Clinical Epidemiology and Global Health 7 (2019) 253–260

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

journal homepage: www.elsevier.com/locate/cegh

Survival analysis in longitudinal studies for recurrent events: Applications and challenges

Mani Thenmozhi[a], Visalakshi Jeyaseelan[a,*], Lakshmanan Jeyaseelan[a], Rita Isaac[b], Rupa Vedantam[c]

[a] Department of Biostatistics, Christian Medical College, Vellore, India
[b] Department of RUSHA, Christian Medical College and Hospital, Vellore, India
[c] Department of ENT Unit 3, Christian Medical College & Hospital, Vellore, 632004. India

ARTICLE INFO

Keywords:
Recurrent events
Upper respiratory infection
Data structure
Extended Cox model
Frailty model

ABSTRACT

Background and objective: Recurrent event data analysis is most commonly used in biomedical research. However, the researchers dealing with recurrent events in survival analysis have ignored the assumption that the recurrent events are correlated. There are methods available that takes into account dependency between recurrent events. The main objective of this study was to demonstrate the recurrent event models using upper respiratory infection (URI) among Indian infants.

Methods: The frequency of URI among a birth cohort of 210 babies was evaluated monthly with nasopharyngeal swabbing. Data on 11 potential risk factors were noted. The extended Cox models, such as Andersen-Gill counting process (CP), Prentice-Williams-Peterson (PWP-CP), PWP–Gap time model, Marginal Model and Cox frailty model were applied. The better model was assessed based on Loglikehood test statistics.

Results: Of the four models, PWP-CP model had lower log likelihood value. The URI incidence rate was estimated to be 2.27 (95%CI: 1.70–3.03)times significantly higher in the month of July–October and 1.43 (95%CI: 1.19–1.71) times higher in the month of November–February as compared to March–June ($p < 0.001$). Nasopharyngeal colonization with S. pneumoniae was also another important risk factor (HR = 1.18, 95%CI:1.01–1.39, p = 0.03).

Conclusion: In the current study, PWP-CP model was found to be better model as compared to other models. Also biologically appropriate as subsequent events depend on the first event of URI. Hence, the choice of an appropriate method for analyzing the recurrent event data should not be decided only on statistical basis but also based on the research question.

1. Introduction

Cox Proportional Hazard Model has been widely used by most researchers in the recent past due to its versatility and simplicity in nature for interpreting the results. This model is used in recurrent events such as repeated asthma attacks, episodes of upper respiratory infections, repeated myocardial infarctions, recurrent urinary tract infection among the renal transplant patients, etc. are very common in medical research. However, the researchers dealing with recurrent events in survival analysis have ignored the assumption that the recurrent events are correlated.[1–3] In such situation either they have used the latest event and the time related to that event as outcome or, they have assumed the recurrent events are independent and analysed data using survival analysis.[4,5] If the correlations between the recurrent events are ignored, then the null hypothesis is mostly rejected, because the Cox model does not incorporate within subject correlation. However, methods have been developed that make use of all available data, while accounting for the lack of independence of recurrent events within subjects. The two popular approaches are namely, "Variance-corrected Cox based models" and "Frailty/random effects" models.[6–8] Variance-corrected models were developed to account for correlation by using robust (sandwich) standard errors. However, the theory behind frailty models is that some subjects are intrinsically more or less prone to experience events of interest than others; frailty can be considered as a random covariate in the model that corrects dependence among recurrent event times. Limitations of applying these variance-corrected

* Corresponding author. Department of Biostatistics, Christian Medical College, Vellore, 632002, Tamil Nadu, India.
E-mail addresses: mani.thenmozhi@gmail.com (M. Thenmozhi), visali_pv@hotmail.com (V. Jeyaseelan), ljey@hotmail.com (L. Jeyaseelan), rita.isaac@cmcvellore.ac.in (R. Isaac), rupavedantam@cmcvellore.ac.in (R. Vedantam).

https://doi.org/10.1016/j.cegh.2019.01.013
Received 16 August 2018; Received in revised form 24 January 2019; Accepted 31 January 2019
Available online 02 February 2019

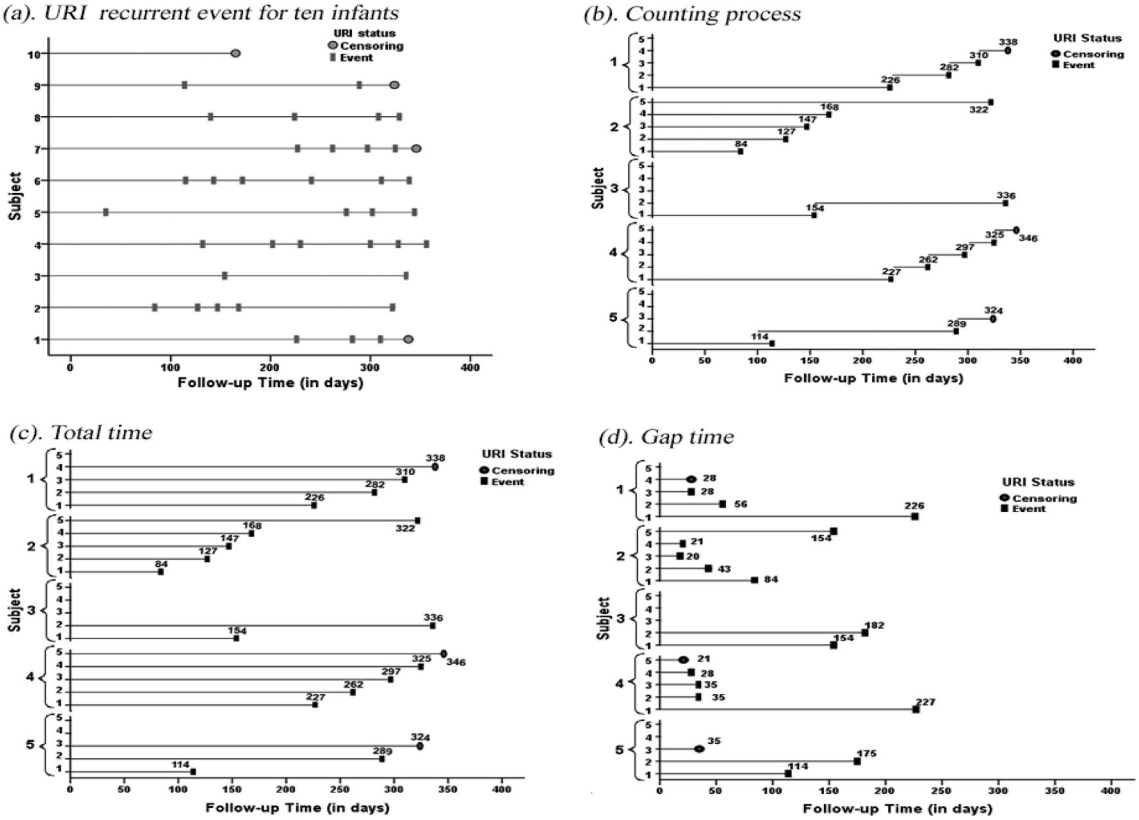

Fig. 1. Upper respiratory infection recurrent time to event data of birth cohort in the first year of life.

models and frailty models include their complexity and difficulty in implementation. Generalised Estimating Equation (GEE) analysis is a variance corrected model that requires the specification of a working correlation matrix but are still inefficient as the information on time to event is ignored. Most statistical models have been developed for analysing the recurrent event data which largely focused on binary outcome having discontinuous risk intervals using GEE.[7,9] Secondly, longitudinal data are analyzed using binary logistic regression ignoring the time, which is inappropriate. We intend to formulate a model which will consider the actual time of recurrence of each outcome. Though these concepts have been disseminated in statistical journals this is seldom practised in developing countries. Hence the aim of this paper is to summarize various methods for modelling recurrent event data. We would also show the differences in estimation and interpretation of recurrent event approaches, as well as to sensitize appropriate models, based on research objectives for the longitudinal study.

2. Materials and methods

2.1. Data

This study was conducted in K.V Kuppam rural development block, which belongs to the service area of RUHSA (Rural Unit for Health and Social Affairs) Christian Medical College, Vellore, India between February 2009 to August 2010. After taking an informed consent, a detailed socio-demographic history was obtained. Patient information was obtained from their parents who were interviewed at each visit regarding recurrent colds, allergic symptoms, overcrowding, family size, breast feeding, smoke exposure and day care attendance. At birth and at monthly scheduled visits, nasopharyngeal swabbing was performed with a calcium alginate swab stick. Then, the presence of upper respiratory infection was noted.[10]

2.2. Standard Cox PH model

The standard Cox proportional hazard model for the survival data specifies the hazard of ith individual as,

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}x_i\} \quad (1)$$

Where (t) $\lambda_0(t)$ is an unspecified baseline hazard function and β is the vector of regression coefficients, x_i is the vector of covariates of the i^{th} subject.

2.3. Extended Cox model

The extended Cox models were used to model recurrent time-to-event outcomes within a subject comprehensively than the Cox model. The extended Cox models were: 1) the Andersen-Gill counting process (CP), 2) the Prentice-Williams-Peterson (PWP-CP/Total time),3) PWP – Gap time (PWP-GT) model, 4) Marginal (Wei, Lin and Weissfeld)Model and 5) Cox frailty model.

2.4. Andersen Gill model

Andersen Gill model assumes that the correlation between event times for a subject can be explained by the past events. AG model is suitable model when correlations among events for each individual are induced by measured covariates. The counting process style of data input is seen in AG model where each subject is represented as series of observation with recurrence time given as $(t_0, t_1]$, $(t_1, t_2]$ … (t_m, last follow-up time] where, each recurrent event for the ith subject is assumed to follow a proportional hazard model is given as

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}_k\, x_i(t)\} \quad (2)$$

Under this model, the risk of recurrent event for a subject follows the usual Cox PH assumption but the number of recurrence is not taken

Table 1
Summary of time between consecutive URI recurrent Events.

	Follow-up time (in days)			No of patient with URI		
	Min	Max	Median	Event	Censored	Total
1st recurrence	2	339	98	185	18	203
2nd recurrence	49	382	197	162	13	175
3rd recurrence	77	393	243	135	12	147
4th recurrence	105	405	276	102	19	121
5th recurrence	152	424	311	69	18	87
6th recurrence	175	398	321	44	9	53
7th recurrence	203	386	338	32	2	34
8th recurrence	231	384	336	11	5	16
9th recurrence	287	363	347	6	1	7
10th recurrence	315	377	346	1	1	2

into account. Every subject risk intervals contribute to the risk set for every event, irrespective of the number of events for each individual (Fig. 1b).[11]

2.5. Prentice, William and Peterson model (PWP)

Another model for analyzing recurrent events is PWP model.[12] PWP CP model (total time) and PWP Gap time model. PWP CP model is similar to the AG-CP model but stratified by events. The baseline hazards vary from event to event, the hazard function for the k^{th} event for the i^{th} subject with the PH form is written as

$$\lambda_{ik}(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}_k\, x_i(t)\} \quad (3)$$

The PWP - GT model describes an intensity process from the occurrence of an immediately preceding event, with the gap time defined as $(t- t_{k-1})$. Both PWP approaches are conditional models as an individual is not considered in the riskset for the k^{th} event until experiencing the $(k-1)^{th}$ event.

$$\lambda_{ik}(t) = \lambda_0(t\text{-}t_{k\text{-}1})\exp\{\boldsymbol{\beta}_k\, x_i(t)\} \quad (4)$$

$\lambda_{0k}(t)$ represents the event-specific baseline hazard for the k^{th} event over time. AG model and both PWP Models are adjusted by estimating the sandwich type estimators and hence they are known as variance corrected models[13,14](Fig. 1d).

2.6. Marginal (Wei, Lin and Weissfeld) model

Wei, Lin and Weissfeld (1989) proposed a Cox-type model to analyse repeated events data.[15] In most applications the analysis has been the "time from the study entry" scale, since all the time intervals start at zero[16] (Fig. 1c). The hazard function for the k^{th} event for the i^{th} individual is,

$$\lambda_{ik}(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}_k\, x_i(t)\} \quad (5)$$

Unlike the AG model, this model allows a separate underlying hazard for each event. When an event is zero, it means that subject is no longer at risk after the last given event.[17,18]

2.7. Cox frailty model

The frailty model is an extension of the Cox PH model, in which, the hazard function depends on an unmeasured random variable.[18,19] The term 'frailty' means that each subject has his/her own disposition to failure, in additional to any effects that will be quantified using regression. Hazard function $\lambda_{ij}(t)$ for the recurrent time of the k^{th} event in the ith subject ($j = 1,2,\ldots k_i$; $i = 1,2,\ldots n$) conditional on the frailty Z_i, follows the PH form and its given by:

$$\lambda_{ik}(t) = \lambda_{0k}(t)Z_i\{\exp\{x_i(t)\beta_k\},\ t > 0 \quad (6)$$

Where, $\lambda_{0k}(t)$ is the common baseline hazard function, X_i is a vector of observable covariates and β is a vector of unknown regression coefficients. Frailty Z_i is the unobserved (random) common risk factors shared by all subjects in cluster 'i' and is assumed to be i.i.d random variable with unit mean and unknown variance θ.[19,20] The Frailty effects occur when the observed sources of variation in the observed or unobserved explanatory variables fail to account for the true difference in risk. That is, when there are other important but omitted variables presented, the effect of omitted variable can be captured by frailty.

3. Results

Number of recurrence experienced by infants ranged between zero to ten during the follow-up period. Seventeen infants (8.1%) out of 210 infants did not return to the study area after birth. The upper respiratory infection recurred at least once in 193 subjects and highest recurrence events (9 and 10 times) were observed in 7 patients.

Table 1 shows a summary of follow-up times and number of patients with and without URI event for the consecutive recurrent events. The median follow-up time to the first URI event were 98 days and starts increasing for the higher consecutive recurrent event.

A total of 163 infants (77.6%) had 6–13 visits where as 30 infants (14.3%) made < 5 visits. The median number of visits for these 193 infants was 9 visits. In thousand days of life, 845 records from 747 upper respiratory patients were followed–up during the study period and three infants died during the period of the study. The socio-demographic data were obtained at birth and child characteristics were presented in Table 2a and Table 2b. More parents resided in tiled/pucca houses than thatched houses (66.7%), and were labourers/unemployed. The majority of parents (61.9%) had at least high school education (84.8% fathers and 86.7% mothers). Majority of household had

Table 2a
Socio demographic and baseline characteristics.

Variables	Baseline (n = 210)
	n%
Sex	
Male	121 (57.6)
Female	89 (42.4)
Type of House	
Thatched	70 (33.3)
Tiled/Terraced/Group House	140 (66.7)
Parental Occupation	
Nil/Laborer	130 (61.9)
Petty Business/Professional/Others	80 (38.1)
Father's Education	
Illiterate/Primary	32 (15.2)
High/Higher Secondary and above	178 (84.8)
Mother's Education	
Illiterate/Primary	28 (13.3)
High/Higher Secondary and above	182 (86.7)
Birth weight (Grams)	
≤ 2500	76 (36.2)
> 2500	134 (63.8)
Smoke	
Yes	15 (7.1)
No	195 (92.9)
No. of members in the house	
≤4	48 (23.1)
> 4	160 (76.9)
Firev	
Yes	85 (40.5)
No	125 (59.5)
Water	
Bore well	124 (59.0)
River/Open Well	86 (41.0)
Nasophryngeal Swab Report	
Positive	8 (3.8)
Negative	201 (96.2)

Table 2b
Socio Demographic and baseline characteristics by URI recurrent events.

Variable	Event 1	Event 2	Event 3	Event 4	Event 5
	n (%)	n (%)	n (%)	n (%)	n (%)
Sex					
Female	85 (41.9)	73 (41.7)	62 (42.2)	50 (41.3)	93 (46.7)
Male	118 (58.1)	102 (58.3)	85 (57.8)	71 (58.7)	106 (53.3)
Type of house					
Tiled/Terraced/Grouped Houses	136 (67.0)	120 (68.6)	104 (70.7)	82 (67.8)	134 (67.3)
Thatched	67 (33.0)	55 (31.4)	43 (29.3)	39 (32.2)	65 (32.7)
Occupation					
Farmer/Bigbusiness/Petty business	75 (36.9)	63 (36.0)	52 (35.4)	39 (32.2)	54 (27.1)
Nil/Labourer	128 (63.1)	112 (64.0)	95 (64.6)	82 (67.8)	145 (72.9)
Father Education					
High school/Secondary and above	173 (85.2)	150 (85.7)	126 (85.7)	102 (84.3)	164 (82.4)
Illiterate/Primary	30 (14.8)	25 (14.3)	21 (14.3)	19 (15.7)	35 (17.6)
Mother Education					
High school/Secondary and above	186 (91.6)	161 (92.0)	135 (91.8)	111 (91.7)	180 (90.5)
Illiterate/Primary	17 (8.4)	14 (8.0)	12 (8.2)	10 (8.3)	19 (9.5)
Birth weight					
> 2.5 kg	130 (64.0)	109 (62.3)	90 (61.2)	71 (58.7)	116 (58.3)
≤2.5 kg	73 (36.0)	66 (37.7)	57 (38.8)	50 (41.3)	83 (41.7)
Smoking					
No	189 (93.1)	163 (93.1)	138 (93.9)	112 (92.6)	180 (90.5)
Yes	14 (6.9)	12 (6.9)	9 (6.1)	9 (7.4)	19 (9.5)
Mem5					
≤4	188 (93.5)	173 (98.9)	145 (98.6)	120 (99.2)	196 (98.5)
> 4	13 (6.5)	2 (1.1)	2 (1.4)	1 (0.8)	3 (1.5)
Fire					
No	123 (60.6)	107 (61.1)	91 (61.9)	75 (62.0)	101 (50.8)
Yes	80 (39.4)	68 (38.9)	56 (38.1)	46 (38.0)	98 (49.2)
Water					
Bore well	121 (59.6)	106 (60.6)	95 (64.6)	78 (64.5)	123 (61.8)
Open well/River	82 (40.4)	69 (39.4)	52 (35.4)	43 (35.5)	76 (38.2)
Swab					
Negative	139 (68.8)	102 (58.3)	80 (54.4)	83 (68.6)	166 (83.4)
Positive	63 (31.2)	73 (41.7)	67 (45.6)	38 (31.4)	33 (16.6)
Season(Months)					
March to June	24 (11.8)	28 (16.0)	42 (28.6)	56 (46.3)	128 (64.3)
July to October	118 (58.1)	56 (32.0)	23 (15.6)	8 (6.6)	21 (10.6)
November to February	61 (30.0)	91 (52.0)	82 (55.8)	57 (47.1)	50 (25.1)

(98.1%) < 3 under-five children, 76.9% of the households had more than 4 family members.

Fig. 1a: We have considered 10 children to demonstrate the risk intervals using URI data. Among those 10 children, seven had at least three events. Remaining three of the children was censored at 365 days; Subject 4 and 6 had largest number of events (6 events). Subject 3 had only two event at times 154 and 336 days.

Variance corrected models and Frailty model results are presented in Table 3, the 'seasonal' variable was the only consistently significant risk factor for an URI. Cox PH model had lower log-likelihood values. However, these lower likelihood values do not represent 'good fit' because it does not consider the subsequent events within each child.

3.1. Variance corrected model comparison

Children were most predisposed to upper respiratory infection in the July–October months and November–February months, which is statistically significant with slight differences in their parameter estimates in all the models. Using AG model, the upper respiratory infection was estimated to be 2.27 (95%CI: 1.70–3.03) significantly higher in July–October months and 1.43 (95%CI: 1.19–1.71) in November–February months as compared to March–June Months ($p < 0.001$). The PWP gap time model showed HR of 2.22 (95%CI: 1.66–3.00) for the July–October months and 1.37 (95%CI: 1.11–1.69) for November–February months respectively compared to March–June Months, which is statistically significant ($p < 0.01$). The WLW model for total time up to the 10th URI recurrence since the study entry yielded a HR of 1.58 (95%CI: 1.05–2.37, P value = 0.027) in July–October months and 2.50 (95%CI: 1.87–3.32) in November–February months as compared to March–June Months ($p < 0.001$). Nasopharyngeal colonization with *S. pneumoniae* was another important risk factor which was significant in all recurrent event models except PWP Gap time model. (AG model: HR = 1.23, 95% CI = 1.07–1.42, p value = 0.003; PWP total time model: HR = 1.18, 95% CI: 1.01–1.39, p value = 0.03; Marginal model: HR = 1.49, 95% CI: 1.14–1.95, p value = 0.003). Birth weight of an infant < 2.5 kgs had a risk of 1.14 (95% CI: 1.04–1.27) times of URI infection as compared to normal birth weight infant (Table 3).

3.2. AG model and frailty model comparison

The parameter estimates obtained from the frailty model with counting process time scale and AG models were almost same without frailty term. In other words, when the frailty model, with a variance almost close to zero ($\theta = 0$) would indicate that the frailty component does not contribute to the model. Based on the AG model, children with nasopharyngeal colonization with *S. pneumoniae* positive had high risk of recurring URI, which was 23% higher as compared to children without nasopharyngeal colonization by *S. pneumoniae*.Children were most susceptible to URI in July–October months (HR: 1.43, 95%CI: 1.70–3.03) and November–February months (HR: 2.27, 95%CI: 1.19–1.71) as compared to March–June months, which is statistically significant ($p < 0.001$). The cumulative hazard plot in Fig. 2a showed that both AG model and frailty model have estimated same cumulative hazard in the study and it clearly showed that if frailty variance is not significant. The frailty variance θ was estimated to be 0 and 0.153 for

Table 3
Risk factors for upper respiratory infection (URI) recurrent event data using Variance-Corrected model and frailty model in the first year of life.

Variable	Model 1 (AG Model)			Model 2 (PWP Total Time Model)			Model 3 (PWP Gap time Model)			Model 4 (Marginal Model)			Model 5 (Cox Frailty Model)		
	HR	95% CI	P Value	HR	95% CI	P Value	HR	95% CI	P Value	HR	95% CI	P Value	HR	95% CI	P Value
Season															
March–June	1.00			1.00			1.00			1.00			1.00		
July to October	2.27	1.70–3.03	< 0.001	2.60	1.14–5.94	0.023	2.22	1.66–3.00	< 0.001	1.58	1.05–2.37	0.027	2.30	1.84–2.86	< 0.001
November to February	1.43	1.19–1.71	< 0.001	1.50	1.12–2.02	0.007	1.37	1.11–1.69	0.003	2.50	1.11–1.69	< 0.001	1.44	1.17–1.76	< 0.001
Sex															
Male	0.95	0.83–1.09	0.492	0.93	0.86–1.01	0.098	0.93	0.81–1.07	0.308	1.09	0.81–1.07	0.504	0.95	0.82–1.11	0.530
Female	1.00			1.00			1.00			1.00			1.00		
Swap															
Positive	1.23	1.07–1.42	0.003	1.18	1.08–1.39	0.039	1.13	0.97–1.31	0.127	1.49	1.14–1.95	0.003	1.22	1.03–1.44	0.014
Negative	1.00			1.00			1.00			1.00			1.00		
Mem5															
≤4	1.01	0.48–2.16	0.966	1.02	0.49–2.11	0.954	0.91	0.42–1.97	0.811	0.97	0.42–1.97	0.849	1.66	0.84–3.31	0.960
> 4	1.00			1.00			1.00			1.00			1.00		
Smoking															
Yes	1.04	0.86–1.25	0.693	1.04	0.83–1.30	0.737	1.05	0.84–1.31	0.658	0.99	0.84–1.31	0.963	1.04	0.79–1.37	0.790
No	1.00			1.00			1.00			1.00			1.00		
Water															
Open well/River	0.97	0.84–1.14	0.748	0.96	0.80–1.16	0.691	1.01	0.87–1.16	0.947	1.08	0.87–1.16	0.562	0.97	0.83–1.12	0.750
Borewell	1.00			1.00			1.00			1.00			1.00		
Fire															
Yes	1.14	0.98–1.33	0.087	1.15	1.04–1.27	0.004	1.18	1.02–1.35	0.022	1.22	1.02–1.35	0.131	1.14	0.98–1.34	0.920
No	1.00			1.00			1.00			1.00			1.00		
Father Education															
Illterate/Primary	1.07	0.89–1.28	0.483	1.07	0.92–1.25	0.372	1.02	0.86–1.23	0.790	1.01	0.86–1.23	0.960	1.07	0.87–1.32	0.540
High school & above	1.00			1.00			1.00			1.00			1.00		
Mother Education															
Illterate/Primary	0.97	0.74–1.28	0.850	0.89	0.71–1.12	0.330	0.99	0.76–1.31	0.977	0.78	0.96–1.27	0.223	0.98	0.75–1.27	0.850
High school & above	1.00			1.00			1.00			1.00			1.00		
Birth_weight															
≤2.5 kg	1.11	0.96–1.28	0.151	1.14	1.04–1.27	0.009	1.11	0.96–1.27	0.145	1.30	1.08–1.27	0.035	1.11	0.95–1.30	0.190
> 2.5 kg	1.00			1.00			1.00			1.00			1.00		
Parent's Occupation															
Nil/Labourer	1.11	0.95–1.30	0.177	1.11	0.97–1.27	0.122	1.08	0.91–1.27	0.367	0.92	0.86–1.16	0.576	1.10	0.94–1.29	0.190
Professional & Others	1.00			1.00			1.00			1.00			1.00		
Type of House															
Pucca/Kacha	0.99	0.85–1.16	0.963	1.01	0.82–1.23	0.943	0.99	0.86–1.16	0.975	0.98	0.86–1.16	0.908	1.02	0.87–1.20	0.960
Tiled/Terraced/Grouped house	1.00			1.00			1.00			1.00			1.00		
Frailty Variance													0.00		0.870
Log likelihood	−3712.08			−2578.19			−2918.27			−5627.17			−3706.00		
R Square	0.083			0.092			0.072			0.129			0.083		

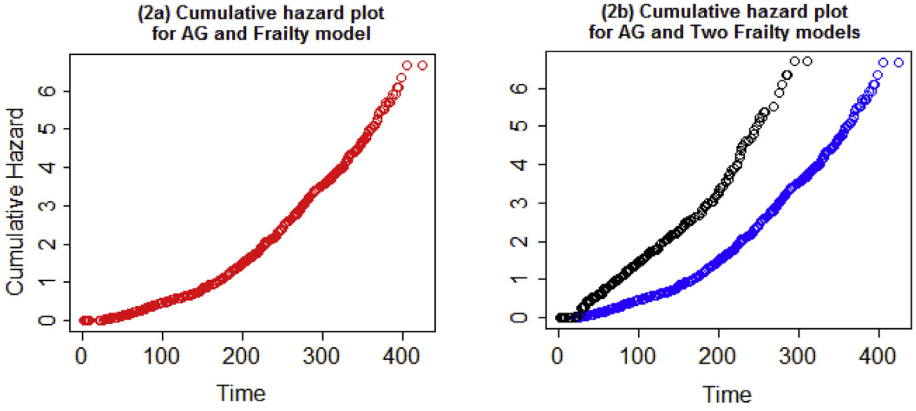

Fig. 2. Cumulative hazard plot for upper respiratory infection recurrence over a time of follow-up for AG model and frailty models.

counting process time scale and gap respectively. Fig. 2b shows that both models have the different estimated cumulative hazard over a time. Recurrent event data structure and how to organize the data for each recurrent event models, R code is given in the Appendix.

4. Discussion

Upper respiratory infections are the most common cause of morbidity in the first year of life among the Indian infants. We described the relevant methods, its importance, how to fit and interpret the results for different methods. Cox PH model does not examine the effect of the risk factors on the number of recurrence over the follow up time.[21,22] Many researchers continuously used logistic regression, GEE, Poisson and negative binomial approaches for estimating the risk factors for recurrent events.[7,21] However, they considered the total number of events per fixed period of time interval, ignoring the actual time to event concept between repeated occurrences.[10,21] Observations from the same child are expected to be correlated and hence Cox PH model is not suitable method to account of the extra variability of the recurrent event data. So, variance corrected models and frailty model were used.

Several methods have been proposed to account for intra-individual correlation that rises from recurrent events setting in survival analysis. The biological reason for the infection/disease is essential when selecting a model for the recurrent events for example if it is possible that after experiencing the first URI infection, the risk to the next infections may increase. If AG model is reasonable to assume that the risk of the repeated infections remains constant, irrespective of the number of previous infections, then the AG model is recommended.[18] AG Model provides more powerful inference for a covariate effect than the Cox model. A robust sandwich estimate is used to evaluate the standard errors.[11]

The PWP models are reasonable to assume that the child can only be at risk for a given event after he/she experiences the previous event. The PWP model means that the underlying hazard function is assumed to be the same from event to event. Also that, when infections increase with subsequent recurrence, the PWP model may be more appropriate than the AG model.

The WLW model overestimates the covariate effects due to the fact that every child has as many records as the maximum number of the event occurred in our data.[10] The variance corrected models handle correlation between the recurrent events occurring in the same patient by only correcting for the variance.[23,24]

Random effect/Frailty models leads to a person specific interpretation of the estimates which is similar to that of mixed models in longitudinal data in order to account for the dependency between the recurrent event and unobserved heterogeneity among patients properly as this cannot be explained by covariates alone.[25] Frailty model gives consistent estimates based on the distribution of the number of events and sample size. A small frailty variance implies very low correlation between the event times.[26] Frailty model is mainly applied for a moderate to large number of events but even for a small number of events, it is quite adequate.[18,27] In this study the within subject correlation is very low (0.07). However, this need not necessarily be the case always, when dealing with data. Therefore use of the entire model is based on the concept of the problems.

There is strong evidence in the literature that if frailty is present but ignored, then the covariate effects will be underestimated.[28,29] If the common baseline hazard between each event time is not appropriate for repeated event data and also, when a robust variance to any of these models does not adequately account for within-subject correlation, then it has been suggested to apply the frailty model which is also a similar finding from the present study.[23] However, if the primary interest of investigation is a measurement of dependence of within subject, then the frailty model is more adequate.[30] The difference between WLW model and frailty models is driven from the fact that, the frailty model is more naturally related to the fundamental performance of recurrences, while the unconditional WLW model does not provide understanding of the interrelationship among recurrences. However, it has been suggested that, if the frailty distribution is correctly defined, then the frailty model is expected to be more efficient than the WLW model.[31]

In summary, the choice of appropriate model for analyzing recurrent event data will be influenced by many factors, such as number of events, relationship among subsequence events, within subject correlation and varying covariates and the sample size. AG model is appropriate, when the assumption of a common underlying hazard over recurrent event observation is reasonable and when only interested in the overall rate of recurrences. When the dependence from the past event is strong and consistent, then the PWP model is appropriate. However, when the distribution of event per subject is small or prediction of time to the next event is of interest, the PWP gap time model is the appropriate method. However, for the present study, each episode of URI within a child is biologically related though the estimated correlation was very small. This was because the risk of infection to the same sero group/sero type was less in subsequent events within a child; thereby the estimated correlation coefficient was close to 0. However, if the researcher was interested to study the measurement of the dependence between recurrent event times within the subject, the frailty model would be appropriate.

5. Conclusion

The present study finding suggests that the choice of an appropriate method for analyzing the recurrent event data should not mainly depend on statistical basis such as model with low likelihood values; rather the selection should also be based on the research question, a

thorough clinical knowledge on the events of interest followed by organization of the data. Thus the PWP-CP model fit the data appropriately while the biological process also suggested the same model.

Conflicts of interest

None.

Appendix. Data structures for modelling recurrent event data

Organisation of the dataset is a more complication than the usual discontinuous risk intervals. Each subject is represented by several rows of data dependent on number of events child had, with time organized as intervals that represent (entry time, 1^{st} event], (1^{st} event, 2^{nd} event],……, (k^{th} event, end time]. A key difference for fitting recurrent event models is the creation of appropriate datasets. To show up the important features of data structure we present some information about the datasets that we used in the present paper.

Let's consider the example of first five children details from the Upper Respiratory Infection data are presented below:

Study ID	Start	Stop	URI Status	Gap	Sex	Swab	Months	URI count
1	0	226	1	226	1	0	3	1
1	226	282	1	56	1	0	3	2
1	282	310	1	28	1	1	3	3
1	310	338	0	28	1	0	1	4
2	0	84	1	84	1	0	1	1
2	84	127	1	43	1	0	2	2
2	127	147	1	20	1	0	2	3
2	147	168	1	21	1	0	2	4
2	168	322	1	154	1	0	3	5
3	0	132	1	132	2	1	1	1
3	132	202	1	70	2	1	2	2
3	202	230	1	28	2	1	2	3
3	230	300	1	70	2	0	3	4
3	300	328	1	28	2	0	3	5
3	328	356	1	28	2	1	3	5
4	0	154	1	154	1	0	1	1
4	154	336	1	182	1	0	3	2
5	0	35	1	35	1	0	1	1
5	35	276	1	241	1	0	3	2
5	276	302	1	26	1	0	3	3
5	302	344	1	42	1	1	3	4

A pair of variable ***(start, stop)*** is used to define the time interval of the URI. The start time is generally equal to 0 for the 1st URI and equals to the last recurrence time for further URI. The stop time is a recurrent URI time ***(URI status = 1)*** or censored time ***(URI-status = 0)***. The ***study ID*** variable identifies the child's. 1st child study ID = 1 from 226 to 282, 310 and 338 – with start time equal to 0 and stop time equal to follow up time, while child have four rows (study ID = 1 and 5). Child with no censoring in the end of the follow-up but the child five have 3 event and end of the visit he/she became censoring. Child study ID 4 had an event time at 154 days and second event time at 336 days. For five child data corresponding covariates are presented in the following column in the above table. This structure of the data can used to fit AG model and Frailty Model. In the PWP total time model, Gap time model addition information we added as URI counts based on the number of URI occurrence in the study duration, which is going to be used for stratification. The PWP Gap time model and frailty gap time model the time is defined as stop minus start time and the data structure as same.

Marginal approach focuses on total survival time from study entry until the occurrence of a specific (e.g., Kth) event. Suggested when recurrent events are viewed to be of different types also. Each subject is considered to be at risk for all failures that might occur, regardless of no: of events a subject actually experienced. For example in our study, every child to be at risk as the maximum number of recurrent events occurred in the study (k = 10) event if a child has one recurrence. i.e, every child has 10 observations, one in each stratum. In this data the URI event indicator, which is going to be used for stratification. Strata will correspond to the number of URI. Risk set determined from time since study entry. Marginal model is stratified model. The below data structure can be used to fit the marginal models.

Study ID	Start	Stop	URI Status	Sex	Swab	Months	URI count
1	0	226	1	1	0	3	1
1	0	282	1	1	0	3	2
1	0	310	1	1	1	3	3
1	0	338	0	1	0	1	4
1	0	338	1	1	0	1	5
1	0	338	1	1	0	1	6
1	0	338	1	1	0	1	7
1	0	338	1	1	0	1	8
1	0	338	1	1	0	1	9
1	0	338	1	1	0	1	10

R Code for the entire Model:

The library survival in R allows all recurrent event models, which is discussed in this paper.

```
library (survival)
```

```
library (foreign)
uri < -read.spss ("file location", use.value.labels = TRUE, to.data.frame = FALSE)
```

AG Model:

```
AG_Model < -coxph (Surv (Start, Stop,URI_status) ~ Mon_R + Sex_r + Swap_r.
+smk_r + water_r + fire_r + bwt_r + Pocc_r2+Toh_r + cluster (StudyID), data = uri)
summary (AG_Model)
```

Stratification Models: For the below models are stratified Models, the argument strata (URI_Count) identifies stratification variable to obtain their estimates. Estimates are obtained for event-specific effects for each covariates.

1. PWP-Total time Model:

```
PWP_TT < -coxph (Surv (Start, Stop,URI_status) ~ Mon_R + Sex_r + Swap_r.
+smk_r + water_r + fire_r + bwt_r + Pocc_r2+Toh_r + cluster (StudyID)+Strata (URI_Count),data = uri)
summary (PWP_TT)
```

2. PWP-Gap time Model:

```
PWP_GT < -coxph (Surv (Stop-Start,URI_status) ~ Mon_R + Sex_r+
Swap_r + smk_r + water_r + fire_r + bwt_r + Pocc_r2+Toh_r + cluster.
(StudyID)+Strata (URI_Count),data = uri)
summary (PWP_GT)
```

3. Marginal Model:

```
Marginal < -coxph (Surv (Start, Stop,URI_status) ~ Mon_R + Sex_r+
Swap_r + smk_r + water_r + fire_r + bwt_r + Pocc_r2+Toh_r + cluster (StudyID)+Strata (URI_Count),data = uri)
Summary (Marginal)
```

Frailty Model: By default *gamma* distribution is associated to the random effect for the frailty model in R software. However, we can specify the distributions such as gamma and Gaussian. Other way frailty.gamma (Study_ID) and frailty.gaussian (Study_ID) instead of frailty (id,-dist = ”gamma”)

```
Frailty < -coxph (Surv (Start, Stop,URI_status) ~ Mon_R + Sex_r + Swap_r.
+mem5_r + smk_r + water_r + fire_r + Fathedu_r + MothEdu_r + bwt_r + Pocc_r2+Toh_r + frailty (StudyID, dist = "gamma"),
data = uri)
summary (Frailty)
```

References

1. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol.* 1972;34(2):187–220.
2. Anker SD, McMurray JJ. Time to move on from ‘time-to-first’: should all events be included in the analysis of clinical trials? *Eur Heart J.* 2012;33(22):2764–2765.
3. Twisk J, Smidt N, de Vente W. Applied analysis of recurrent events: a practical overview. *J Epidemiol Community Health.* 2005;59(8):706–710.
4. Purroy F, Jiménez Caballero PE, Gorospe A, et al. Recurrent transient ischaemic attack and early risk of stroke: data from the PROMAPA study. *J Neurol Neurosurg Psychiatry.* 2013;84(6):596–603.
5. Gill DP, Zou GY, Jones GR, Speechley M. Comparison of regression models for the analysis of fall risk factors in older veterans. *Ann Epidemiol.* 2009;19(8):523–530.
6. Box-Steffensmeier JM, De Boef S. Repeated events survival models: the conditional frailty model. *Stat Med.* 2006;25(20):3518–3533.
7. Guo Z, Gill TM, Allore HG. Modeling repeated time-to-event health conditions with discontinuous risk intervals: an example of a longitudinal study of functional disability among older persons. *Methods Inf Med.* 2008;47(2):107–116.
8. Ullah S, Gabbett TJ, Finch CF. Statistical modelling for recurrent events: an application to sports injuries. *Br J Sports Med.* 2014;48(17):1287–1293.
9. Amorim leila D.A.F, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol.* 2015;44(1):324–333.
10. Rupa V, Isaac R, Manoharan A, Jalagandeeswaran R, Thenmozhi M. Risk factors for upper respiratory infection in the first year of life in a birth cohort. *Int J Pediatr Otorhinolaryngol.* 2012;76(12):1835–1839.
11. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat.* 1982;10(4):1100–1120.
12. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika.* 1981;68(2):373–379.
13. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text.* Springer Science & Business Media; 2005:616.
14. Clayton D. Some approaches to the analysis of recurrent event data. *Stat Methods Med Res.* 1994;3(3):244–262.
15. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc.* 1989;84(408):1065–1073.
16. Hosmer Jr David W, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data.* 2 edition Hoboken, N.J: Wiley-Interscience; 2008 416 p.
17. Cook RJ, Lawless J. *The Statistical Analysis of Recurrent Events.* Springer Science & Business Media; 2007 415 p.
18. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model.* Springer Science & Business Media; 2000:372.
19. Hougaard P. *Analysis of Multivariate Survival Data.* New York: Springer; 2001 542 p.
20. Duchateau L, Janssen P. *The Frailty Model.* 2008 edition New York: Springer; 2008 316 p.
21. Pandeya N, Purdie DM, Green A, Williams G. Repeated occurrence of basal cell carcinoma of the skin and multifailure survival analysis: follow-up data from the Nambour Skin Cancer Prevention Trial. *Am J Epidemiol.* 2005;161(8):748–754.
22. Dancourt V, Quantin C, Abrahamowicz M, Binquet C, Alioum A, Faivre J. Modeling recurrence in colorectal cancer. *J Clin Epidemiol.* 2004;57(3):243–251.
23. Kelly PJ. A review of software packages for analyzing correlated survival data. *Am Stat.* 2004;58(4):337–342.
24. Lee E, Wei L, Amato D. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein J, Goel P, eds. *Survival Analysis: State of the Art.* Dordrecht: Kluwer Academic Publishers; 1992:237–247.
25. Guo G. Use of sibling data to estimate family mortality effects in Guatemala. *Demography.* 1993;30(1):15–32.
26. Chen LM, Ibrahim JG, Chu H. Sample size determination in shared frailty models for multivariate time-to-event data. *J Biopharm Stat.* 2014;24(4):908–923.
27. Wienke A. *Frailty Models in Survival Analysis.* CRC Press; 2010 322 p.
28. Finkelstein DM, Schoenfeld DA, Stamenovic E. Analysis of multiple failure time data from an aids clinical trial. *Stat Med.* 1997;16(8):951–961.
29. Pickles A, Crouchley R. A comparison of frailty models for multivariate survival data. *Stat Med.* 1995;14(13):1447–1461.
30. Hougaard P. Frailty models for survival data. *Lifetime Data Anal.* 1995;1(3):255–273.
31. Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. *Stat Med.* 1994;13(21):2233–2247.