Tropical Agricultural Research Vol. 28(2): 200 – 210 (2017)
Short Communication

Competing Risk Model for Dengue Epidemiology in Sri Lanka: Modeling Length of Stay in Hospital

G.H.S. Karunarathna* and M. R. Sooriyarachchi

Department of Statistics
University of Colombo
Sri Lanka

ABSTRACT. *This paper focuses on exploring methods for analyzing length of stay in hospital and the discharge endpoint for dengue patients reported from high risk areas of Sri Lanka during the period of 2006-2008. Due to the length of stay being related to different endpoints, the response of length of stay is modeled with the concept of competing risk. A popular competing risk model of subdistribution proportional hazard model was fitted to find out the associated factors for different end points of the discharge competing risk model. Moreover this study concludes that the effect of age, ethnicity, dengue classification, district and platelet count are important to the discharge status of the competing risk model.*

Keywords: *Competing risk model, Hazard, Length of stay, Subdistribution proportional hazards model*

INTRODUCTION

Dengue is a common mosquito-borne infection and is a growing health problem in tropical and sub-tropical regions around the world. According to the Epidemiological Unit (2009), Sri Lanka, Dengue was first documented in 1962 in Sri Lanka and in the profile of communicable disease in Sri Lanka, dengue still remains a major problem despite the resurgence of leptospirosis in the past few years. In recent epidemics, dengue is a significant burden and stress on health care facilities caused by the increment of dengue patients. Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) has become a prominent cause of hospitalization and death in Sri Lanka since 1996. Dengue virus is related to significant morbidity and mortality. The identification of factors associated with hospital stay provides supportive documentation for special care setup for society and it is very helpful towards reducing the morbidity and mortality. Therefore, this study is mainly focused on investigating the risk factors for length of stay for the hospitalized dengue patients with the different end points which are discharge, transfer and death.

In several hospitals epidemiological research tends to be focused on hospital length of stay and discharge with different disciplines. Length of stay is the result of the interaction between patients' characteristics, hospital characteristics, and social characteristics (Sa *et al*., 2007). Moreover, literature often presents differing outcomes regarding the effect of these characteristics on length of stay. Early stage of study of the length of study focused on linear and nonlinear models and was estimated by ordinary least squares. The results of Fenn

* Corresponding Author: hasani@stat.cmb.ac.lk

and Davies (1990) exhibited that the deviation of length of stay should be identified as the conditional probability of discharge of the patients via the log duration model.

Hospital length of stay (LOS) cooperates with different destinations like death, transfers, etc., competing risks, which precludes the occurrence of the other events of interest, (Kalbfleisch & Prentice, 2002) is used to handle length of stay of dengue patients when length of stay has different end points. A key assumption, events of interest will eventually occur for all patients in the population in Kaplan-Meier estimates to handle time to event data is violated in the presence of competing events. Thus, this paper is woven on a more appropriate technique called competing risk for handling mortality and transfers as competing risk to investigate the associate factors of time to discharge.

The data set was obtained from Epidemiology Unit, Medical Statistics Bureau, Colombo, Sri Lanka and it consists of details about dengue patients reported from high risk districts during the period of 2006-2008.

Numerous modeling methodologies are available for assessing the effects of covariates on the cause-specific outcomes in competing-risk data (Prentice, *et al*., 1978; Larson & Dinse, 1985; Fine & Gray, 1999).Two popular approaches, cause-specific hazard and proportional subdistribution hazard (PSH), have been proposed for modeling competing risk with different aims, while modeling cause-specific hazard targets aetiological research, which investigate the causal relationship between risk factors or determinants and given an outcome, PSH model is beneficial for medical decision making and prognosis research, as it models the absolute risk of an event (predict the probability of a given outcome at a given time for an individual patient) (Noordzij, *et al*., 2013; Kohl*et a*l., 2015). Also, cause-specific hazard model each event separately by applying the standard cox regression model and PSH is an extension of cox regression that models the cumulative incidence function. Therefore, in this study emphasis is on PSH model to handle LOS of dengue patients since the researcher is interested in getting an absolute risk.

METHODOLOGY

In this study, the authors were interested in investigating associated factors for the time to discharge of dengue patients. Death and transfers set up a competing event, before patient discharge. The considered period for the study is 2006 – 2008 and the data set includes 8695 patients in Sri Lanka. The data set consists of high incidence districts, Colombo, Galle, Gampaha, Kalutara, Kandy, Kegalle, Kurunegala, Matara, Puttalam and Ratnapura. These were selected to investigate the associated factors for length of stay.

According to the dengue clinical course of dengue days being classified into three categories, 0-4 days, 4-6 days and 6-10 days and being called as febrile phase, critical phase and recovery phase respectively (Yacoub*et al*., 2014). This data set also investigates each individual studied within 10 days of the admission without considering categories to this study by considering the response length of stay as a continuous variable. For the patients who are admitted and discharged on different days, duration is calculated as the difference between the discharge and admission dates. There were 4 different discharge destinations, k, is equal to zero if the observation is censored (LOS >10 days), 1, if the individuals is discharged, 2, if the individual is transferred to another hospital and 3, if the individual dies in the hospital.

Before moving to the analysis, the dataset was cleaned to make it appropriate for analysis since the type of missingness (the chance of observations being missing) mechanism affects the validity of 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 in 1976. According to Rubin (1976), missingness mechanisms can be classified as Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). At the cleaning stage, commonly seen missing observation in epidemiology (Barzi & Woodward, 2004) was seen in this data set. So, missing mechanism was applied according to Rubin, 1976.

Then, descriptive analysis was carried out to gain further understanding about the variables of the data set. The data on dengue patients were analyzed descriptively to visualize the features in the data. For this purpose Cumulative Incidence Function (CIF) was used to identify whether the CIFs change according to the levels of each variable. All the continuous variables were categorized according to their percentiles in order to avoid the problems of non-linearity between these and the response in modeling (Wickramasuriya & Sooriyararachchi, 2013) to capture the CIF variation of variables (Table 1).

Table 1. Categorizing details of the data

Variable	Category	Code
Age	<18 years	1
	18-31 years	2
	>31 years	3
Sex	Male	1
	Female	0
Ethnicity	Sinhala	1
	Tamil	2
	Moor	3
	Other	4
Place Treaded Initially	Government hospital	1
	Private hospital	2
	Other	3
White Blood cell	<3100	1-Low
	3100-4700	2-Moderate
	>4700	3-High
Platelet Count	<36000	1-Low
	36000-72000	2-Moderate
	>72000	3-High
Packed Cell Count	<40	1-Low
	40-45	2-Moderate
	>45	3-High
Classification	DF	1
	DHF1	2
	DHF2	3
District	Colombo	1
	Gampaha	2
	Kegalle	3
	Kalutara	4
	Matara	5
	Kandy	6
	Ratnapura	7
	Kurunegala	8
	Galle	9
	Puttalam	10

As a part of the descriptive analysis, CIF was used to identify, changes according to the levels of each variable. CIF graphs also illustrate the difference between the CIFs of the levels of covariates, thus giving an idea whether the covariates are significant or not. But this would be only a univariate approach since the covariates are unadjusted for the effect of other covariates. Before fitting the model, Log Cumulative Hazard (LCH) plots, which is a similar interpretation as a probability plot for visually examining the proportionality assumption of hazard, was used to graphically check each covariate.

To evaluate the transmissions from hospital admission, Fine &Gray's proportional hazard regression for subdistribution approach was used. This approach extends the Cox proportional hazard model to account for the presence of competing events by modeling the effects of the covariates on the subdistributional hazard (Taylor *et al*., 2015). Then PSH model was used to analyse the competing event. In the PSH model, event of interest was considered as length of stay (i.e. Time to discharge), coded as 1. Censored observations coded as 0 and competing event, transfers, 2 and dead as 3. Without categorizing, original data variables were used for this model to find out absolute risk factors. Therefore age, platelet, WBC, PCV was used as a continuous variable as the original. Here SAS macro *%pshreg* was used to build up the model by using *PROC PHREG* function. (Kohl *et al*., 2015). The model was fitted using forward selection, backward selection and stepwise selection to check whether the same model.The factors such as age, sex, place treated initially, white blood cell count, platelet count, packed cell volume, district and type of dengue (classification) were considered as explanatory variables for analyzing LOS of dengue patients.

Model validation of proportional hazard model mainly focuses on checking the validity of the assumptions of proportionality of hazards. The Schoenfeld Residual Test proposed by Schoenfeld, 1982 was followed to check the independence between residuals and time. Hence it was used to test the proportional hazrd assumption in Cox model.

Non Parametric Cumulative Incidence Function (CIF)

As described by Pintilie, 2006, let $t_1 < t_2 < t_3 \ldots < t_r$ be the unique ordered uncensored time points. Let d_{kj} be the number of events of type k that occur at t_j. An individual is at risk t_j if his ordered time, whether censored or not, is t_j or larger. Let n_j be the number at risk at t_j and $\widehat{S(t)}$ be the Kaplan-Meier estimator of the probability of being free of any time t. The CIF can be obtained by assuming over all t_j, the probabilities of observing event k at time t_j,while the individual is still at risk (did not experience any event prior to t_j. $\widehat{S(t_{j-1})}$ is the probability of remaining event-free prior to time$t_j$and $\lambda_j$is the cause-specific hazard for event k at t_j,

$$\widehat{F_k(t)} = \sum_{all\ k, t_j \le t} \left[\widehat{\lambda_{kj}} \widehat{S(t)}_{j-1}\right] \sum_{all\ k, t_j \le t} \frac{d_{kj}}{n_j} \left[\widehat{S(t)}_{j-1}\right] \qquad [1]$$

is the CIF, representing the probability that an individual will experience an event of type I by time t, where $\widehat{S(t_0)} = 1.$

Log Cumulative Hazard (LCH) plots

These are plots of *log (-log (1-F))* against log (time), where *F* is the CIF for the event of interest. LCH plots are used to investigate the proportionality of the hazard assumption of a Cox proportional hazard model. If the lines in the LCH plot are nearby parallel, then, the assumption of the proportionality could be accepted.

Proportional Subdistribution Hazard Regression

The proportional subdistribution hazard model was proposed by Fine & Gray, 1999 with the aim of estimating the effect of covariates on the cumulative incidence of the event of interest.

Let T be the (partialy unobservable) random variable describing the time at which the first event of any type occurs in an individual,C=1,2,…..and *k* is an event of interest type related to that time. The subdistributional hazard of event type k,$h_k(t,X)$ defined as,

$$\boldsymbol{h}_k(t,X) = \lim_{\Delta t \to 0} [\frac{1}{\Delta t} pr\,[t \le T \le t + \Delta t], C = k, \Big| (T \ge t \; or \; (T \le t, C \ne k), X)\Big] \quad [2]$$

Equation [1] can be modified as a function of parameter vector β through,

$$\boldsymbol{h}_k(t,X) = \boldsymbol{h}_{0k}(t)e^{X\beta} \quad [3]$$

Where, $h_{0k}$is an unspecified baseline subdistribution hazard function.

Fine & Gray, 1999 showed that the partial likelihood approach is valid for estimation. let $t_1 < t_2 < t_3 ..... < t_r$ be the unique ordered uncensored time points,the partial likelihood of the proportional subdistribution hazards models was defined by Fine and Gray as,

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(X_{(j)}\beta)}{\sum_{i \in R(t_{(j)})} w_i [(t]_{(j)}) \exp(X_{(j)}\beta)} \quad [4]$$

Where, X_i- covariate row vector of the subject

$R_{(t_{(j)})}$- risk set for cause k at time t = { I; $t_i \ge t$ or ($t_i \le t$ and $C \ne k$)}

$w_i [(t]_{(j)})$ -weights given to an individual

$$= \begin{cases} 1 & \text{if } t_i \ge t \\ \dfrac{\widehat{G_{(t)}}}{\widehat{G_{(t_i)}}} & \text{if } C \ne k \text{ and } ti < t \end{cases} \quad [5]$$

$\widehat{G_{(t)}}$ - estimator of the survival function of the censoring distribution at t.

Testing proportionally Assumption

Scheoenfeld-type residuals and weighted schoenfeld-type residuals can be inspected in order to check the assumption of validity of the proportional subdistribution hazard assumption. For the event time $t_{(j)}$,a raw vector of schoenfeld type residuals is defined by,

$$\widetilde{U_{(j)}} = X_{(j)} - \bar{X}\left(\hat{\beta}, t\right) \quad [6]$$

$$\textbf{Where, } \bar{X}(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}, \hat{\beta} \textit{ is the maximum likelihood estimate of } \beta$$

$$S^{(0)}(\beta, t) = \sum_{i \epsilon R_{(t)}} W_i(t) \exp(x_i \beta), \qquad S^{(1)}(\beta, t) = \sum_{i \epsilon R_{(t)}} X_i W_i(t) \exp(x_i \beta),$$

and$w_i(t)$ is the weight of subject i at time t as defined in Eq. [5].

RESULTS AND DISCUSSION

Initially, missing value portions were calculated to understand the structure in this study. The main reason for the missing observation was wrongly input data at the data entering stage. Therefore, it can be classified as MCAR. Then, it was identified that all the variable shown in table 1, have less than 5% missingness. According to Harell, 2001, case wise deletion was applied for all 9 variables since there was less than 5% missing observations. But the variables, occupation and some laborarty data, lg M and lg G has a very high missing ness over 50%, and the variables were deleted as recommended by Van Buuren *et al*., 1999, Rathnayake & Sooriyarachchi, 2014. Then the analysis was carried out for the 7,990 individuals.

Out of the 7,990 patients, 7,603 (95.2%) were discharged from the hospital within 10 days. The highest number of dengue cases was reordered from Colombo district. Among the patients who were recorded from Colombo, 95.5% were discharged during the clinical course. In the clinical course categories, discharge rates are 97.5% at the febrile, 98.2% at the critical phase and 98.1% at the recovery phase.

To find out the probability that an event of type I occurs at or before time t, CIF was calculated and SAS was used for the analysis. The plot of estimated CIF against failure time graphically illustrates this probability. CIF plots indicate a higher probability to discharge DF patients and higher probability dead DHF2 patients. Also it shows when, age is between 18-31 years, platelet count >72,000, WBC range is 3100-4700, PCV as 40-45 and initial place treated is Government hospital, there is a higher probability to discharge a patient.

LCH plots were then used to investigate the proportionality assumption before fitting the model. This approach does not give reliable results on the actual situation as it is a univariate approach, due to its inability to grasp the dependencies among different failure events or among the covariates. But this approach is followed to get a preliminary view of the nature of the proportionality of hazard, due to inadequacy of the proper method to do on. R software was used for obtaining LCH plots. Since there was no significant difference in log cumulative hazard values for categories in each variable, a worthy conclusion couldn't be

had from this analysis. But the graphs clearly indicated that age and classification variable holds the proportionality assumption.

From the competing risk model, the patients who are having a competing event, are changed and create several disjoint episodes with time dependent weights decreasing from 1 to 0 (Kohl *et al*., 2015). Because of these additional episodes (multiple observations for each patient with a competing event), new data set consists with 14,452.The results before and after processing by macro is given in Figure 1 and 2.

LOS	Status	_id_ ▲	_censcrr_	_start_	_stop_
1	1	1	1	0	1
3	1	2	1	0	3
6	1	3	1	0	6
2	1	4	1	0	2
1	1	5	1	0	1
7	1	6	1	0	7
3	1	7	1	0	3
3	1	8	1	0	3
10	1	9	1	0	10
2	1	10	1	0	2
5	1	11	1	0	5
5	1	12	1	0	5
4	1	13	1	0	4
7	1	14	1	0	7
6	1	15	1	0	6
4	1	16	1	0	4

Fig.1. Extract of the input SAS data set

LOS	Status	_id_ ▲	_censcrr_	_start_	_stop_
4	1	13	1	0	4
7	1	14	1	0	7
6	1	15	1	0	6
4	1	16	1	0	4
0	3	17	0	13.1	14.1
0	3	17	0	23.1	24.1
0	3	17	0	127.1	128.1
0	3	17	0	24.1	26.1
0	3	17	0	67.1	68.1
0	3	17	0	32.1	33.1
0	3	17	0	95.1	106.1
0	3	17	0	39.1	62.1
0	3	17	0	128.1	155.1
0	3	17	0	11.1	12.1
0	3	17	0	19.1	20.1
0	3	17	0	36.1	37.1

Fig.2. Extract of the new SAS data set created by %pshreg

According to figures 1 and 2, status shows as 1 (discharge) up to the 16th individual. After the 16th individual, multiple observation were created for the competing event, transfer (status=2) and dead (status=3). Those events are considered as censored observations in the newly created data set.

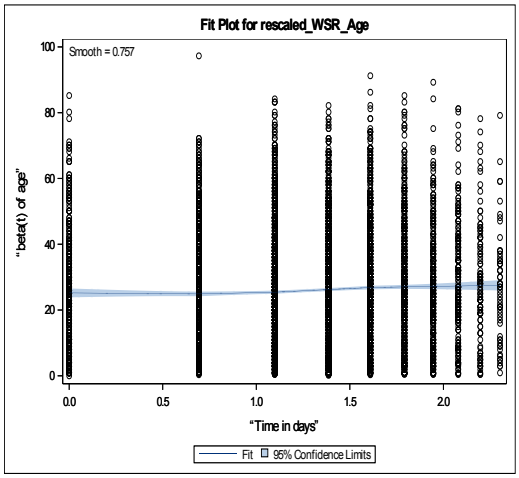
Same model was fitted from each selection method at 10 % significance level. The variables, age, ethnicity, classification, platelet and districts were found to be significant. Then the proportional odds assumption was checked by including two way interactions between covariate and time variable in the model. By examining these two-way interactions, it was evident that no interaction was significant at 10% level of significance. This indicates that the proportionality assumption is valid for all the variables. Therefore, there were no time-dependent effect covariates in the model. Estimates for the final model are given in Table 2.

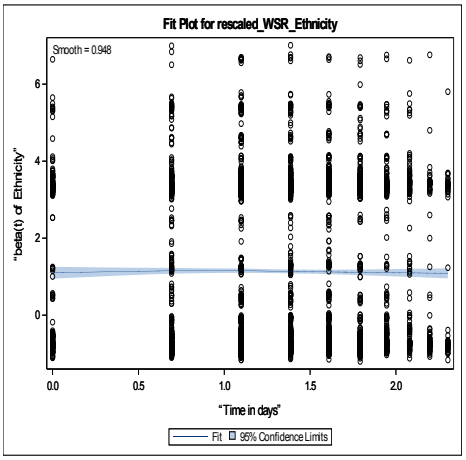
Table 2. Parameter Estimates in Final Fitted Model

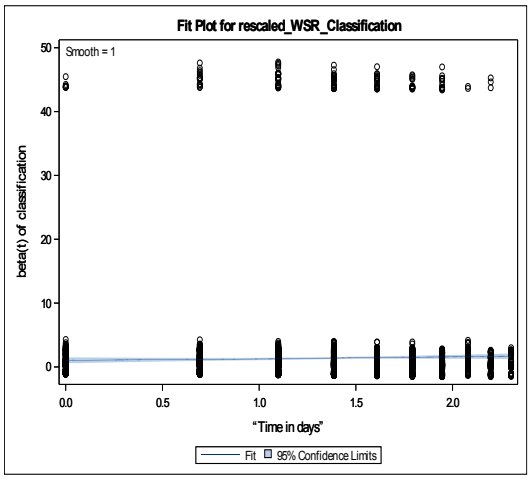
Parameter		DF	Parameter Estimate	Standard Error	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Age		1	-0.00259	0.0006072	<0.0001	0.997	0.996	0.999
Ethnicity	**2**	1	-0.18889	0.06497	0.0036	0.828	0.729	0.940
Ethnicity	**3**	1	-0.10001	0.05213	0.0551	0.905	0.817	1.002
Ethnicity	**4**	1	-0.18579	0.09196	0.0433	0.830	0.693	0.994
Classification	**2**	1	-0.08885	0.03734	0.0173	0.915	0.850	0.984
Classification	**3**	1	-0.17957	0.02833	<0.0001	0.836	0.790	0.883
platelet		1	1.47037E-6	2.1193E-7	<0.0001	1.000	1.000	1.000
DIS	**2**	1	-0.20329	0.02548	<0.0001	0.816	0.776	0.858
DIS	**3**	1	-0.14490	0.04073	0.0004	0.865	0.799	0.937
DIS	**4**	1	0.07060	0.03364	0.0358	1.073	1.005	1.146
DIS	**5**	1	-0.19404	0.04703	<0.0001	0.824	0.751	0.903
DIS	**6**	1	-0.31462	0.05661	<0.0001	0.730	0.653	0.816
DIS	**7**	1	-0.19673	0.08823	0.0258	0.821	0.691	0.976
DIS	**8**	1	-0.11582	0.05285	0.0284	0.891	0.803	0.988
DIS	**9**	1	-0.22791	0.07899	0.0039	0.796	0.682	0.930
DIS	**10**	1	-0.07852	0.05881	0.1818	0.924	0.824	1.037

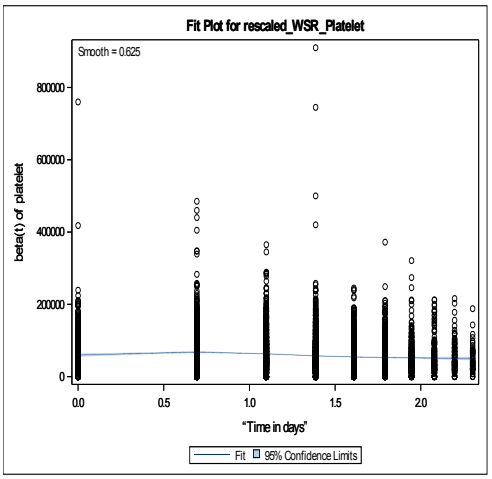
Table 2 depicts that age, ethnicity, classification, platelet and districts are significant at 10% level of significance. According to the hazard ratios, classification 2 is significant with a hazard ratio of 0.915. It indicates that compared to patients with DHF2 (classification 2), the patients with DF have a lower hazard by an amount 9% and patients with DHF3 (classification 3), the patients with DF have a lower hazard by an amount 16%. Also, since its estimates shows negative values, DHF2 and DHF3 show lower discharge probability and more likely to stay in the hospital. DHF3 has lower discharge than DHF2. Also, it says that instantaneous probability of length of stay is higher in older people and instantaneous probability of length of stay is lower in Sinhalese rather than Tamils and Muslims and people who live in Kalutara district. When considering the platelet count, if any person signposts lower platelet count, probability of discharge of that person is lower.

Schoenfeld type residuals were used for checking the model adequacy. Figure 3 shows the Schoenfeld residuals in the final model with the explanatory variable in the final model.









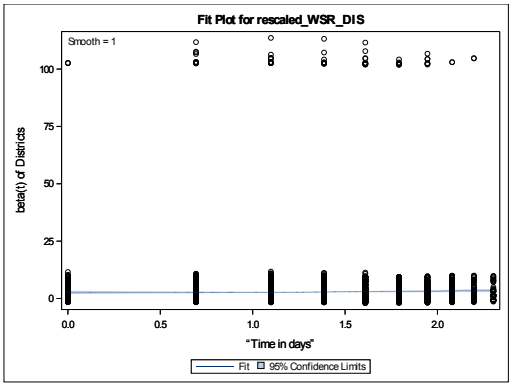


Fig.3.Schoefeld Residual plot for explanatory variable

According the residual plots, there is no evidence to suggest that it shows the decreasing / increasing importance of that variable. Therefore, it implies that proportionality assumption holds for this model and the fitted model is adequate.

CONCLUSIONS

The Subdistribution proportional hazard model was verified as a good model for handling the length of stay when there are competing events. Model implies that age, ethnicity, classification, platelet and districts are associated factors for length of stay for dengue patients. This study reveals that if older person lives in Kalutara, his/her platelet count is lower and if that person has DHF3, then probability of discharge is lower or length of stay in a hospital is higher in dengue patients. Finally, Model validity was checked by using schoenfeld residuals.

ACKNOWLEDGMENT

The authors would like to thank the University of Colombo for funding this research study, under its 'Research Grant 2014'. Further, we are grateful to 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.

REFERENCES

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: results from imputation of serum cholesterol in 28 cohort studies. American Journal of Epidemilogy, *160(1),* 34-45.

Epidemelogical Unit, Sri Lanka. (2009). Outpatient and First Contact Management of Dengue Fever/DHF. *36*. Epidemiology Unit.

Fenn, P., & Davies, P. (1990). Variation of Length of Stay. Journal of Health Economic, *9(2),* 223-243.

Fine, J.P., & Gray, R.J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of American Statistical Association, *94* (446). doi:10.1080/01621459.1999.10474144

Harell, F.E. (2001). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer.

Kalbfleisch, J.D., & Prentice, R. (2002). Statistical Analysis of Failure Time Data:2nd Edition. Willey_interscience.

Kohl, M., Plischke, M., Leffondre, K., & Heinze, G. (2015). PSHREG: A SAS macro for proportional and non proportional subdistribution hazards regression. Computer Methods and Programs in Biomedicine, *118,* 218-233.

Larson, M.G., & Dinse, G.E. (1985). A Mixutre Model for Regression Analysis of Competing Risks Data. Applied Statistics, *34*, 201-211.

Noordzij, M., Leffondre, K., van Stralen, K.J., Zoccali, C., Dekker, F.W., & Jager, K.J. (2013). When do we need competing risks methods for survival analysis in nephrology? Nephrology Dialysis Transplantation, 1-8. doi:10.1093/ndt/gft355

Pintilie, M. (2006). Analysisng and Interpreting Competing Risk Data. Willey, *26*(6), 1360-1367. doi:10.1002/sim.2655

Prentice, R.L., Kalbfleish, J.D., Peterson, A.V., Flurnoy, N., Farewell, V.T., & Breslow, N. E. (1978). The Analysis of Failure Times in the presence of Cometing Risks. Biometrics, 541-544.

Rathnayake, G.I., & Sooriyarachchi, R.M. (2014). Automated Statistical Information System (ASIS) for Diagnosis and Prognosis of LIfe-Threatening Viral Diseases. Sri Lankan Journal of Applied Statistics, *15(3)*, 185-210.

Rubin, D. B. (1976). Inference and Missing Data. Biometrika, *63(2)*, 581-592.

Sa, C., Dismuke, C.E., & Guimaares, P. (2007). Survival Analysis and Competing Risk Models of Hospital Length of Stay and Discharge Destination. ResearchGate. doi:10.1007/s10742-007-0020-9

Schoenfeld, D. (1982). Partial Residuals for the Proportional Hazard Regression Model. Biometrika, *69(1)*, 239-241.

Taylor, S.L., Sen, S., David, G.G., & Lawless, M. (2015). A Competing Risk Analaysis for Hospital Length of Stay in Patients with Burns. JAMA Surg.

Van Buuren, S., Boshuizen, H.C., & Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicin, *18(6)*, 681-694.

Wickramasuriya, S.L., & Sooriyararachchi, R. (2013). A Multilevel Analysis to Determine the Impact of Demographic, Clinical and Climatological Factors on Type of Dengue. International Journal of Biological Science and Enginering (LJBE).

Yacoub, S., Wertheim, H., Simmons, P.C., Screaton, G., & Wills, B. (2014). Cardiovascular manifestations of the emerging dengue pandemic. Nature Reviews Cardiology, *11*, 335-345. doi:10.1038/nrcardio.2014.40