.0474	0.0270
σ_v^2	0.0182	0.0064	0.0083	0.0186	0.0065	0.0082
Correlation coefficient (ρ)				0.6668	0.1513	0.0002
-2log-likelihood (both parts)	4277.8			4269.2		

Abbreviations: AFT, accelerated failure time; SE, standard error.

The sensitivity analyses showed that the transformation of zero EBF durations in Part 2 time-to-event analysis did not affect the robustness of parameter estimation and model fit. Parameter estimates and the goodness-of-fit were exactly equivalent between the model without zero EBF durations and the model including transformed zero EBF durations as censored observations.

4. Discussion

A two-part mixed-effects modeling approach was proposed for analyzing clustered time-to-event data with clumping at zero. This approach takes into account the correlation within each clustering unit by incorporating random effects in each part. The method further takes into account the possible correlation between the two random effects, that is, the correlation between the baseline prevalence and the followed-up time-to-event outcomes. Compared to the conventional survival analysis, this approach makes full use of all available information at baseline and during the follow-up. In addition to breastfeeding studies, the methodology has potential applications in a wide range of research areas, such as social science, finance and health care, in which clustered/ longitudinal time-to-event outcomes with clumping at zero arise. A notable research area is duration analysis in political science, where political scientists have proposed solutions to selection bias derived from nonrandom sample selection, while a correlated random effects approach has not been developed and implemented yet.[19]

The two-part mixed-effects modeling approach has been discussed widely in the literature,[25] but most existing works have been focused on longitudinal semi-continuous data[9,10,13,26] and longitudinal count data with excess zeros.[12,27] Applications of the two-part mixed-effects model on clustered time-to-event data with clumping at zero are very

sparse. To our best knowledge, our proposed two-part mixed-effects model is the first attempt to deal with such data.

As discussed in the work by Su, Tom, Farewell [26], an incorrect assumption about the correlation between random effects in two parts can introduce bias in the two-part modeling of semi-continuous data. The results obtained from the illustrative example in our study showed a consistent conclusion that the variance component was underestimated in Part 2 survival modeling if the model was misspecified as independent random effects when the correlation between two parts actually existed.

Clustered time-to-event data are often encountered in longitudinal epidemiological studies. The proposed correlated two-part mixed-effects modeling approach takes into account the correlations possibly presenting in a two-level hierarchical data structure, i.e., subjects nested within different clustering units, via random effect terms u_i and v_i in the model. The method can be extended to handle higher level hierarchical or a multilevel data structure, by specifying a corresponding variance-covariance structure for depicting correlations between random effects.

Conflicts of interest statement

The authors of this paper declare no conflicts of interest.

Statement of ethical approval

The illustrative breastfeeding study was approved by the Human Research Ethics Committee of Curtin University, Australia (HR 184/2013), and the Nepal Health Research Council, Nepal (773/2014).

Funding

This study was partially supported by China Scholarship Council (Grant NO: 201406240008).

References

1. Gartner LM, Morton J, Lawrence RA, et al. Breastfeeding and the use of human milk. *Pediatrics.* 2005;115(2):496-506.
2. Jordan SJ, Siskind V, A CG, Whiteman DC, Webb PM. Breastfeeding and risk of epithelial ovarian cancer. *Cancer Causes Control.* 2010;21(1):109-116.
3. Kull I, Almqvist C, Lilja G, Pershagen G, Wickman M. Breast-feeding reduces the risk of asthma during the first 4 years of life. *J Allergy Clin Immunol.* 2004;114(4):755-760.
4. Palmer JR, Viscidi E, Troester MA, et al. Parity, lactation, and breast cancer subtypes in African American women: results from the AMBER Consortium. *J Natl Cancer Inst.* 2014;106(10).
5. World Health Organization. Breastfeeding. 2018; http://www.who.int/topics/breastfeeding/en/. Accessed February 10,, 2018.
6. Scott JA, Binns CW, Oddy WH, Graham KI. Predictors of breastfeeding duration: evidence from a cohort study. *Pediatrics.* 2006;117(4):e646-655.
7. Qiu L, Zhao Y, Binns CW, Lee AH, Xie X. Initiation of breastfeeding and prevalence of exclusive breastfeeding at hospital discharge in urban, suburban and rural areas of Zhejiang China. *International breastfeeding journal.* 2009;4:1.

8. Vila-Candel R, Duke K, Soriano-Vidal FJ, Castro-Sanchez E. Effect of Early Skin-to-Skin Mother-Infant Contact in the Maintenance of Exclusive Breastfeeding. *J Hum Lact.* 2017:890334416676469.
9. Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *J Am Stat Assoc.* 2001;96(454):730-745.
10. Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Stat Methods Med Res.* 2002;11(4):341-355.
11. Lambert D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics.* 1992;34(1):1-14.
12. Yau KK, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Stat Med.* 2001;20(19):2907-2920.
13. Lee AH, Xiang L. Mixture analysis of heterogeneous physical activity outcomes. *Ann Epidemiol.* 2011;21(10):780-786.
14. Tian L, Huang J. A two-part model for censored medical cost data. *Stat Med.* 2007;26(23):4273-4292.
15. Liu L, Strawderman RL, Cowen ME, Shih YC. A flexible two-part random effects model for correlated medical costs. *J Health Econ.* 2010;29(1):110-123.
16. Xing D, Huang Y, Chen H, Zhu Y, Dagne GA, Baldwin J. Bayesian inference for two-part mixed-effects model using skew distributions, with application to longitudinal semicontinuous alcohol data. *Stat Methods Med Res.* 2017;26(4):1838-1853.
17. Brown S, Ghosh P, Su L, Taylor K. Modelling household finances: A Bayesian approach to a multivariate two-part model. *J Empir Finance.* 2015;33:190-207.

18. Brockett PL, Golden LL, Guillen M, Nielsen JP, Parner J, Perez-Marin AM. Survival Analysis of a Household Portfolio of Insurance Policies: How Much Time Do You Have to Stop Total Customer Defection? *The Journal of Risk and Insurance.* 2008;75(3):713-737.

19. Boehmke FJ, Morey DS, Shannon M. Selection bias and continuous-time duration models: Consequences and a proposed solution. *Am J Polit Sci.* 2006;50(1):192-207.

20. Cox C, Chu H, Schneider MF, Munoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat Med.* 2007;26(23):4352-4374.

21. Nardi A, Schemper M. Comparing Cox and parametric models in clinical studies. *Statistics in medicine.* 2003;22(23):3597-3610.

22. Duchateau L, Janssen P. *The Frailty Model.* New York: Springer; 2008.

23. Khanal V, Lee AH, Karkee R, Binns CW. Postpartum Breastfeeding Promotion and Duration of Exclusive Breastfeeding in Western Nepal. *Birth.* 2015;42(4):329-336.

24. Kiernan K, Tao J, Gibbs P. Tips and strategies for mixed modeling with SAS/STAT® procedures. Paper presented at: SAS Global Forum2012.

25. Farewell VT, Long DL, Tom BDM, Yiu S, Su L. Two-Part and Related Regression Models for Longitudinal Data. *Annu Rev Stat Appl.* 2017;4:283-315.

26. Su L, Tom BD, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics.* 2009;10(2):374-389.

27. Lee AH, Wang K, Scott JA, Yau KK, McLachlan GJ. Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. *Stat Methods Med Res.* 2006;15(1):47-61.

Appendix: SAS sample codes

```
proc nlmixed data= tpmvk;
bounds gamma>0;
***Logistic component***;
eta1=betal_0+betal_1*birthmode+betal_2*grandmapref+betal_3*bonding+ran
l;/*ranl is the random effect in the logistic part*/
p=exp(eta1)/(1+exp(eta1));
if ebfbase=0 then loglik=log(1-p);/*log likelihood for the first
Logistic regression part*/
***Survival component***;
if ebfbase=1 then do;
eta2=betas_0+betas_1*birthmode+betas_2*grandmapref+betas_3*bonding+ran
s;/*rans is the random effect in the survival part*/
alpha=exp(-eta2);
loglik=log(p)-(alpha*ebfduration)**gamma+(censor=0)*(-
gamma*eta2+(gamma-1)*log(ebfduration)+log(gamma));/*log likelihood for
the Weibull survival part censor=1 indicates censored observation*/
end;
model ebfduration~general(loglik);
random ranl rans ~
normal([0,0],[exp(2*logsigl),cov_l_s,exp(2*logsigs)])
subject=CommunityCode;
estimate 'correlation coeffecient(ranl_rans_rho)'
cov_l_s/(exp(logsigl)*exp(logsigs));
estimate 'variancel' exp(2*logsigl);
estimate 'variances' exp(2*logsigs);
run;
```