International Journal of Mosquito Research

ISSN: **2348-5906**
CODEN: **IJMRK2**
IJMR 2015; 2 (3): 114-121

Received: 08-07-2015
Accepted: 09-08-2015

Wimarsha Jayanetti
Department of Statistics
University of Colombo
Colombo 3, Sri Lanka

Roshini Sooriyarachchi
Department of Statistics
University of Colombo
Colombo 3, Sri Lanka

Correspondence:
Wimarsha Jayanetti
Department of Statistics
University of Colombo
Colombo 3, Sri Lanka

A multilevel study of dengue Epidemiology in Sri Lanka: modeling survival of dengue patients

Wimarsha Jayanetti, Roshini Sooriyarachchi

Abstract
This paper focuses on exploring methods and analyzing the survival pattern of clustered dengue data reported from high risk areas in Sri Lanka during the period 2006 to 2009. Due to dengue cases being clustered within districts resulting in cluster correlation, the response of survival was modeled in a multilevel framework. As the data consists of several missing values this paper further investigates multilevel multiple imputation as a method to handle the partially observed dengue dataset appropriately. A Discrete Time Hazard Model via standard logistic model has been suggested to model the survival of dengue patients. Results indicate that there is an impact from the clustering variable, district and from different types of dengue infections, place treated initially, Packed Cell Volume and White Blood Cell count on the response of interest.

Keywords: Multilevel data, Discrete Time Hazard Model, Multiple imputation, Dengue, Cluster

1. Introduction
Dengue is the most significant mosquito-borne disease, now endemic in most tropical countries, and a major public health concern. Dengue Fever and Dengue Hemorrhagic Fever (DF/DHF) have been classified as leading causes of hospitalization and death in Sri Lanka. According to the Epidemiology Unit in Sri Lanka, since the first reported outbreak of Dengue Fever in 1965, it has become more common after 1980s with progressively large outbreaks occurring more frequently. Thus, in order to prevent dengue, it is vital to conduct research and identify possible contributing factors to the disease in Sri Lanka. When considering the disease an important response of interest is the survival of dengue patients. Therefore, the aim of this paper is to understand the survival pattern of patients getting dengue and determine this response by taking into consideration the correlated nature of the data within the districts (clusters).

In epidemiological studies patients belonging to the same district would tend to have similar characteristics, which in turn results in an obvious hierarchical structure. This type of data typically generates a number of statistical problems, in which clustering is particularly important. Therefore, it is not reasonable to perform a traditional statistical analysis by pooling all the records of dengue in island-wide fashion or by fitting a different regression model within each district. To solve the statistical problems inherent in these, special statistical techniques are required. Thus, this paper is woven on a more appropriate technique known as multilevel analysis, which leads to correct standard errors, confidence intervals and significance tests compared to the traditional techniques, which simply ignore the presence of clustering.

The main data set used in this analysis was obtained from the Epidemiology Unit, Medical Statistic Bureau, Colombo, Sri Lanka. It consists of details about 24400 dengue patients reported from high risk districts during the period 2006 – 2009.

Initially, descriptive analysis was carried out to get a basic understanding of the structure of the variables. All the continuous explanatory variables were categorized according to their percentiles [1]. This is given in Table 1.

Generalized Cochran Mantel Haenszel Test for correlated categorical data [2] was used to identify which categorical variables have significant association with the response variable. Further, as missing observations are common in epidemiological studies, to use the maximum information "multiple imputation" was utilized.

To investigate the factors contributing to the survival of dengue patients, survival analysis was carried out. For this a multilevel discrete time hazard model was fitted by dividing the survival time span into three predetermined intervals based on the literature (Yang and Goldstein, 2003) [17]. By considering the probability that an individual dies in the current period, given that

the individual survived from the previous period, a multilevel discrete-time model, assuming a piecewise constant baseline hazard was fitted as a standard logistic model. Factors such as age, sex, ethnicity, place treated initially, fever, white blood cell count, platelet count, packed cell volume and type of dengue were considered as explanatory variables when fitting the model. Following the model building procedure, residual analysis was used to assess the adequacy of the model. Moreover, internal and external validations of model building and data from year the model were carried out before drawing conclusions.

Data from 2006-2008 were used for drawing conclusions. Data from 2006-2008 were used for model building and data from year 2009 were used to do external validations.

This research is novel with respect to two aspects. Firstly, no study of dengue has been done in Sri Lanka which looks at the survival of dengue patients. Also, studies carried out have been on an island-wide fashion and the correlation in the data has not been adjusted for. Methodologically speaking the literature does not contain dengue studies which combine multilevel multiple imputation together with discrete piecewise hazard modeling of survival data within a multilevel framework

Table 1: Description of the Data

Variable	Notation	Category	Code
Survival time	SURVIVAL	<7 days	1
		7-9 days	2
		>9 days	3
Outcome	OUTCOME	died	1
		discharged	0
Age	AGE	<18 years	1
		18-31 years	2
		>31 years	3
Sex	SEX	Male	1
		Female	0
Ethnic	ETHNIC	Sinhala	1
		Other	0
Place treated initially	PATTREAT	Government hospital	1
		Private hospital	2
Fever	FEVER	Yes	1
		No	0
White Blood Cell	WBCL	<3100	1-Low
		3100-4700	2-Moderate
		>4700	3-High
Platelet count	PLATL	<36000	1-Low
		36000-72000	2-Moderate
		>72000	3-High
Packed Cell Volume	PCVH	<40	1-Low
		40-45	2-Moderate
		>45	3-High
Classification	CLASIFI	DF	1
		DHF 1	2
		DHF 2	3

2. Materials and Methods

All the continuous explanatory variables were categorized based on their percentiles in order to avoid the problem of non-linearity between these and the response in modeling [1]. The survival response variable was categorized into 3 categories based on the literature from the Centers for Disease Control and Prevention of USA [http:// www.cdc.gov/ dengue/clinicalLab/clinical.html]. This leads to dividing the survival time span into three predetermined time intervals (< 7 days, 7-9 days, >9 days) to see during what period of time deaths are more likely to occur.

Initially, descriptive analysis was carried out to visualize the patterns in the data using frequency tables. Graphical displays were used to gain further understanding about the variables. Prior to model fitting it is essential to test the strength of the relationship between the response and explanatory variables through univariate analysis. Since the data set consists of a hierarchical structure, to check the associations, Generalized Cochran Mantel Haenszel test [2, 3] was performed. Missing observations are common in epidemiological studies [4]. Thus, to get the maximum use of costly collected data "imputation" was used. Imputation avoids non-response bias and enhances the precision and power by increasing the sample size. In the current data set missing values are present in health and laboratory variables. It is important to consider the stratification factor during the imputation in order to obtain intuition regarding the data. The REALCOM Impute software can handle the multilevel structure and properly works with categorical as well as normal data. Therefore REALCOM Impute software was preferred over others to impute missing data in demographic variables, and health and laboratory variables. Since the type of missingness (the chance of observations being missing) mechanism affects the validity of our subsequent analyses, it is important to identify what type of missingness is in the dataset. Missingness mechanisms can be classified using a typology first proposed by Rubin [5].

The literature mentions three types of missingness, namely: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) [6,7]. The definitions of missingness are stated by Tan, *et al*. [8]

In this study all the observed variables were used in the imputed model since exclusion of variables may cause bias if these are associated to the imputed value [9] Variables with missing values were used as the responses in the imputation model and variables without missing values were used as auxiliary variables. However, variables that have high missing percentage were excluded as recommended by Van Buuren *et al*. [10] due to the possibility that they may produce biased estimates. Here, a joint modeling approach was adopted using latent normal variables to impute ordinal and categorical data, and allowed for multilevel structure [11] and this is available in the REALCOM-impute software [12] The conventional method is to multiply impute many data sets and carry out a statistical analysis for each of these and then combine the results using Rubin's rules [13]. However, there are two reasons for the application of Rubin's rules being unsuitable for this study. Firstly, this combination rule assumes that the estimates are asymptotically normally distributed [14] However, in multilevel modeling, the second level variance is not normally distributed [15] Therefore Rubin's rule cannot be used in this study to combine the second level variance. Secondly, the Rubin's rules cannot be used to obtain a full data set which is necessary for validation. An alternative approach used by Van Leeuwen *et al*. which overcomes both problems posed by Rubin's rules was adopted [16]. This method multiply imputes many data sets and obtains the average of each observation for each binary variable. For binary variables, the averaged out observations were grouped as 1 if $P_i > 0.5$ or else grouped as 0. For categorical variables, the category having the highest proportion was considered as the category of interest.

A multilevel discrete-time model via the standard logistic model, assuming a piecewise constant baseline hazard was fitted to model survival status (dead or survived) of dengue patients considering the survival time of patients [17]. In doing this the districts were treated as units a level above individuals.

In order to fit a discrete-time model, the data must first be expanded so that every individual's record is replicated as many times as the observed number of time intervals before experiencing the event of interest (death) or being censored (survived). In the discrete time case, time is restructured into intervals. This means that the expanded data set in the discrete time case will be smaller than in the continuous timing case since there will be less risk sets [18].

A random intercept model is helpful to determine whether multilevel models are required in the first place. In this model intercepts are allowed to vary. Therefore the scores on the dependent variable for each individual observation are predicted by the intercept that varies across groups. This model assumes that slopes are fixed. Moreover, this model provides information about intra cluster correlations. In this study the intercept consists of two terms: a fixed component β_0 and district specific component, the random effect u_{0j}.

The multilevel random intercept discrete-time model via standard logistic model can be developed and expressed as follows. Suppose Π_{gij} is the probability that the i^{th} individual in the j^{th} district dies in the current period (g), given that he survived from the last period (g-1) then the two level random intercept model can be written as follows.

$$\text{Log}\left(\frac{\Pi_{gij}}{1-\Pi_{gij}}\right) = \beta_0 + u_{0j} + \sum_{g=1}^{n-1} \alpha_g T_{gij} + \beta X_{ij}$$

Where

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \text{.....(1)}$$

T_g corresponds to Indicators for the n time intervals and X_{ij} corresponds to covariates / factors (Design matrix).

The third term on the right hand side of the model which represents a piecewise constant baseline hazard function can take the form of a continuous polynomial function (Goldstein *et al*., 2002). In this research, blocking factors were used [18] to model the baseline hazard function. Blocking factors are a set of dummy variables corresponding to the risk sets, and take the form

$$\alpha_1 T_1 + \alpha_2 T_2 + \cdots + \alpha_n T_n$$

Where the α's are parameters to be estimated and for g = 1,...n

$$T_g = \begin{cases} 1 & \text{for time interval } g \\ 0 & \text{otherwise} \end{cases} \quad \text{.....(2)}$$

There is a dummy variable corresponding to each risk set. If one risk set is taken as the baseline then there are T_{g-1} dummy variables.

The above mentioned discrete time model needs the proportionality assumption. That assumption is known as the 'proportional odds' assumption when the logit link is used. For it to be valid, effects of the covariates should be same at all time points [18]. The proportional odds assumption can be tested by including interaction terms between predictors and time in the model. If there is a significant interaction that means that covariate is time varying and hence should be included to the final model [17]. The other assumptions in this model are that the hazard rate is assumed constant within the observed time intervals and the random error term at the second level is assumed to be normally distributed.

In order to build up the model gradually, the forward selection method was used together with the Wald statistic and the Deviance Information Criteria (DIC) to select significant variables [19]. MLwiN 2.19 software was used to build the models. It uses a quasi - likelihood procedure to estimate the parameters as maximum likelihood estimation is computationally intensive for discrete response models [15]

Caterpillar plots and normal plots (Rasbash *et al*., 2004) [15] are used to assess the validity of the fitted model. Sensitivity and Specificity calculations and ROC curves (Lalkhen and McCluskey, 2008) were used to internally and externally validate the models.

3. Results and Discussion

The dataset consists of dengue records of individuals in ten districts where dengue incidence is high. These selected districts are Colombo, Galle, Gampaha, Kalutara, Kandy, Kegalle, Kurunegala, Matara, Puttalam and Ratnapura. The period considered for the study is 2006-2009. To check the behavior of the data before modeling several graphing techniques and univariate analysis techniques was used. Due to limitations on space these graphs and detailed univariate tables are not given here, but only some patterns in the graphs and univariate analyses are explained. To capture the full extent of information, the dataset with missing values was used here.

When considering the deaths as a percentage of the total number of cases recorded in each year the highest was observed in the year 2006 and the lowest percentage was observed in the year 2007. The percentage of deaths with respect to total number of dengue cases shows a reduction in the percentage in the year 2009. The highest number of deaths due to dengue was observed in Colombo district and the lowest number of deaths were observed from the Ratnapura district during this time period. When considering the percentage of deaths in each district Colombo, Matara and Kurunegala have highest percentages of deaths in descending order.

After this descriptive analysis the univariate analysis was performed in order to enhance the findings obtained in the graphical analysis. Here the main objective is to gain a basic idea of the corresponding relationships between the response and the explanatory variables. The dataset concerned in this study violates the independence assumption as it takes a hierarchical structure. Violation of the independence assumption severely affects using traditional univariate techniques. Hence, as mentioned in the methods section an alternative method to the traditional chi squared test was adopted, namely, Generalized Cochran Mantel Haenszel (GCMH) Test for correlated categorical data (De Silva and Sooriyarachchi, 2012; Zhang and Boos, 1997) [2, 3].

Survival of dengue patients takes a hierarchical form with respect to patients being clustered within districts. Therefore, in the univariate analysis 'District' can be considered as the stratification factor, according to which patients are clustered. Here the response variable termed as 'Survival time together with outcome' refers to a variable with six categories (table 2). When the survival of dengue patients is considered both survival time and the outcome of a patient (time to cure or death) should be considered.

Table 2: Combined levels of survival time and outcome

	Outcome	
Survival time	**Discharged - "0"**	**Died - "1"**
1 - " < 7"	1	2
2 – " 7 - 9"	3	4
3 – " > 9"	5	6

The dataset contains nine explanatory variables at the patient level, namely 'Age', 'Sex', 'Ethnic', 'Place treated', 'Fever', 'WBC', 'Platelet', 'PCV' and 'Classification'. All nine variables are categorical variables with the Age, WBC, Platelet, PCV and classification being ordinal categorical. The univariate analysis was carried out for patient level using the 2006-2008 dataset (before doing imputation). The GCMH test was carried out to check the significance of patient level factors in the presence of the district as the stratification factor. According to these results, it is clear that all patient level variables significantly affect the response variable at 5% level of significance except 'Sex', 'Ethnic' and 'WBC'. Moreover, 'Age', 'Fever' and 'Platelet' variables are highly significant.

3.1 Multiple imputation of missing data

Multiple imputation was considered for the missing data of the variables. This needs an understanding of the structure of missing values in this study. Thus initially missing value proportions were analyzed. Table 3 shows the missing percentages associated with each of the variables.

It is well known that the validity of statistical analyses when there is missing data depend on the reason why data is missing in the dataset. Therefore, it is needed to make an assumption about the missing data mechanism, and based on this assumption the analysis will be carried out to meet the objectives. The epidemiologist involved with this study was consulted in order to determine the missing mechanism of the variables with missing values. According to Rathnayake and Sooriyarachchi [20], justification of the missing situation can be explained as below.

As shown in Table 3 the variables age, sex, outcome, fever, classification and the month have less than 5% missingness and as explained by the epidemiologist the missingness mechanism could be classified as MCAR. Therefore, as Harrell [21] recommended case wise deletion was applied to these six variables thus 427 records were deleted. Then the size of the data set becomes 9685. The occupation and the laboratory information such as Ig M and Ig G has a very high missing percentage over 50% and as recommended by Van Buuren *et al*. [10] these variables were removed from the analysis. Variables place treated initially and ethnic group indicated 5.3% and 22.2% missingness respectively. Reasons for this missingness as explained by the epidemiologist could be classified as MCAR.

Table 3: Description of the Missing Data

Variable	Missing cases	Complete cases	Missing percentage
Month	17	10095	0.2%
Survival time	1222	8890	12.1%
Age	92	10020	0.9%
Sex	1	10111	0.0%
Ethnic	2244	7868	22.2%
Occupation	7326	2786	72.4%
Place treated initially	532	9580	5.3%
Outcome	127	9985	1.3%
Fever	117	9995	1.2%
White Blood Cell	3764	6348	37.2%
Platelet count	1112	9000	11.0%
Packed Cell Volume	2917	7195	28.8%
Ig M	6098	4014	60.3%
Ig G	6416	3696	63.4%
Classification	107	10005	1.1%

Case wise deletion is a possibility in this situation, but according to Enders [22] eliminating this amount of data is wasteful therefore this variable was selected for imputation mainly to enhance the precision and power by increasing the sample size and to get the maximum use of costly collected and valuable data. The survival time of the patient, is one of the important variables of this study. The survival time contains some missing values due to the date of discharge being missing and date of discharge being before date of onset (recording errors). The missing percentage in survival time is 12.1%. According to the epidemiologist, this missingness is a mixture of MAR or MCAR. Since MCAR has more limited assumptions it is reasonable to consider the missing mechanism of that variable as MAR. Therefore, it was decided to impute this variable to get rid of non-response bias and also to increase the sample size. When considering laboratory variables such as white blood cell count, platelet count and packed cell volume the epidemiologist indicated that the missing mechanism could be MAR. Therefore, laboratory variables, blood cell count, platelet count and packed cell volume were selected for imputation mainly to avoid non-response bias caused by MAR missing mechanism and to increase the sample size.

A suitable option is to use multiple imputation if the missing data mechanism is MCAR or MAR by taking advantage of case wise deletion whenever possible [20] Therefore, after identifying missing mechanism, six variables were selected to apply multiple imputation to, namely, survival, ethnic, place treated, WBC, Platelet, and Packed cell volume.

Finally, missing values of Survival time, Ethnic, Place treated, WBC, Platelet, and Packed cell volume were imputed using variables without missing values (Age, Sex, Outcome, Fever and Classification) as predictor variables using REALCOM Impute software together with MLwiN. The REALCOM Impute model uses a latent normal structure via a probit link function, with the probit analogue of the proportional odds model for ordinal data and for unordered data it uses the maximum indicant model [23] for imputation. Once the model is specified, software fits the model using Markov Chain Monte Carlo approach [12].

In the pooling phase of multiple imputation the results from m complete data sets are combined for the inference. For this phase most of the studies have used Rubin's rule of averaging out the estimated coefficients of m imputations by using standard formulae developed by Rubin [13]. However, due to the two problems explained in the methods section Van Leeuwen *et al*. (2007) [16] method was used as specified in the methods section. In this study imputation was done using REALCOM-Impute software as it can handle multilevel structure and 100 multiply imputed data sets were combined using Van Leeuwen *et al*. [16] method.

Comparison of before and after imputation results can help to identify major structural differences after imputation if such exist. In order to have meaningful interpretation, it was decided to construct a single response variable to represent the combined levels of the response variables, survival time and outcome as explained previously in this section. Afterwards Generalized Cochran Mantel Haenszel (GCMH) Test was performed separately for before and after imputation dataset to identify which categorical variables have significant association with the response variables [2, 3]. Table 4 shows the p values of univariate analysis for the association between different explanatory variables, with respect to the survival time together with the outcome variable before and after

imputation. The test carried out was the GCMH test. The district to which the patient belongs is used as the second level variable for stratified data accordingly.

It can be observed from Table 4 that the variables that were considered to be insignificant remained to be so while those that were significant remained to be significant apart from the variable WBC which had become significant after the imputation. This may be due to the fact that since the sample size increases the standard error has decreased hence increasing the power of the analysis. Therefore, it can be concluded that no major structural differences could be observed after the imputation [2]

3.2 Modeling interval censored dengue survival data

The model building was done using MLwiN v2.19. First, the 2nd order Penalized Quasi-Likelihood (PQL) procedure was chosen to estimate parameters of the discrete time hazard model as it yields more accurate estimates (Goldstein, 2011). Moreover, after convergence with Restrictive Iterative Generalized Least Squares (RIGLS) procedure, the method was switched into Monte Carlo Markov Chain (MCMC), since it gives a Deviance Information Criteria (DIC) to compare competing models (Rasbash *et al*., 2004) [15]. To determine the best model, forward selection procedure was adopted.

Table 4: T_p statistic test results for response - Survival time together with outcome

Explanatory variable	Before			After		
	T_p	DF	P Value	T_p	DF	P Value
Age	53.8995	10	<0.0001	73.5187	10	<0.0001
Sex	5.6681	5	0.3399	8.6008	5	0.1261
Ethnic	2.4282	5	0.7873	1.8066	5	0.8752
Place treated	14.1115	5	0.0149	92.4880	5	<0.0001
Fever	27.5692	5	<0.0001	25.3546	5	0.0001
WBC	13.6298	10	0.1906	66.6395	10	<0.0001
Platelet	108.3871	10	<0.0001	131.3365	10	<0.0001
PCV	32.9568	10	0.0003	78.6684	10	<0.0001
Classification	34.8400	10	0.0001	103.9032	10	<0.0001

At each stage the DIC value was used to check the statistical significance of the added variables.

3.3 Multilevel Discrete Time Hazard Model

First, Survival time span was divided into 3 predetermined intervals, namely, <7 days (1), 7-9 days (2) and >9 days (3) and an indicator variable (T1, T2, T3 respectively) was introduced for each time interval [17]. The first time interval was taken as the base category. Then the data were restructured in time intervals corresponding to times when events occur. The data set was restructured as explained in the following example. Consider the example of 5 observations with survival time category (1, 1*, 2, 3*, 3), where '*' denotes censored (discharged/survived) observations. This restructured data set is given in table 5.

In each time interval the response has a code 1 if an individual died in the time period, and 0 otherwise. Now in the expanded dataset each individual has a line of data corresponding to every risk set they survive until either censoring or the event of interest occurs. The dataset of 9685 was expanded as in table 5. Then the size of the final data set was 17, 567.

At each stage, variables were added one at a time to the best model at the previous stage. Model with minimum DIC was selected as the best model at each stage. This was carried out until there was no decrement in the DIC value at which point the forward selection procedure was terminated. The variables, classification, place treated, WBC, PCV and Time interval were found to be significant. Then the Proportional Odds assumption was checked by including two–way interactions between covariates and time in the model. By examining these two-way interactions it was evident that the "Time *Classification" and **"**Time* Place treated initially**"** interactions were significant at the 5% level of significance. This indicates that the proportionality assumption is not valid for these variables. Therefore, these two interaction terms were included in the final main effects model. Figure 1 gives final discrete time hazard model

Table 5: Expanded data structure

Time interval (g)	Individual (i)	Response (d_{gi})	Indicator T1	Indicator T2	Indicator T3
1	1	1	1	0	0
1	2	0	1	0	0
1	3	0	1	0	0
1	4	0	1	0	0
1	5	0	1	0	0
2	3	1	0	1	0
2	4	0	0	1	0
2	5	0	0	1	0
3	4	0	0	0	1
3	5	1	0	0	1

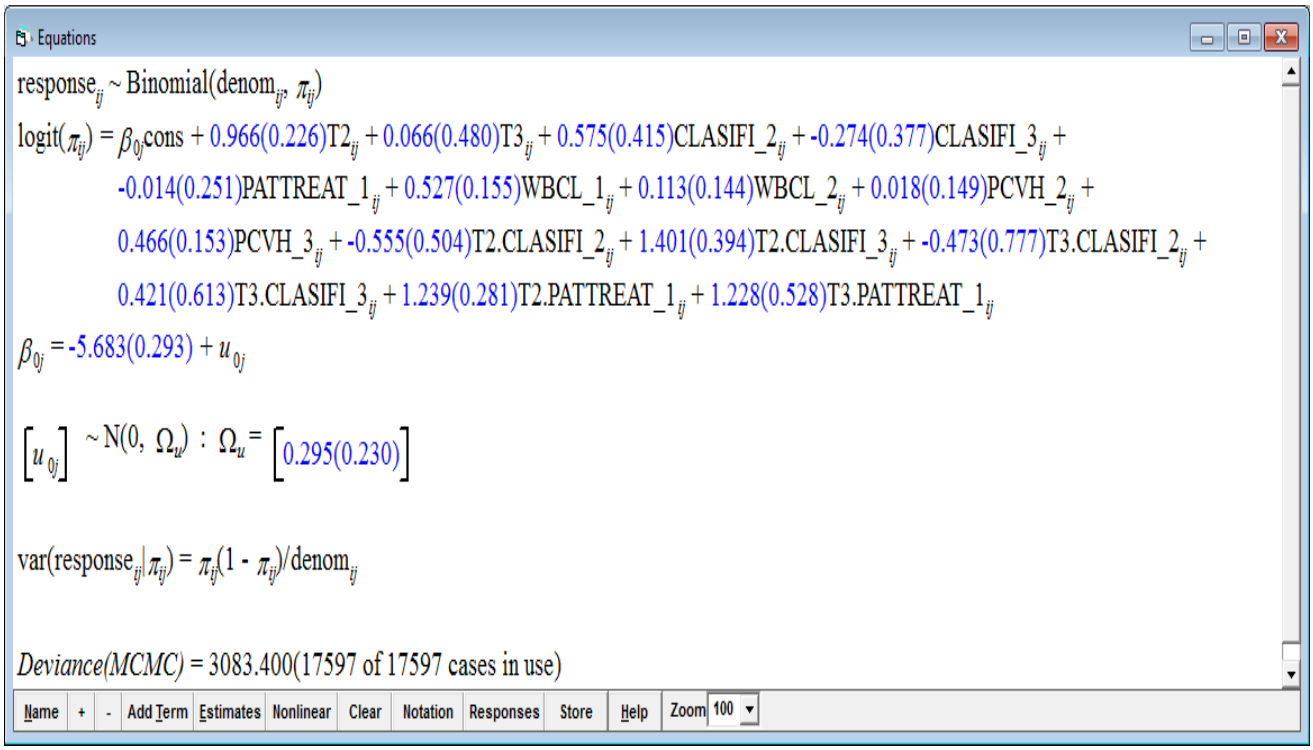


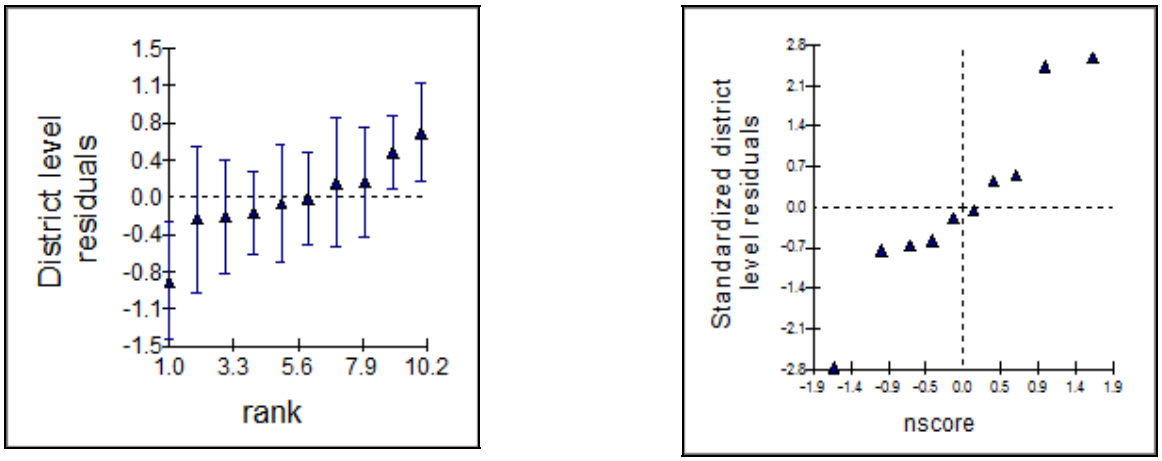
Fig 1: Discrete Time Hazard Final Mode

Table 6: Results of the final model

Factor	Category	Coefficient	SE	Wald Statistic	p-value
Classification	DHF1	0.575	0.415	1.919727	0.165887
	DHF2	-0.274	0.377	0.528224	0.467354
Place treated	Government	-0.014	0.251	0.003111	0.95552
WBC	Low	0.527	0.155	11.56	0.000674*
	Moderate	0.113	0.144	0.615789	0.432616
PCV	Moderate	0.018	0.149	0.014594	0.903845
	High	0.466	0.153	9.276603	0.002321*
Time interval	T2	0.966	0.226	18.26995	0.000019*
	T3	0.066	0.48	0.018906	0.890636
Time* Classification	DHF1*T2	-0.555	0.504	1.21262	0.270814
	DHF2*T2	1.401	0.394	12.64398	0.000377*
	DHF1*T3	-0.473	0.777	0.370579	0.542689
	DHF2*T3	0.421	0.613	0.471675	0.492217
Time*Place treated initially	Government*T2	1.239	0.281	19.44151	0.00001*
	Government*T3	1.228	0.528	5.409148	0.020031*
Cons		-5.683	0.293	376.2011	0.000000*
Deviance Information Criteria (DIC) : **3105.542**					
* significant at the 5% level					

It is essential to check the suitability of the multilevel concept for the final model by testing the significance of the district level variance. It would be equivalent to fitting a single level model if between districts variance is zero. Then the multilevel model will not be required as there will be no level 2 variation. The estimate of the district level variance and its 95% Bayesian credible interval is given by 0.295 and (0.074, 0.859) respectively. As the value zero does not lie within the 95% confidence band (0.074, 0.859). It can be concluded that it is appropriate to apply the multilevel concept.

Figure 2 gives a Caterpillar Plot, Normal Plot and Anderson Darling Test Results for testing the adequacy of the final Model



AD statistic: 0.318

P-value 0.474 > 0.05

Fig 2: Caterpillar plot, normal plot and Anderson Darling test results of the final model

According to the caterpillar plot only Kegalle (1), Colombo (9), and Matara (10) districts exhibit 95% confidence intervals that do not include zero. This implies that these districts show significant differences from the overall mean predicted by the fixed part of the model. When the Anderson Darling test was carried out, a p-value of 0.474 was obtained. Hence, it can be concluded that there is no evidence to suggest that the district-level residuals are not normally distributed in the fitted final model.

3.4 Internal and External Validation

Specificity and sensitivity of several cutoff values ranging from 0 to 1 were calculated (Lalkhen and McCluskey, 2008). Predictions were calculated for the internal and external datasets. Afterwards observed and predicted outcome for each patient was compared by considering the time interval using a written SAS code. The calculated values of sensitivity and specificity depict that when sensitivity increases specificity decreases and vice versa. To get excellent classification between survival statuses (outcome) both sensitivity and specificity should be high. But this rarely occurs in practice. The main purpose of this study is to identify the factors associating to the hazard of dengue. The prediction was carried out in order to validate the model. In this study, we should be more concerned about the sensitivity of this classification test in order to identify deaths of dengue patients. Hence sensitivity should be higher on the other hand specificity should also be of acceptable value since incorrectly classifying a surviving patient as dying will be an incorrect decision and will involve the waste of time and resources in conducting further investigations (Rathnayake and Sooriyarachchi)[20] . Therefore, in order to satisfy all mentioned criteria cutoff value of (0.030, 0.013, 0.030) was chosen as the most appropriate cutoff. The classification results were obtained for internal data using the selected cutoff value. It can be seen that this cutoff gives a sensitivity of approximately 73% and specificity of 65%. That is probability of correctly identifying deaths using this classification rule is 0.73 while probability of correctly identifying survived patients is 0.65.

The classification results obtained for external data using previously selected best cutoff value indicates sensitivity of approximately 65% and specificity of 63%. That is the probability of correctly identifying deaths using this classification rule is 0.65 while probability of correctly identifying survived patients is 0.63.

3.5 Interpreting the parameters of the final model

Hazard ratios with their corresponding 95% confidence intervals were calculated for the final model. In the final model, variables which are not involved in interactions are PCV and WBC. Therefore the main effects of those variables can be interpreted based on the base category. Hazard of death for a patient with high PCV compared to a patient with low PCV is 1.5936. Hazard of death for a patient with low WBC compared to a patient with high WBC is 1.6938.

Survival time span is considered to be a factor which consists of three levels, namely T1 (< 7 days), T2 (7-9 days) and T3 (> 9 days). Place treated is a factor with two levels, namely Government and Private. Hazard of death for a patient having survival time in the 2nd time interval (T2) treated in government hospital compared to a similar patient treated in a private hospital is 3.4042. Hazard of death for a patient having survival time in the 3rd time interval (T3) treated in government hospital compared to a similar patient treated in a private hospital is 3.3669. For government hospital, hazard of death for a patient having survival time in the 2nd time interval (T2) with respect to 1st time interval (T1) is 9. For government hospital, hazard of death for a patient having survival time in the 3rd time interval (T3) with respect to 1st time interval (T1) is 3.6473. For government hospital, hazard of death for a patient having survival time in the 3rd time interval (T3) with respect to 2nd time interval (T2) is 0.4021. For private hospital, hazard of death for a patient having survival time in the 2nd time interval (T2) with respect to 1st time interval (T1) is 2.6274. For private hospital, hazard of death for a patient having survival time in the 3rd time interval (T3) with respect to 2nd time interval (T2) is 0.4066.

Survival time span is considered to be a factor which consists of three levels, namely T1 (< 7 days), T2 (7-9 days) and T3 (> 9 days). Classification is also a factor with three levels, namely DF, DHF1 and DHF2. For the 2nd time interval (T2), hazard of death of DHF2 patients with respect to DF patient is 3.0864. For the 2nd time interval (T2), hazard of death of DHF2 patients with respect to DHF1 is 3.0253. Hazard of death for DHF2 patient having survival time in the 2nd time interval (T2) with respect to a similar patient having survival time in the 1st time interval (T1) is 10.665. Hazard of death for DHF2 patient having survival time in the 3rd time interval (T3) with respect to a similar patient having survival time in the 2nd time interval (T2) is 0.1526. Hazard of death for DF patient having survival time in the 2nd time interval (T2) with respect to a similar

patient having survival time in the 1st time interval (T1) is 2.6274. Hazard of death for DF patient having survival time in the 3rd time interval (T3) with respect to a similar patient having survival time in the 2nd time interval (T2) is 0.4066

4. Conclusions

Much literature is not available with respect to modeling survival data of rare events with a large proportion of censored observations in multilevel framework. Hence, modeling survival of dengue patients in a multilevel structure via discrete time hazard model proved to be a new and challenging experience. The objective of this part of analysis is to identify the factors associated with the survival of dengue patients. The data relevant to dengue patients from 2006-2008 was chosen to perform model fitting. The model was first validated using internal data and then it was validated using external data (2009). Advanced analysis showed some deviations compared to GCMH test (De Silva and Sooriyarachchi, 2012) [2]. Even though Age was significant in the univariate analysis, it fails to become significant in the advanced analysis. Moreover, the variable WBC has become significant during the advanced analysis, whereas it was not significant in the univariate analysis before imputation. However, the variables Sex and Ethnic are still insignificant as obtained under the univariate analysis.

The Proportional Odds Assumption was checked by including two–way interactions between covariates and survival time in the model. "Time *Classification" and "Time*Place treated initially" interactions were found to be significant. Therefore, these two interaction terms were included in the final main effects model as they indicate that the proportionality assumption is not valid for those variables.

The Discrete Time Hazard Model sets out the clinical, demographic and time based risk factors for death by dengue. When patients have clinical symptoms and demographic risk factors contributing to a high hazard, especially in the high risk time periods extra vigilance on the patient is required. Moreover, the findings of this study can contribute to policy changes in the health sector.

5. Acknowledgements

Our sincere thanks go out to the staff of Epidemiology unit, Medical Statistics Bureau, Colombo 10 and Dengue Management Unit, IDH, Colombo for helping us to have the necessary data to carry out this study.

6. References

1. Wickramasuriya SL, Sooriyarachchi MR. A Multilevel Analysis to Determine the Impact of Demographic, Clinical and Climatological Factors on Type of Dengue. International Journal of Biological Sciences and Engineering (IJBE). 2013.
2. De Silva DBUS, Sooriyarachchi MR. Generalized Cochran Mantel Haenszel test for multilevel correlated categorical data: an algorithm and R function. Journal of the National Science Foundation of Sri Lanka, 2012; 40(2):137-148.
3. Zhang J, Boos DD. Generalized Cochran-Mantel-Haenszel test statistics for correlated categorical data. Communications in Statistics-Theory and Methods, 1997; 26(8):1813-1837.
4. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. American Journal of Epidemiology. 2004; 160(1):34-45.
5. Rubin DB. Inference and missing data. Biometrika, 1976; 63(3):581-592.
6. Little RJ, Rubin DB. Statistical analysis with missing data. Wiley-Interscience; 2nd edition. 2002.
7. Acock AC. Working with missing values. Journal of Marriage and Family, 2005; 67(4):1012-1028.
8. Tan MT, Tian GL, Ng KW. Bayesian missing data problems: EM, data augmentation and noniterative computation. CRC Press. 2010.
9. Bartlett B, Carpenter J. Missing Data – Concepts. London School of Hygiene & Tropical Medicine. Retrieved from http://www.missingdata.org.uk., 2012.
10. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine, 1999; 18(6):681-694.
11. Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. Statistical Modelling. 2009; 9(3):173-197.
12. Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. Journal of Statistical Software, 45(5), 1-14.
13. Rubin DB. Multiple Imputation for Nonresponse in Surveys, New York: John Wiley & Sons, Inc, 1987.
14. Ratitch B, Lipkovich L, Kelly M. Combining Analysis Results from Multiply Imputed Categorical Data. Pharma SUG 2013 – Paper SP03, 2013.
15. Rasbash J, Steele F, Browne W, Goldstein H. A user's guide to MLwiN, version. Centre for Multilevel Modelling, University of Bristol, 2004.
16. Van Leeuwen M, Zweers EJ, Opmeer BC, Van Ballegooie E, TerBrugge HG, De Valk HW, *et al*. Comparison of accuracy measures of two screening tests for gestational diabetes mellitus. Diabetes Care. 2007; 30(11)2779-2784.
17. Yang M, Goldstein H. Modelling survival data in MLwiN 1.20. London: Institute of Education University of London. 2003.
18. Stewart CH. Multilevel modelling of event history data: comparing methods appropriate for large datasets, Doctoral dissertation, University of Glasgow, 2010.
19. Agresti A. Categorical data analysis (Vol. 359). John Wiley & Sons, 2002.
20. Rathnayake G, Sooriyarachchi M. Automated Statistical Information System (ASIS) for Diagnosis and Prognosis of Life-threatening Viral Diseases. Sri Lankan Journal of Applied Statistics, doi:10.4038/sljastats.v15i3.7796, 2014; 15(3):185-210.
21. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001.
22. Enders CK. Applied missing data analysis. Guilford Press. 2010.
23. Aitchison J, Bennett JA. Polychotomous quantal response by maximum indicant. Biometrika, 1970; 57(2):253.