~ CHAPTER 3 ~

DATA AND METHODOLOGY

3.1 Overview

This chapter elucidates and elaborates on the data as well as the various statistical techniques employed in this study. It addresses issue of missing data which has greatly impaired the quality of the data herein. The generation of new variables is further explained. The chapter also describes the procedure for testing the relevant assumptions, estimating the model, assessing the fit and diagnostic method for the multiple logistic regression central to the paper.

3.2 Data

The dengue data studied in this paper was originally collected by the Department of Pediatrics, Faculty of Medicine, University of Malaya, Kuala Lumpur. The data collection took place at the University Malaya Medical Centre (previously known as the University Hospital) during a dengue epidemic which lasted for about three and a half months from mid-July to October 2002. The collection did not follow any specific sampling procedure. It basically covered all the out-patients and in-patients, adults and children (aged 12 and below) suspected of dengue infection during the said period. A total of 53 variables with 734 observations were collected in this study. Of these variables, 35 of them were analyzed. The description of the variables used in this paper is provided in Table 3.1.

Table 3.1: Listing of variables and description

No.	Variables	Description
1	Sex	Gender of male or female
2	Race	Ethnicity of Malay, Chinese, Indian or others
3	Admission	Patient admitted or not
4	Age	Age of patient
5	Child or Adult	Child aged 12 and below or adult otherwise
6	Length of Stay	Length of stay at UMMC in days
7	Fever duration	Duration of fever in days
8	Fever	Fever (yes or no)
9	Vomit	Vomiting (yes or no)
10	Giddy	Giddiness (yes or no)
11	Headache	Headache (yes or no)
12	Skin Rash	Skin rashes (yes or no)
13	Eye Pain	Pain behind the eyes, retro-orbital pain (yes or no)
14	Muscle & Joint Pain	Muscle or joint pain, myalgia or arthralgia (yes or no)
15	Bleeding	Haemorrhagic evidence (yes or no)
16	Shock Evidence	Evidence of shock or poor perfusion (yes or no)
17	Heart Rate	Heart rate per minute
18	Hepatomegaly	Palpable liver (yes or no)
19	Rash / Petechiae	Petechiae test for rash (positive or negative)
20	Abdominal Pain	Abdominal pain (yes or no)
21	Dehydration	Body dehydration (yes or no)
22	Hematocrit change	Percentage change in hematocrit.
23	Haemoconcentration_20	Hematocrit changes 20% or greater
24	Haemoconcentration_50	Hematocrit changes 50% or greater
25	Platelet count at admission	Count of platelet cells in thousand upon admission
26	Thrombocytopenia_100	Platelet count of 100,000 cells per mm^3 or less
27	Thrombocytopenia_50	Platelet count of 50,000 cells per mm^3 or less
28	Serology Test	Dengue serology test done / not done
29	First Serology Test	First dengue serology test result positive or negative
30	Second Serology Test	Second dengue serology test result positive or negative
31	Final Serology Test	Combined first and second test result
32	Notification	Cases notified as DF or DHF or not notified
33	Clinical Diagnosis	Clinical Diagnosis as DF, DHF or DSS
34	Clinical Diagnosis_2	Clinical Diagnosis as DF or DHF
35	WHO Diagnosis	Diagnosis of DF or DHF as per the WHO guidelines

The management and analysis of the data were performed with the help of SPSS (Statistical Package for the Social Sciences) for Windows, Release 11.5.0. Microsoft Excel 2002 was employed in some simple spreadsheet analysis.

3.3 Missing Data

The dataset suffers a great amount of missing data which render many cases and variables unusable. Variables with severely missing data were excluded from the analysis. These variables are *Family with Confirmed Dengue*, *Neighbour with Confirmed Dengue*, *Relationship of Family Member with Confirmed Dengue* and *Virus Serotype*. For subjects with missing data, the computer program selectively processed subjects with complete data only when performing the multivariate analysis.

A few categorical variables were almost constant in their response with more than 99% of the response in one category. These variables were removed from the analysis and they are *Excess Thirst, Reduced Urine Output, Plasma Leakage, Hess Test, Overweight, Abdominal Consciousness, Jaundice, Pleural Effusion, Ascites, Postural Hypotension* and *Hypoprotenaemia.*

Variable *Heart Rate* contained a huge block of observations with zero reading. It was assumed that the heart rate of these patients was not taken (instead of having zero heart beat) and hence was treated as missing data. One adult patient with heart rate of 267 per minute was treated as outlier since the maximum heart rate for human was estimated at 220 per minute less the age of the patient (Wikipedia, 2005).

Variables *Place of Admission*, *Platelet Count at Discharge* and *Follow Up with Patient* are not examined here as they are faintly relevant to the study objectives.

3.4 New Variables

A few new variables were created using the existing variables in the dataset. Variable *Age* was created using the date of birth by subtracting it from the year 2002, which is the year the data collection took place. *Fever Duration* was categorized into a two-group variable (*Fever*) comprising those with or without such symptom. Variable *Hematocrit Change* was created with the information on the high and low of patient's hematocrit readings. The said variable was then translated into two new variables, *Haemoconcentration_20* and *Haemoconcentration_50* dichotomized at hematocrit changes of equal to or greater than 20% and 50% respectively. Variables *Thrombocytopenia_100* and *Thrombocytopenia_50* were created using information on the platelet count at admission categorized at the level of 100,000 cells per mm^3 and 50,000 cells per mm^3 respectively. Variable *WHO Diagnosis* of DF and DHF was created based on the information on *thrombocytopenia_100* and change of hematocrit (greater than 20%) as per the WHO case definition. Due to the non-significance of the difference between DSS and DHF and the fact that the former is a severe form of DHF, variable *Clinical Diagnosis_2* was created by collapsing the category of DSS into DHF, resulting in only two categories – DF and DHF. *Final Serology Test* was created by taking into consideration the first and second results of the dengue serology test.

3.5 General Statistical Method

The characteristics and shape of the distribution for all quantitative variables were examined through their respective histograms as well as the descriptive statistics. Boxplot and scatterplots were used to identify potential outliers. The former graphical method was also useful in examining group differences.

Multicollinearity between numerical variables was examined using the Pearson's correlation and the bivariate scatterplots of the variables in question.

Pearson's Chi-square (with continuity correction for 2x2 table) was computed to understand the relationship between two categorical variables of interest. Such measure of association was frequently used throughout the paper since most independent and dependent variables of interest were dichotomous.

Odds Ratios with 95% confidence intervals were also computed in attempt to understand the relative risk for a given clinical feature. They also provide an avenue to comprehend how much likely children are to suffer certain symptoms compared to adults.

The t-test was carried out to compare the means of the numerical variables for two groups. Prior to running such test, the normality assumption was assessed via the Kolmogorov-Smirnov test, while the homogeneity of the group variances was examined through the Levene's test. Whenever the assumption was violated, the non-parametric

equivalent of the two-independent-sample t-test, the Mann-Whitney U test, was use instead.

To compare the means of more than 2 groups, one-way Analysis of Variance (ANOVA) was employed wherein the sources of error were examined. In cases where at least two group means were found different, Tukey, Scheffe and Dunnet's post hoc multiple comparison tests were carried out to identify the pair of group means that differed significantly. Levene's test was executed at the very first step of this process in assessing the homogeneity of group variances.

3.6 Multiple Logistic Regression

The multiple logistic regression method is appropriate for many analyses in this paper given that all the dependent variables considered in this study are dichotomous. According to Hair et al. (1998), the logistic regression is not restricted by the assumptions of multivariate normality and equal variance-covariance matrices across groups. In fact, it is much more robust when the two assumptions are not met. Therefore, no remedy or transformation of variable was required for this study.

Although the discriminant analysis – a familiar alternative to logistic regression – is also capable of handling categorical dependent variables, it is not suitable here given the use of categorical independent variables in this paper which may cause problem with the variance-covariance equalities – an assumption of the method (Hair et al, 1998).

In meeting the second objective of this paper, the method of multiple logistic regression is employed to construct a parsimonious model consisting of independent variables that can significantly predict and classify the incidence of dengue serology test outcome. In this case, the dependent variable of the logistic regression assumes the binary outcome of the *final serology test* result; coding of 1 for positive and 0 for negative dengue infection.

For the classification of clinical DF and DHF (third objective of the paper), the outcome variable was *Clinical Diagnosis_2* wherein the clinical DF cases were coded 0 while the DHF as 1. The logistic regression method allows the construction of a model that can classify the suspected dengue patients into clinical DF and DHF based on a set of predictor variables significant in differentiating DHF from DF.

The model estimation strategy for the logistic regression essentially centers on the criteria of reducing the log likelihood ratio. In selecting potential independent variables for inclusion in the logistic model, the univariate likelihood ratio test was performed for each independent variable and those with significance of 25% or less are selected as the potential candidates in the subsequent stepwise procedure (Hosmer and Lemeshow, 2000). The Score statistic, which tests if a coefficient is different from zero based on the change in the log-likelihood associated with the effect, is the variable selection criterion in the stepwise procedure. Variable with the highest Score statistic was chosen for entry into the model. The remaining independent variables were then examined for potential inclusion. At each step, the change in the -2 log-likelihood value from the previous step

was assessed using the Chi-square test. The process stopped when the change in the log-likelihood was no longer significant.

Once a model with the main effects was obtained, plausible interactions between the main effects were tested.

For a given set of value for variable X_i, where $\hat{\beta}_i$ is the estimated regression coefficient for effect *i* and α the constant, the predicted risk for positive dengue infection (Y=1) can be obtained from the fitted logistic model as follows:

$$\hat{P}(Y=1 \mid X_i) = \left[1 + e^{-(\alpha + \sum \hat{\beta}_i X_i)} \right]^{-1}$$

Controlling for other independent variables in the equation, the risk odds ratio for effect *i* can be calculated as such:

$$\text{Risk Odds Ratio (ROR)} = e^{\hat{\beta}_i (X_{1i} - X_{0i})}$$

where X_1 and X_0 are (1,0) for categorical independent variable or numerical value for metric variable.

The overall model significance was examined using the Chi-square test for the change in the -2 log-likelihood value from the base model. In addition, the Hosmer and Lemeshow measure of overall fit which tests the null hypothesis of no difference between the observed and predicted classification was also used in assessing the significance of

the final model. Various R-square measures such as the Cox and Snell R-square and the Nagelkerke R-square were used. Throughout the whole process, Wald statistic was used to assess the significance of the individual estimated regression coefficients as well as the constant.

Classification accuracy of the model was compared against the maximum chance criterion and proportional chance criterion for unequal group size (Hair et al., 1998). The maximum chance criterion is determined by calculating the percentage of the larger sample size of the two groups in the total sample. The proportional chance criterion is given by:

$$C_p = p^2 + (1-p)^2$$

where C_p is the proportional chance, p is the proportion of subject in a specific group.

In addition, the discriminatory power of the classification matrix was also assessed via the Press's Q statistic given by:

$$\text{Press's Q} = \frac{[N-(nK)]^2}{N(K-1)}$$

where N is the total sample size, n is the number of observations correctly classified and K is the number of groups. The statistic, which compares the number of correct classifications with the total sample size and the number of groups, is distributed as Chi-square with 1 degree of freedom (Hair et al., 1998).

Besides looking at the sensitivity and specificity of the classification performance, the Receiver Operating Characteristic (ROC) curve was drawn in order to better

understand the discriminative power of the logistic model. The area under the curve essentially provides "a measure of the model's ability to discriminate between subjects who experience the outcome of interest versus those who do not" (Hosmer and Lemeshow, 2000).

Model diagnostic was carried out to detect and remove any influential observations in order to improve the estimated model. The measures for such purpose were the studentized residuals, hat values, DFBETA and Cook's Distance. The threshold value specifications for the said measures are given in Table 3.2 (Hair et al., 1998).

Table 3.2 Threshold value specification for the diagnostic test

Diagnostic Measure	Threshold value specification
Studentized Residuals	Critical t value at specified confidence interval
Hat values	Medium to large sample: $2(k+1) / n$
Cook's Distance	$4 / (n - k - 1)$
DFBETA	$2 / \sqrt{n}$

Note: k is the number of independent variable, n is the sample size.

3.7 Other Consideration

In classifying the suspected dengue cases into DF and DHF, children and adults were modeled separately in the logistic regression due to the symptomatic differences between the two groups as per the literature.