Automated Statistical Information System (ASIS) for Diagnosis and Prognosis of Life-threatening Viral Diseases

G.I. Rathnayake and M. R. Sooriyarachchi *

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

*Corresponding Author: roshinis@hotmail.com

Received: 25, February 2014 / Revised: 17, September 2014/ Accepted: 5, November 2014

ABSTRACT

Diagnosis of life-threatening viral diseases, such as Meningitis, Viral Hepatitis, Japanese Encephalitis, Dengue, Leptospirosis (Rat Fever) to name a few, is extremely challenging particularly in low-resource settings, because the clinical presentation of such diseases cannot accurately be differentiated from that of other types of viral fever and laboratory tests need to be done to confirm the diagnosis. Due to limitations on cost or availability of diagnostics, or lack of access to laboratory facilities for specimen testing, it may not be possible to conduct diagnostic testing nationwide on all recorded suspected disease cases. Therefore epidemiologists will select a subset of such suspected cases for further investigation based on a rule of thumb. Thus a classification rule is vital to assist doctors in order to do this selection. In addition to diagnosis, it is also important to determine the prognosis of such patients as the concern is on life threatening diseases. Determining diagnosis and prognosis is often further complicated by the presence of missing values. The major objective of this study was to develop a user friendly Automated Statistical Information System (ASIS) that will output the diagnosis and prognosis of the patient when details regarding risk factors are given. In order to satisfy each of these objectives logistic modeling, survival modeling and Missing value imputation was used. Once the appropriate models were fitted, these models were combined using a Hierarchical Statistical Decision model (HSDM) to aid in developing the ASIS. The methodology developed was illustrated on a dataset of Acute Encephalitis Syndrome (AES) patients. The developed ASIS is applicable to any life threatening viral disease and it will help the epidemiologist to make quick decisions particularly in low income settings where there are low funds for sophisticated diagnostics.

Keywords: Hierarchical Statistical Decision model (HSDM), logistic model, survival model, Missing value imputation.

1. Introduction

1.1 Background

Life threatening viral disease surveillance systems in a country are developed based on the surveillance standards developed by WHO (Department of Vaccines and Biologicals, WHO, Geneva, 2003) which recommends that syndromic surveillance should be conducted nationwide with all health facilities reporting done at the national level on cases that meet a specified clinical case definition of probable disease. This provides a national estimate of probable disease. Whenever possible, laboratory tests should then be conducted to specifically identify the disease under consideration and differentiate it from viral fever. As stated in Advanced Immunization management (http://aim.path.org\) "Countries are requested to report the diseased cases to the WHO, but official notifications substantially underestimate the incidence due to limitations on cost or availability of diagnostics, or lack of access to laboratory facilities for specimen testing". Collection and analysis of data as accurately as conditions and resources allow is important as a first step in understanding disease burden. The classification rule developed by this study can be used to assist doctors and epidemiologists in order to select a subset of probable diseased cases for further investigation so that maximum benefit can be obtained from limited funds and laboratory resources. Determining the prognosis of patients and comparing their prognosis with that of other viral fever patients is also required for treatment of patients. A user friendly automated system that will determine the diagnosis and prognosis of patients will enable the epidemiologist and the medical doctor to obtain an initial understanding of the patient's condition before further investigation.

1.2 Objectives

The five primary objectives of this study were as follows:

- Discuss multiple imputation methods when data regarding several variables are missing.
- Develop a classification method using a Linear Logistic model to diagnose probable patients and differentiate them from viral fever patients. Develop a survival model to determine the prognosis of patients.
- Discuss how to identify factors that affect the diagnosis and prognosis of patients.

- Develop an automated system connecting the classification rule based on the logistic model and survival model which can output the outcome of the patient when details are given.
- Illustrate the methods developed on a set of Acute Encephalitis Syndrome (AES) patients from Sri Lanka.

This research sets out to achieve novel objectives in the sense that it automates both the diagnosis and prognosis of disease while most studies are only concerned with diagnosis. Another novel aspect of this study is its handling of non-response bias. Most previous studies have only resulted in a complete case analysis, thus wasting a large quantity of data. This type of system has wide applicability and can be used in any epidemiological study dealing with a life threatening viral disease.

1.3 Statistical Modeling

As the concern is to differentiate life threatening viral disease patients from viral fever patients this response is binary. Thus a logistic model was used to model the relationship between disease confirmation and the other risk factors with the objective of developing a classification rule.

The survival data was initially modeled using the Cox model (Cox, 1972), but the results revealed that the data does not fit well to the model. Alternatively the Accelerated Failure Time (AFT) model was used as suggested by Collett (Collett, 2003).

Finally a hierarchical multi-attribute statistical decision model (HSDM) was developed combining the classification rule developed based on the logistic model and the survival model in order to build a user friendly Automated Statistical Information System (ASIS) that will output the outcome of the patient when details regarding risk factors are fed into the system.

2. Materials and Methods

2.1. Missing Values and Multiple Imputation

Perusal of this type of data usually shows a number of observations with missing values for several variables. The usual method used by many analysts is to do a complete case analysis, throwing away the incomplete data. Complete case analysis restricts attention to cases where all variables are present; any observations with missing values for any of the covariates are deleted. Complete case analysis assumes missing values in the covariates are not associated with the outcome. It is also known as listwise deletion or casewise deletion. Generally this is unsuitable for several reasons. First it is a waste of data, next it may bring in non-response bias into the results and thirdly when dealing with relatively small data sets, it is impossible to reduce any observations. However casewise deletion is advantageous to be applied to variables having lower than 5% missingness where the missingness mechanism could be categorized as Missing Completely at Random (MCAR) (Harrell, 2001). When the missing data mechanism falls into one of the categories, missing completely at random (MCAR) or missing at random (MAR) (Little and Rubin, 2002), a suitable option is thus to use multiple imputation of the missing values (Greenland and Finkle, 1995 ; Janssen and Donders, 2010) together with casewise deletion whenever advantageous. The advantages of missing value imputation are that it makes maximum use of costly collected data, especially in the case of small samples and it helps in dealing with the problem of non-response bias. Certain groups of scientist argue against the concept emphasizing that it may give spurious results and may not always be practical, particularly with very large data sets. Therefore even though the use of multiple imputation has increased as pointed out by Van Buuren (Van Burren, 2007), it needs to be applied carefully to avoid misleading conclusions.

2.2 Diagnosis of Disease, using Logistic Modeling

Justification

The outcome variable is binary describing whether the patient is having the disease of interest or viral fever. Thus the logistic model was used since the logit link has been promoted for the case of binary data by many authors in the literature (Agresti, 2007). Further, with case–control studies, it is not possible to estimate effects in binary models with link functions other than the logit. This provides an important advantage of the logit link over links such as the probit. It is a major reason why logistic regression surpasses other models in popularity for biomedical studies. Collett (Collett, 2002) explains this model in detail.

Receiver Operator Characteristic (ROC) Curves

Hosmer and Lemeshow (Hosmer and Lemeshow , 2000) explain in their book that the Area under the ROC Curve (AUC), which ranges from 0.5 to one, provides a measure of the logistic model's ability to discriminate between those subjects who experienced the outcome of interest versus those who did not. The authors (Hosmer and Lemeshow, 2000) point out criteria for determining the discriminatory power of the model as a general rule of thumb. Literature reveals that ROC curves are extensively used for clinical decisions, and for determining the optimal cutoff (Tabatta and Shinchiro, 2009; Lin and Lee, 2002).

2.3 Determining a suitable survival distribution and form of model

Probability plots can be used for the purpose of checking the distributional assumptions. The distribution for which the probability plot is linear can be considered as a suitable distribution to model the failure time (Meeker and Escobar, 1998; Gan and Koehler, 1991). This becomes even more useful by plotting, in addition, simultaneous confidence bands (Meeker and Escobar, 1998). Based on the available data; any possible distribution function ($F(t)$) within these bands is statistically consistent with the data, indicating graphical goodness of fit.

There are two families of models which are commonly used for modeling survival data. These are the Cox proportional hazards family and the Accelerated Failure time family (Collett, 2003). Log Cumulative Hazard (LCH) plots can be utilized to identify if the Cox model is appropriate. If the LCH plots for the two groups of data are parallel, then the Cox PH model is appropriate (Collett, 2003). A quantile-quantile (Q-Q) plot provides an explanatory method for assessing the validity of an AFT model for two groups of survival data. If the points in the Q-Q plot fall on a line that is reasonably straight, this suggests that the AFT model is appropriate (Collett, 2003).

The predictive performance of the survival model was evaluated by considering measures of discrimination and calibration (Clarke, Bradburn et al, 2003)

2.4 Hierarchical Statistical Decision model (HSDM)

In order to develop the ASIS, the diagnosis (logistic) and prognosis (survival) models should be combined. A HSDM as explained by Bohanec, Zupan, and Rajkovic (Bohanec, Zupan, and Rajkovic, 2000) is used for this.

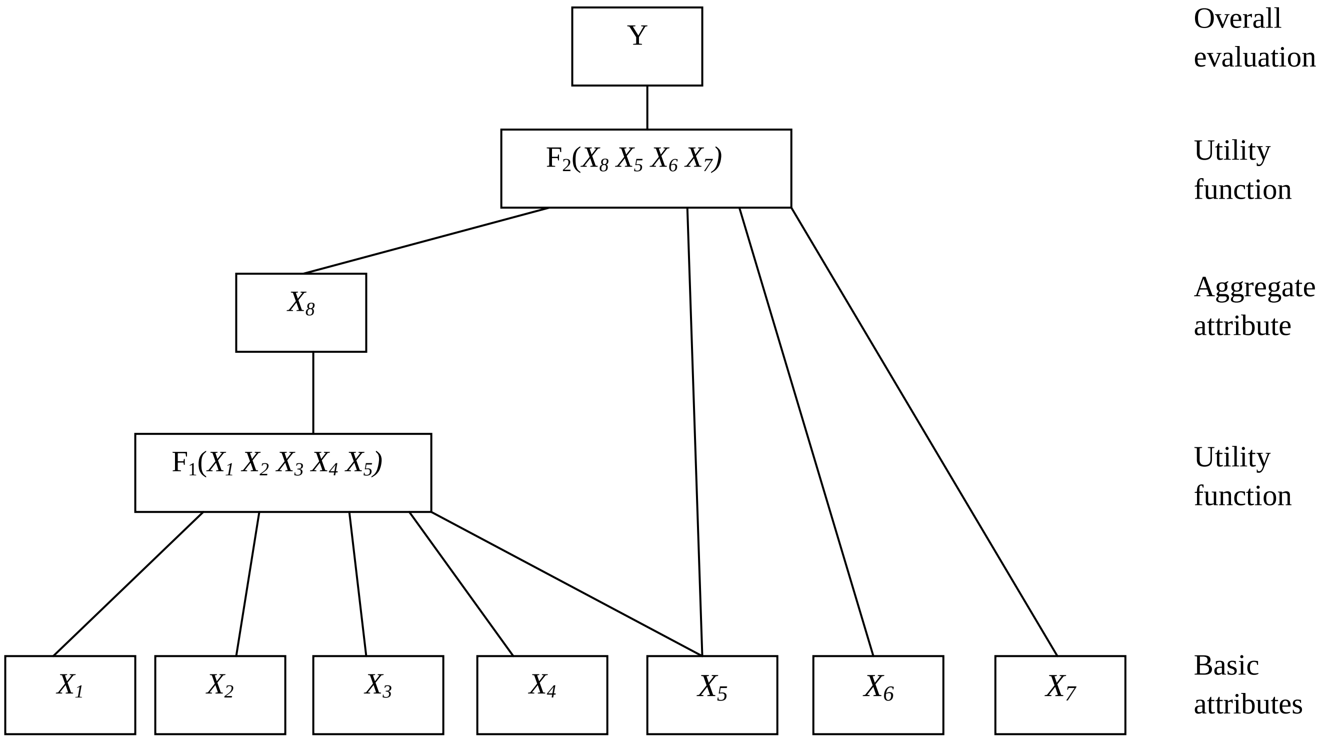


Figure 1: HSDM Example

purpose. A HSDM is composed of attributes X_i (i=1,2…,I) and utility functions F_j(j=1,2,…J). Attributes(sometimes also referred to as performance variables or parameters) are variables that represent decision subproblems. They are organized hierarchically so that the attributes that occur on higher levels of the hierarchy depend on lower level attributes. Fig. 1 illustrates an example of an HSDM where I=8 and J=2. Here F_1, the utility function 1 is modelled by variables X_1 to X_5. From this model an aggregate attribute X_8 is obtained. Utility function F_2 is modeled by attributes X_5 to X_8.Y gives the overall evaluation from the system.

2.5. Automated Statistical Information System (ASIS)

As clearly stated in the introduction, the main objective of this research was to develop a classification rule that will assist the epidemiologist to select a subset of probable patients from recorded viral disease cases and to determine the prognosis of these patients. The end users of the final results are mainly the epidemiologist or the medical doctor who are non-statisticians, so results should be presented in a user friendly format for effective and efficient use.

Therefore, for better presentation of the results, finally, the classification rule developed based on the logistic model was combined with the survival model to build a user friendly ASIS that will output the outcome of the patient when details regarding risk factors are fed into the system. The ASIS includes a Graphical User Interface (GUI) which was developed with Java programming language using NetBeans 6.9 IDE. Apart from the default available libraries in Java, jsc.jar library was imported to obtain required Cumulative Distribution Functions.

The user has to input the required variables and click the submit button in the interface to get the default output which includes whether the patient is a Probable case, Median Survival Time, Hazard Function and Survival Function at Median Survival Time. If the user requires to calculate any percentile other than the Median, it can be done with the Survival Function and Hazard Function at that percentile.

3. An Example

3.1. Available Data

A Data set of electronic medical records on Acute Encephalitis Syndrome (AES) patients for the years 2005 to 2009 were obtained from the Epidemiological Unit, Colombo, Sri Lanka. The variables can be categorized into main sections such as; Particulars of the patient, Present illness/outcome, Clinical Data, Laboratory Data, Japanese Encephalitis (JE) Vaccination status, Information on risk factors, and JE confirmation status. A total of 383 observations were available for analysis. Of these 383 observations, several contained missing values for some of the variables. There were two response variables JE confirmation (binary) and Survival time with the outcome of disease being the censoring indicator. The explanatory variables were province, age, gender, ethnic group, occupation, fever, drowsy, lethagic, coma, meningereal signs, convolutions, headache, nausea, tremors, Fit Definition of AES, WBC(highest count), WBC(Total count), WBC(Neutrophils), Patient Vaccination, Abundance of paddy fields, Presence of piggeries and History of travel.

3.2. Missing Value Imputation

Description of Missingness Situation

In this study, to understand the structure of missingness missing value proportions were analyzed. Table 1 shows the missing percentages associated with each of the variables. The type of variable indicates whether the variable is an outcome (response) variable or an explanatory variable. In table 1, if the 'survival' variable is missing, this indicates that the survival time is missing, while if the 'Outcome of disease' variable is missing the censoring indicator is missing.

Table 1: Percentage and Counts for Missingness.

Variable	Type of Variable	Missing Cases	Complete cases	Percentage Missingness
Outcome of disease	Dead/alive (Censor)	62	321	16.2
Survival time	Response	42	341	11.0
Province	Explanatory	0	383	0.0
Age	Explanatory	4	379	1.0
Gender	Explanatory	3	380	0.8
Ethnic group	Explanatory	24	359	6.3
Occupation	Explanatory	213	170	55.6
JE Confirmation	Response	174	209	45.4
Fever	Explanatory	1	382	0.3
Drowsy	Explanatory	1	382	0.3
Lethargic	Explanatory	1	382	0.3
Coma	Explanatory	1	382	0.3
Meningeal signs	Explanatory	1	382	0.3
Convulsions	Explanatory	1	382	0.3

Headache	Explanatory	1	382	0.3
Nausea	Explanatory	1	382	0.3
Tremors	Explanatory	1	382	0.3
Fit Definition of AES	Explanatory	1	382	0.3
WBC(highest count)	Explanatory	267	116	69.7
WBC(Total count)	Explanatory	238	145	62.1
WBC(Neutrophils)	Explanatory	274	109	71.5
Patient Vaccinations	Explanatory	0	382	0.0
Abundance fields	Explanatory	186	197	48.6
Presenceof piggeries	Explanatory	189	194	49.3
History of travel	Explanatory	197	186	51.4

When handling the missing data in this study, different approaches which best suited the situation were selected based on the percentage missing and missing data mechanism. Initially, variables with missing data were checked for MCAR or MAR by questioning the epidemiologist as to how the missing data occurred.

All variables with missing data were found to satisfy the MCAR or MAR criteria as explained below. Next, the missing percentage of each variable was considered. Casewise deletion which pertains to deleting the entire series of observations (records) having missing values was applied to variables having lower than 5% missingness, where the missingness mechanism could be categorized as MCAR (Harrel, 2001). Variables with around 50% or more missingness were removed and were not considered for imputation since these may produce biased results due to their high missing percentages (Van Burren and Boshuizen, et al, 1999). Based on these rules, casewise deletion was applied to Age and Gender thus deleting 6 records. The Epidemiologist in charge of the reporting procedure explained that these values were missing simply due to medical staff not being able to complete the records. Therefore missingness mechanism could be categorized as MCAR. In one patient’s record, none of the values of the variables were recorded. The epidemiologist emphasized that this may be due to a data entry error and therefore it could be categorized as MCAR and it was decided to delete this record. Now the data set was reduced to 376 from 383. Variables Occupation, Abundance of paddy fields, Presence of piggeries and History of recent travel to endemic areas indicated high missing percentages. While occupation and history of recent travel had marginally over 50% missing, abundance of paddy fields and presence of piggeries had marginally less than 50% missing. The epidemiologist mentioned that apart from medical staff being ignorant, missing values in these variables may be due to the respondent not knowing the answer. Therefore missing mechanism of these variables could be categorized as MCAR. Thus these variables were removed from the analysis considering the fact that if imputed, these may produce biased estimates due to their high missing percentages, and also due to the fact that they are MCAR.

For most of the cases, the laboratory information is available only after further investigations regarding the patient is carried out. However for some cases, laboratory data were recorded prior to further investigation, but due to high missing percentages, it was decided that it would be best to remove these variables. Ethnic group indicated 6% missingness. The reason for this missingness as explained by the doctor was the ignorance of the respondent (patient) or the medical staff. Therefore it could be categorized as MCAR. Listwise deletion is an option in this situation, but as explained by Enders (Enders, 2010), eliminating 6% of data is wasteful. Therefore this variable was selected for imputation.

Missingness of the JE confirmation was due to a had a different reason. The confirmation status was missing mainly for patients on whom further investigation was not done and Laboratory information was not available. MCAR assumption is valid if subjects are randomly selected to undergo more extensive physical examination. However in this situation, patients were selected for further investigation based on the observed symptoms and the rule of thumb used by the epidemiologist according to their experience. Thus missingness in JE confirmation totally depended on the observed symptoms. Therefore confirmation may be missing for a certain subset of patients which indicates that the missing mechanism can be considered as MAR (Rubin, 1987). Therefore JE confirmation status was selected for imputation mainly to avoid nonresponse bias caused by MAR missing mechanism.

Survival time and outcome of the disease indicates missing percentages of 11% and 16% respectively. The epidemiologist stated that the reason for this missingness may be due to administration inefficiencies at the final stage of questionnaire completion and medical staff being ignorant. On the other hand, if the patient is selected for further investigations, then these information will also be completed based on the availability. This indicates that the missing mechanism in these variables is a mixture of MAR and MCAR. Since MCAR is a more restrictive assumption, it is justifiable to consider the missing mechanism of these two variables as MAR. Thus it was decided that these two variables need to be imputed to avoid nonresponse bias and also to improve the sample size.

Simulation Study

As the missing percentage was as high as 45% for JE confirmation a simulation study was considered in order to examine whether it was feasible to use this variable in the study. The simulation study was designed so that the truth is known and there are no missing values. This was to examine the impact of varying levels of missing values of the response JE confirmation which is an important variable used in the ASIS. This would be helpful especially as 45% of JE confirmation is missing and this percentage rate will not be unusual in other diseases as well.

Three correlated binary variables X_1 and X_2 and X_3 are used as explanatory variables. The event probabilities are simulated from a model:

$$\text{logit}(P)=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where β_0 was taken to be 0.5, β_1 was taken to be 0.2, β_2 was taken to be 0.3 and β_3 was taken to be 0.4. The variables X_1 was simulated from a Bernoulli distribution with parameter 0.25, X_2 was simulated from a Bernoulli distribution with parameter 0.6 and X_3 was simulated from a Bernoulli distribution with parameter 0.75, from the event probabilities. The binary observations were obtained by comparing the event probabilities with simulated values from an Uniform[0,1] If the event probability was larger than the uniform variable, then the value was taken as 1 and otherwise as zero. Effect of 0%, 10%, 20%, 30% and 45% missingness in the variables were looked at. A sample of size 400 observations was simulated in line with our sample size of 386. This was done a 1000 times and the correct prediction proportion was averaged for consistency. As in the true situation there was a proportion of 0.8 successes (1's), the cutoff was taken as 0.8. For each case, the response was predicted and compared with the true response. The same model as simulated from was used. For 0% missing there were 69.32% correct predictions, for 10% missing there were 67.37% correct predictions, for 20% missing there were 66.63% correct predictions , for 30% missing there were 64.34% correct predictions and for 45% missing there were 61.91% correct predictions. This clearly indicates that the missing values have only a small impact on the correct prediction percentage. Also a missing percentage of 45% has reduced the correct prediction percentage only from 69% to 62%, which indicates that even with a missing percentage of 45% JE confirmation can be used in the ASIS for a data set of approximately size 400. The simulations were done using SAS version 9.0 statistical package. The SAS code is given as supplementary material.

Multiple Imputation

As explained above, four variables were selected to apply multiple imputation to. These were, namely, Ethnic group, JE confirmation, Outcome and Survival time. Fig. 2 presents the missing data patterns observed in this study for the 4 variables selected for imputation. Dot indicates that the value is missing.

Group	Out	Eth	Conf	Dtd
1	X	X	X	X
2	X	X	X	.
3	X	X	.	X
4	X	X	.	.
5	X	.	.	X
6	.	X	X	X
7	.	X	X	.
8	.	X	.	X
9	.	X	.	.
10	.	.	X	.

Figure 2: Missing Data Patterns

Fig. 2 indicates that the 10 missing data patterns observed in this study do not follow any specific pattern and hence can be categorized as a general or arbitrary pattern. For imputation purposes, SAS macro IVEware which was developed by Raghunathan et. al. Raghunathan et. al, 2001) was used. Conditional regression models used by IVEware depend on the form of the variable that is being imputed. Ethnic group, JE confirmation and outcome are binary variables. Therefore IVEware will use logistic regression and this is the model thought to be appropriate for further analysis. The only continuous variable that was selected for imputation is the survival time. This is the only continuous variable in the data set. For continuous variables, IVEware like most of the other algorithms assumes normality of the variable. Therefore as proposed by Royston (Royston, 2007), a natural logarithm transformation is applied to overcome the non-normality problem after adding one for cases with zero survival time.

The data set consists of another important factor due to the doctor's effect. Therefore a binary dummy variable named as 'doctor's effect' was introduced to the imputation model to account for this factor.

All the variables observed, apart from the variables that were decided to be removed due to the reasons explained above were used in the imputation model. Each imputation was done for 10 rounds as Raghunathan (Raghunathan et al, 2001) suggested and 100 such multiple imputed data sets were created using SAS macro IVEware, and these were averaged.

Comparison of Results

As suggested by Sterne et al. (2009) (Sterne and White, 2009) comparison before and after imputation should be included to determine that no major structural differences after imputation exist. Table 2 gives the proportions of JE confirmation before and after imputation. It can be observed for most of the cases, that the change is approximately lower than 0.1, apart from the continuous variable survival time. In addition, following the imputation of missing responses for variables, a univariate analysis was carried out to determine any changes in the association between the prognostic factors and the response. It could be observed that variables that were significant before the imputation remained more or less unchanged after the imputation.

Table.2: Proportions of JE Confirmation Before and After Imputation

Description	**Level**	**Proportion J/E confirmation**		**Difference**
		Before	After	
Outcome	1-Died	0.2963	0.2281	0.0682
	0-Cured	0.3385	0.3354	0.0031
Ethnic group	Sinhala-1	0.3333	0.3533	-0.02
	Other-0	0.2439	0.1842	0.0597
Doctor's effect	1-(2007-2009)	0.2222	0.1895	0.0327
	0-(2005 & 2006)	0.6	0.5703	0.0297
Gender	1-Female	0.3407	0.2768	0.0639
	0-Male	0.4144	0.3568	0.0576
Province	1-Western	0.5510	0.5111	0.0399
	2-Central,north western,uva	0.4211	0.3293	0.0918
	3-Southern	0.2083	0.2	0.0083

	4-North,estern,north central		0.2667	0.2254	0.0413
	5-Sabaragamuwa		0.3696	0.2530	0.1166
Fever	1-Yes		0.2690	0.2343	0.0347
	0-No		0.6667	0.5889	0.0778
Drowsy	1-Yes		0.2154	0.1803	0.0351
	0-No		0.4599	0.3858	0.0741
Lethargic	1-Yes		0.2222	0.1864	0.0358
	0-No		0.4157	0.3438	0.0719
Coma	1-Yes		0.2889	0.2877	0.0012
	0-No		0.4076	0.3267	0.0809
Managerial signs	1-Yes		0.2286	0.2131	0.0155
	0-No		0.4132	0.3397	0.0735
Convolutions	1-Yes		0.2791	0.2025	0.0766
	0-No		0.4088	0.3502	0.0586
Headache	1-Yes		0.2672	0.2076	0.0596
	0-No		0.5349	0.5071	0.0278
Nausea	1-Yes		0.2333	0.1786	0.0547
	0-No		0.5	0.4327	0.0673
Tremors	1-Yes		0.3333	0.3125	0.0208
	0-No		0.3663	0.3194	0.0469
Fit into the case definition of AES	1-Yes		0.3436	0.3189	0.0247
	0-No		0.5385	0.32	0.2185
Patient Vaccination Vaccination	1-Yes		0.0435	0.0213	0.0222
	0-No		0.4246	0.3617	0.0629
Age(years)	1	Age<=1	0.4545	0.5	-0.0455
	2	Age>1 & Age<=2	0.3077	0.1957	0.112
	3	Age>2 & Age<=12	0.1081	0.1	0.0081
	4	Age>12 & Age<=20	0.4063	0.3065	0.0998
	5	Age>20 & Age<=40	0.4898	0.4146	0.0752
	6	Age>40& Age<=60	0.4483	0.36	0.0883
	7	Age>60	0.5556	0.6111	-0.0555
Survival time	Mean(days)		17.2192	16.45	0.7692

Table 3 shows the p-values for the significance of the different explanatory variables, with respect to JE confirmation before and after imputation. The tests carried out were the Log-Rank test for difference in medians for continuous variables (Survival time), the Pearson's Chi-squared test for the variables with expected values greater than 5 (i.e. Age and Province) and the Fisher's Exact test for the other remaining categorical variables with small expected values.

Table 3: P-values for JE confirmation Comparison Tests Before and After Imputation

Variable	Before Imputation	After Imputation
Doctor's effect	<.0001	<.0001
Fever	<.0001	<.0001
Headache	<.0001	<.0001
Nausea	<.0001	<.0001
Patient Vaccination Status	<.0001	<.0001
Drowsy	0.001	<.0001
Age	0.0064	<.0001
Province	0.0188	0.0001
Fit in to case Definition	0.0285	0.0482
Lethargic	0.037	0.0218
Managerial signs	0.0547	0.0709
Survival time	0.8287	0.7141
Convolutions	0.1567	0.0142
Coma	0.1667	0.5774
Gender	0.3102	0.1206
Ethnic Group	0.3406	0.0056
Outcome	0.823	0.1241
Tremors	0.9999	0.9999

It can be observed from Table 3 that the variables that were considered to be insignificant remained to be so, while those that were considered to be significant remained to be significant apart from the variables Ethnic Group and Convolutions which had become significant after the imputation. This could very likely be due to the fact that since the sample size increased the standard error decreased, hence increasing the power of the analysis.

It is also important to see whether there is any change between survival patterns before and after imputation. Thus the Log Rank Test of checking the existence of significant differences was carried out after the imputation and compared with the results before the imputation. The results are presented in Table 4 below.

Table 4: P-values for Log Rank Test for testing significant difference in survival experience between groups

Description	**Log Rank Test**	
	Before Imputation p-value	After Imputation p-value
JE confirmation	0.829	0.253
Ethnic Group	0.215	0.409
Gender	0.168	0.086
Province	0.130	0.086
Fever	0.357	0.749
Drowsy	0.479	0.505
Lethargic	0.059	0.054
Coma	0.159	0.039
Managerial signs	0.541	0.548
Convolutions	0.002	0.000
Headache	0.000	0.004
Nausea	0.773	0.993
Tremors	0.870	0.998
Fit into the case definition of AES	0.098	0.660
Patient Vaccination	0.689	0.603
Age(years)	0.123	0.000

It can be clearly seen from Table 4 that most of the variables that were significant before the imputation were significant after the imputation and the variables which were not significant before the imputation remained insignificant after the imputation, apart from the variables Coma and Age which have become significant after the imputation. This could very likely be due to the fact that since the sample sizes increase and standard errors decrease after imputation, this results in the increasing of the power of the analysis.

It is interesting to note that no major structural differences could be observed after the imputation. Thus it could be concluded that the imputation is satisfactory.

3.2. Logistic Model Building Procedure

One of the main objectives of this study was to develop a classification method to select probable JE patients for further analysis using a linear logistic model. Therefore as the first step, a logistic model was fitted to identify the relationship between JE confirmation and other risk factors. Stepwise selection method (Agresti, 2007) in SAS was used for variable selection.

The final model selected

The variables for the model were chosen as, Lethargic, Age, Fit into the case definition of AES, Nausea, Drowsy and the interaction term between Nausea and Drowsy. This model can be represented by the following equation,

$$\text{logit}[P_{ijklm}] = \text{contant} + \beta_i^{\text{Leth}} + \beta_j^{\text{Age}} + \beta_k^{\text{Fit}} + \beta_l^{\text{Nau}} + \beta_m^{\text{Drow}} + \beta_{lm}^{\text{Nau}\times\text{ Drow}} \qquad (1)$$

The deviance (Collett, 2002) related to this model was 166.6153 on 234 degrees of freedom. As the p-value associated with this deviance is 0.9967(>0.05) the model fits well. Residual analysis indicated the adequacy of the linear predictor and the absence of large outliers and high leverage values. A test for the adequacy of the link function indicated the adequacy of the logistic link (Collett, 2002).

3.3 Survival Analysis

Another of the main objectives of this study was to identify the factors affecting the survival of Encephalitis patients with the main emphasis on identifying whether there is a significant difference between the prognosis of the JE confirmed patient group and the other viral Encephalitis (OVE) patient group. The Kaplan-Meier Survival curves (Lin and Lee, 2002) of the two groups cross and also the LCH plot (Lin and Lee, 2002) for the two groups of patients showed non parallel lines, indicating the violation of the proportional hazards (PH) assumption (Collett, 2003).This pattern remained unchanged after the imputation as well.

It could be identified from the preliminary analysis that most of the other variables also violated the PH assumption and this was evident after the imputation as well. When PH assumption is violated, then the standard Cox model should not be used. It may entail serious bias and loss of power if used. Therefore as recommended by Collett (2003) [4] a parametric AFT model was considered in this study. A parametric univariate comparison of survival experience between the two groups of patients was also carried out.

3.3.1 Selecting the Appropriate Parametric Distribution

From the probability plot in Fig. 3 it can be observed that the lognormal distribution fits the observed survival time better than other distributions considered but not presented here. The log likelihood test, AIC and BIC Collett (2003) indicate that the lognormal distribution is the most appropriate. Hence in this study the lognormal model was selected as the best distribution to model the observed survival times.

3.3.2 Adequacy of AFT Model: Quantile-Quantile Plot (Q-Q plot)

Q-Q plot Collett (2003), provides an explanatory method for assessing the validity of an AFT model for two groups of survival data. The points fall on a line that is reasonably straight, suggesting that the AFT model would not be inappropriate. However, this conclusion must be regarded with some caution in view of the limited number of points in the graph.

3.3.3 Model Fitting Procedure

Based on the above analysis it was decided to fit a Lognormal AFT model to model the survival times of Encephalitis patients in the presence of the other risk factors in this study. As explained in Collett (2003), a stepwise selection procedure was carried out. Only the interactions with the group indicator variable JE_conf was considered since the main objective was to investigate any significant difference between the two groups of patients and also due to the fact that it will reduce the complexity of the selection procedure and minimize the risk of over fitting.

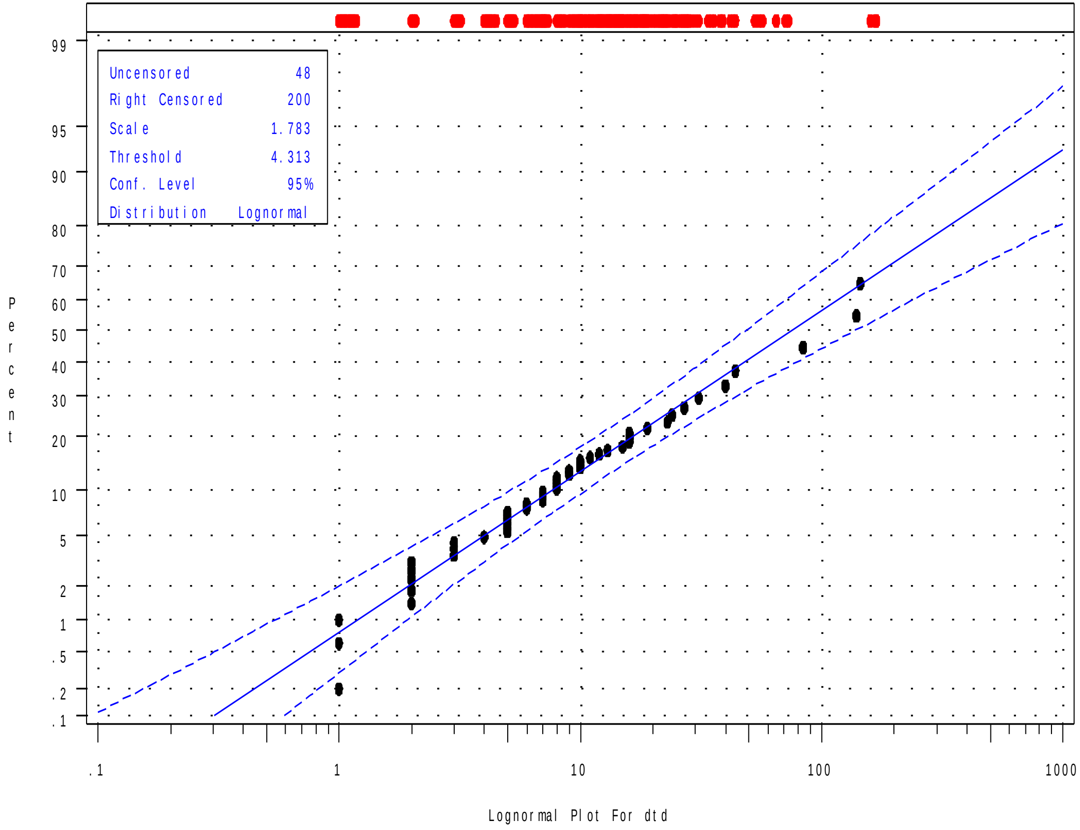


Figure 3: Lognormal Probability plot for observed survival time

3.3.4 Final model with parameters

According to the log linear form of the model, the Lognormal AFT model fitted can be represented by the following equation where T_i is the random variable associated with the survival time of the i^{th} patient.

$$\begin{aligned}\log T_i = \; & 1.8703 - 1.0232{*}Conv_i + 1.6077{*}Head_i + \\ & 1.9067{*}Agem1_i + 2.0603{*}Agem2_i + 2.2040{*}Agem3_i + \\ & 2.2098{*}Agem4_i + 1.3206{*}Agem5_i + 1.016{*}Agem6_i + \\ & 1.0435{*}JE\_conf_i - 0.0573{*}Nau_i - 1.8310{*}Nau_i{*} \\ & JE\_conf_i + 1.5504{*}\epsilon_i\end{aligned} \tag{2}$$

Here ϵ_i has a standard normal distribution, $Conv_i$,$Head_i$, Nau_i and $Agemj_{i,}$ are the values of Convulsions, Headache, Nausea and the Age Category for the i^{th} individual, and JE_conf is zero if the patient belongs to the OVE group and one if the patient the belongs to JE group.

3.4. The ROC Curve

ROC curve analysis was used to identify the best cutoff point for the classification test developed based on the logistic model. The cutoff point is chosen so that the sensitivity is high and the specificity is also of acceptable value. As explained clearly in the introduction, the main purpose of developing a classification test in this study is to correctly identify probable JE patients. This is evaluated using the sensitivity of the classification test. Therefore while sensitivity should be higher specificity should also be of an acceptable value since incorrectly classifying other viral encephalitis patients as JE will result in a waste of time and resources in conducting further investigations. Therefore to satisfy both criteria, the cutoff pertaining to a sensitivity value of 0.83 and a specificity value of 0.74 was selected. This resulted in a cutoff value of 0.13967 as the best trade off.

3.5 Validation Procedure for Lognormal AFT Prognostic Model for Overall Survival

The predictive performance of the lognormal AFT model was evaluated using the c-index and slope shrinkage respectively. These measures were calculated using R version 2.13.1 by using the package “rms” (regression modeling strategies) developed by Harrell (Harrell, 2009) A slope-shrinkage of 0.7852 was obtained. This is somewhat close to 1 indicating little evidence of over-fitting. A reasonably large value for the C-index of 0.74505 was obtained indicating good predictive discrimination. Therefore it can be concluded that the set of prognostic factors explain the variation in the outcomes reasonably well, and this implies good prediction for individual patients.

3.6 Hierarchical Statistical Decision Model (HSDM) for the Data

The end users of the final results are mainly the epidemiologists or the medical doctors who are non-statisticians, so results should be presented in a user friendly format for effective and efficient use. Therefore for better presentation of the results, the classification rule developed based on the logistic model was finally combined with the survival model using a Hierarchical Statistical Decision model (HSDM) [18] to aid in building a user friendly ASIS that will output the outcome of the patient when details regarding risk factors are fed into the system. The HSDM for this study is illustrated in Fig. 4.

3.7 Encephalitis Patient Information System

The following section gives a description of how to use the Automated Statistical Information System which is named as “Encephalitis Patient Information System”. The user has to input the required variables and click the submit button in the interface to get the default output which includes whether the patient is a probable case of JE, Median survival Time, Hazard Function and Survival Function at Median Survival time, and if the user requires to calculate any percentile other than the Median, it can also be calculated with the Survival Function and Hazard Function at that percentile. Fig. 5 shows the developed ASIS. Initially it is unfilled without any values. The top part of figure 5 shows the values input by the user for the required variables. When the submit button is clicked in the interface, the default output shown at the bottom of figure 5 is obtained.

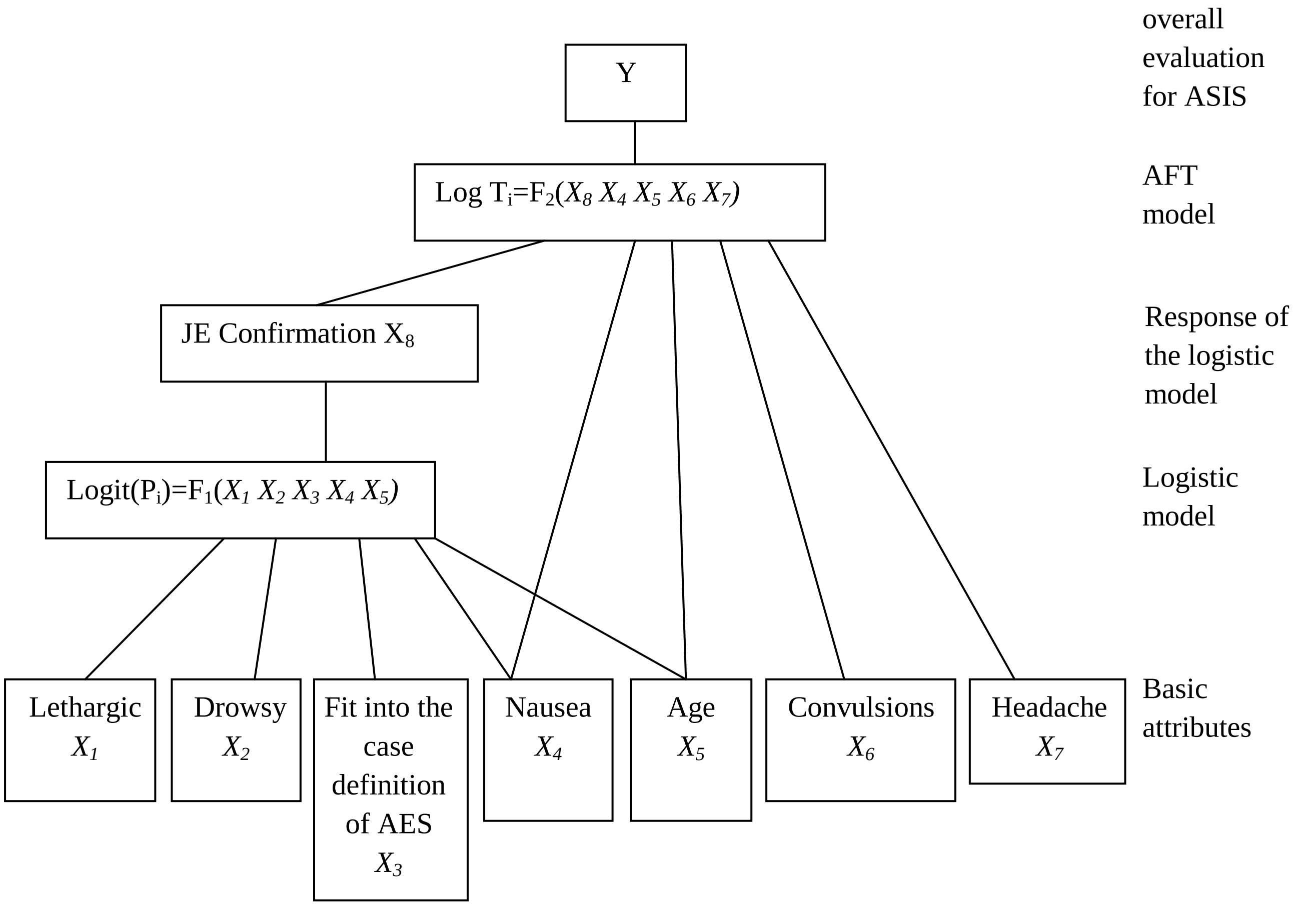
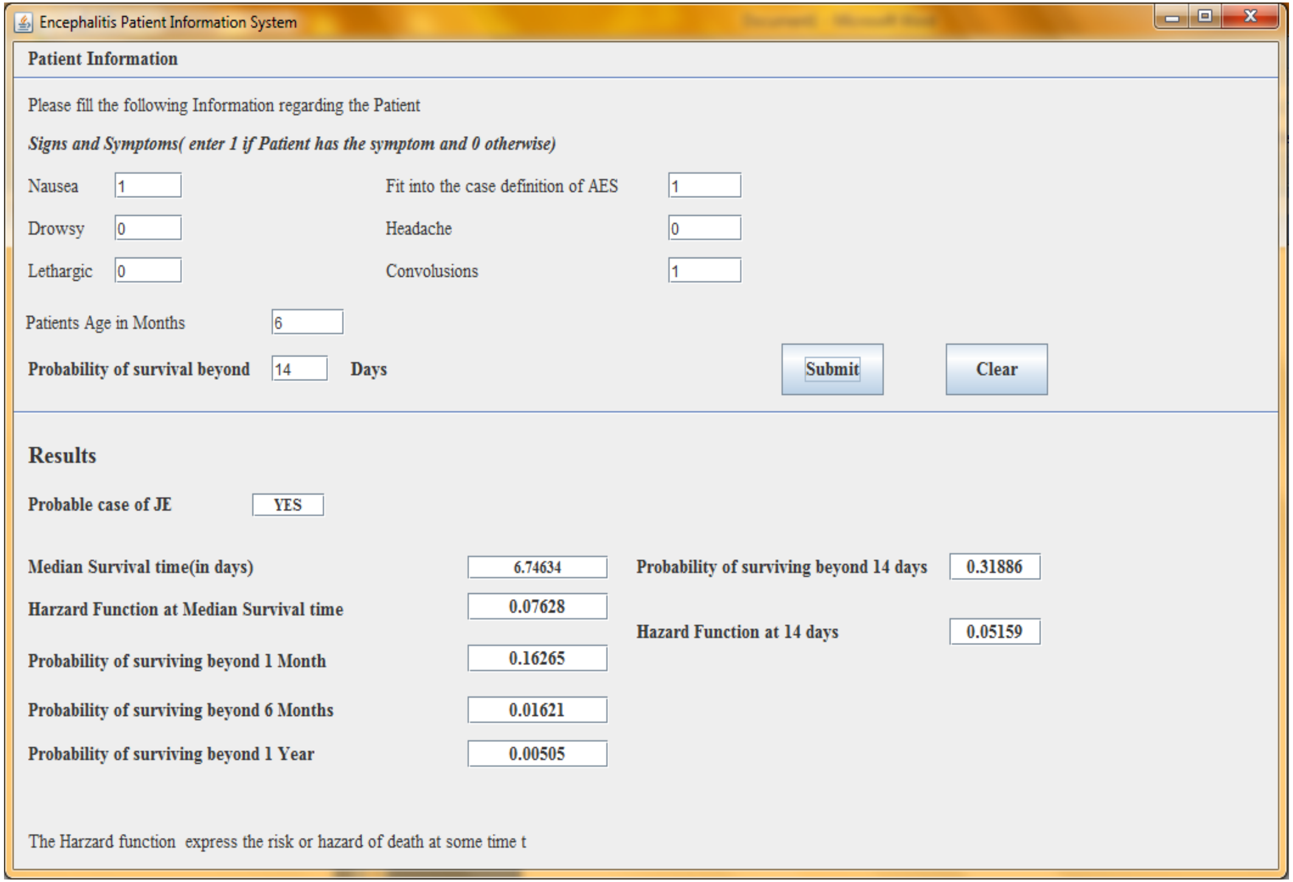

Y
overall evaluation for ASIS
Log T_i=F_2(X_8 X_4 X_5 X_6 X_7)
AFT model
JE Confirmation X_8
Response of the logistic model
Logit(P_i)=F_1(X_1 X_2 X_3 X_4 X_5)
Logistic model
Lethargic X_1
Drowsy X_2
Fit into the case definition of AES X_3
Nausea X_4
Age X_5
Convulsions X_6
Headache X_7
Basic attributes

Figure 4: HSDM for this Study



Encephalitis Patient Information System

Patient Information

Please fill the following Information regarding the Patient

Signs and Symptoms(enter 1 if Patient has the symptom and 0 otherwise)

Nausea 1
Drowsy 0
Lethargic 0

Fit into the case definition of AES 1
Headache 0
Convolusions 1

Patients Age in Months 6

Probability of survival beyond 14 **Days**

Results

Probable case of JE YES

Median Survival time(in days) 6.74634
Harzard Function at Median Survival time 0.07628
Probability of surviving beyond 1 Month 0.16265
Probability of surviving beyond 6 Months 0.01621
Probability of surviving beyond 1 Year 0.00505

Probability of surviving beyond 14 days 0.31886
Hazard Function at 14 days 0.05159

The Harzard function express the risk or hazard of death at some time t

Figure 5: The GUI, the data input and final output

4. Discussion

Countries are requested to report cases of disease to the WHO, but official notifications substantially underestimate the incidence of the disease. In some countries, lack of data can lead to the conclusion that the disease is not a problem. Other countries report all clinically suspected disease cases, which may overestimate the burden of the disease (http://aim.path.org/). One important result of this study is that, it enables epidemiologists to provide answers to the above problems to a certain extent. The classification rule developed in this study can be used by doctors to select a subset of probable diseased cases for further investigation so that maximum benefit can be obtained from limited funds and laboratory resources. It will also assist the epidemiologist to select a subsample of viral disease cases whose disease confirmation is not recorded, in order to carry out further investigations. Further, the results of the survival model can be used to determine the prognosis of patients and to compare this with the prognosis of viral fever patients.

The work carried out in this paper is important as medical practitioners and epidemiologists can now have an easily usable, interpretable and attractive system to determine both the diagnosis and the prognosis of life threatening viral disease patients. This will help them to obtain an initial understanding of the patient before further investigation. Another novelty of this research is the use of multiple imputation to resolve the problem of non-response bias. In modeling the diagnosis and prognosis, the correlation between the two was not taken into account. As further work we suggest the development of bivariate models for this purpose.

Acknowledgements

The authors are grateful to Dr. Paba Palihawardene, Chief Epidemiologist and Dr. T. S. R. Peiris , Asst. Epidemiologist of the Epidemiological unit for giving permission to access the data set collected by Epidemiological Unit regarding surveillance of Encephalitis Patients in Sri Lanka. The authors would also like to thank Dr. Romaine Jayawardene, of the Mathematics Department of the University of Colombo for editing this manuscript.

References

1. *Advanced Immunization Management:Assessing disease burden: Data for decision making.* (2009). Retrieved may 12, 2011, from aim.path: http://aim.path.org/ .

2. Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* New Jersey: John Wiley & Sons, Inc. DOI : 10.1002/0470114754
3. Bohanec, M., Zupan, B. et al. (2000). Applications of qualitative multi – attribute decision models in health care. *International Journal of Medical Informatics , 58-59,* 191-205. DOI : 10.1016/S1386-5056(00)00087-3
4. Clark, T. G., Bradburn, M. J. et al. (2003). Survival Analysis Part IV: Further concepts and methods in survival analysis. *British Journal of Cancer , 89,* 781 – 786. DOI : 10.1038/sj.bjc.6601117
5. Collett, D. (2003). *Modeling Survival Data in Medical Research.* New York: Chapman & Hall/CRC.
6. Collett, D. (2002). *Modeling Binary Data.* New York: Chapman & Hall/CRC.
7. Cox, D.R.(1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 34, No. 2. Pages 187-220.
8. Department of Vaccines and Biologicals of World Health Organization. (2003). *WHO-recommended Standards for Surveillance of Selected Vaccine preventable Diseases.* Geneva: WHO.
9. Enders, C. K. (2010). *Applied Missing Data Analysis.* (T. D. Little, Ed.) New York, United States of America: Guilford Publications
10. Gan, F., Koehler, K., et al. (1991), Probability Plots and Distribution Curves for Assessing the Fit of Probability Models. *The American Statistician.* 45(1) 14-21. DOI : 10.1080/00031305.1991.10475759
11. Greenland, S., & Finkle, W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology , 142* (12), 1255-1264
12. Harrell, F. E. (2001). *Regression Modeling Strategies With Applications to Linear Models,Logistic Regression and Survival Analysis.* New York: Springer-Verlag
13. Harrell, F. E. (2009). *Usser mannual of the Package 'rms'.* Retrieved June 2011, from http://biostat.mc.vanderbilt.edu/rms.
14. Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* United States of America: John Wiley & sons, Inc.
DOI : 10.1002/0471722146

15. Janssen, K. J., Donders, A. R. et al. (2010). Missing covariate data in medical research: to impute is better than to ignore. *PubMed , 63* (7).

16. Lin, W. Y., Lee, L. T., et al. (2002). Optimal cut-off values for obesity: using simple anthropometric indices to predict cardiovascular risk factors in Taiwan. *International Journal of obesity , 26* (9), 1232-1238.
DOI : 10.1038/sj.ijo.0802040

17. Little, R.J.A. and Rubin, D. B (2002). *Statistical Analysis with missing data.* Wiley publications. DOI : 10.1002/9781119013563

18. Meeker, W.Q. and Escobar, L.A. (1998), Statistical Methods for Reliability Data. John Wiley and Sons.

19. Raghunathan, T. E., Lepkowski, J. M., et al.(2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology, Statistics Canada , 27* (1), 85-95.

20. Royston, P. (2007). Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *The Stata Journal , 7* (4), 445–464.

21. Rubin, D. D. (1987). *Multiple Imputation for nonresponses in Surveys.* Hoboken,New Jersey: John Wiley & sons.
DOI : 10.1002/9780470316696

22. Sterne, J. A., White, I. R., et al.(2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal.* DOI : 10.1136/bmj.b2393

23. Tabata, S., Shinichiro, Y et al.(2009). Waist circumference and insulin resistance: a cross-sectional study of Japanese men. *BioMed Central Endocrine Disorders , 9* (1), 1472-6823.

24. Van Buuren, S., Boshuizen, H. C. et al. (1999), Multiple imputation of missing Blood pressure covariates in Survival Analysis. *Statistics in Medicine , 18,* 681-694.
DOI : 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.3.CO;2-I

25. Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research , 16,* 219–242.