Chapter 4

General Discussion and Conclusions

Detailed discussion of the findings was presented in two previous manuscripts (Appendices 1 and 2). The results of the two previous manuscripts require synthesis and interpretation. Therefore, this chapter includes a general discussion including statistical methods used, overall findings, implications for further research and for health policies, and advantages and limitations of the studies. In addition, conclusions and recommendations are given.

4.1 General Discussion

4.1.1 Statistical methods used

In this dissertation, different statistical models were used according to each outcome characteristic and data structure. The reasons, advantages, disadvantages and limitations of each statistical method are discussed in the following sections.

4.1.1.1 Manuscript I

4.1.1.1.1 Negative binomial model

Most public health data are analysed with a focus on disease counts, proportions or rates as an outcome variable rather than a continuous outcome. The large counts or rates are more likely to follow the assumptions of linear models whereas spatial analyses tend to focus on counts from small areas with rather few cases. Thus, appropriate models for count or rate outcomes were required (Waller and Gotway, 2004). The increased computational abilities available at the present time enable the users to choose the appropriate available modeling approaches (Hu et al, 2003).

In the first part of our study, Poisson modelling was first considered for forecasting the count of deaths from HIV/AIDS and other infectious diseases as it is commonly used for analyzing count data with Poisson distribution. This model is strictly on the equality of variance and mean. In our data, many zero counts of deaths occurred. The Poisson model was not appropriate for these data due to over-dispersion. The negative binomial model, which generalizes the Poisson model by allowing for over-dispersion, was then considered. Autoregressive terms were considered as two-month periods for 3 periods, which were lag 1-3. Dormann et al (2007) reported that regions close to each other show more similar values than those further apart. Therefore the lag terms for the previous two months in neighbouring districts were taken into account in the model. Also, the higher incidence of death rates among males in the previous two months may affect the rates among females and vice versa. Gender, place of death, two-month period, lag 1-3, spatial and gender correlations were included in the HIV model where diagnosis (unspecified septicemia or other infectious diseases), two-month period, lag 1-3 and spatial correlation were included in the other infectious disease model. The negative binomial with lag observation-driven GLM can be used for accurately forecasting the incidence of mortality for the next period in this study. This method is a reasonable statistical method to use for diminishing spatial-temporal correlation before calculating the actual incidence mortality rates. However it is not so useful for identifying risk factors for historical prevalence. This is because when lag observations are included in the model the error variance is reduced substantially, but the outcome is no longer the prevailing mortality rate but the incidence of new cases.

If the lag observations were not considered, another statistical method for analysing such data and taking account for the spatial correlation between districts is the generalised estimating equation (GEE) model (Carl and Kühn, 2007). This approach is appropriate for parameter estimation rather than prediction (Augustin et al, 2005). The use of this method for spatial data has also been reviewed by Dormann et al (2007).

The other approach is the spatial-temporal autologistic regression model for binary data which was used by Zhu et al (2008) for predicting mountain pine beetle outbreak. However, fitting this method is time consuming and does not correct for available degrees of freedom (Zhu et al, 2008; Carl and Kühn, 2007). Carl and Kühn (2007) suggested not using autologistic regression for taking account of spatial autocorrelation. For the assessment of risk factors of the time series data of the haemorrhagic fever with renal syndrome (HFRS) transmission, Hu et al (2003) compared three regression models. They were: least squares linear time series model, generalized linear model with a Poisson link time series model and generalized additive model with Poisson link time series spline model. He suggested that the last model may be the most suitable but it needs to be examined in future research. All of the above methods were found to be appropriated for parameter prediction but not for forecasting.

4.1.1.2 Manuscript II

Several statistical methods have been used for analysis of the association between determinants and LOS. The most appropriate statistical method for analyzing length of hospital stay is still questionable as many studies gave different results. The problems for LOS data are almost the same as for medical expense data. Three characteristics of medical expense data that need to be accounted for in modeling are a large numbers of zero, the highly positive skewed distribution and the frequently violated homoscedasticity (Blough and Ramsey, 2000; Xie et al, 2004). The model for predicting LOS with less prediction error is commonly considered as a suitable model for analyzing LOS data. The correctness of predicting LOS is important for managing hospital resource utilization, promoting effectiveness of the health care system and comparing across institutes or countries (Austin et al, 2002; Kulinskaya et al, 2005). Thus model selection is a significant step. In the second part of our study, linear model and logistic model were used. The reasons for, and advantages and disadvantages of, using these two methods are discussed as follows.

4.1.1.2.1 Linear model

Even though linear model is not frequently used in epidemiology studies, it is still required to analyze for continuous outcome. Hanson (1973) tested log-normal model in 3 different hospitals on the data of length of hospitalization of mental patients and found that data from several samples of patients supported the log-normal function. Marazzi et al (1998) examined the adequacy of models based on lognormal, Weibull and gamma distributions. He recommended that the lognormal model was most appropriate, but cases with LOS less than one day were omitted from their analysis because of computational problems when the LOS is zero. Another approach was carried out by Xie et al (2004) to examine three kinds of models, linear regression, logistic regression and mixed distribution models (logistic regression for admitting or not admitting to the hospital and linear regression for length of stay in hospital) for fitting length hospital stay of severe mental illness patients. They concluded that the mixed-distribution model provides greater specification to fit these data and leads to better interpretation of the results. The examination of alternative estimators for log model in term of bias and precision had been studied by Manning and Mullahy (2001).

Several statistical methods apart from lognormal model had been suggested by many authors. Austin et al (2002) compared seven different statistical strategies for analyzing LOS in a cohort of patients undergoing coronary artery bypass graft surgery (CABG) including regression; linear regression with log-transformed length of stay; generalized linear models with Poisson, negative binomial, normal and gamma distribution, and semi-parametric survival models. All model had similar ability to predict LOS in the actual data, except for Cox regression. He suggested that generalized linear models were best able to predict patient length of stay. Such models were performed under Monte Carlo simulations. Whereas Kulinskaya et al (2005) compared standard general linear model (GLM) with truncated maximum likelihood (TML) methods. Before performing standard GLM methods, they excluded outliers using 1.5 times interquartile range rule and then transformed data by taking logarithm. TML methods computed highly robust estimated and rejected observations that were unlikely under the estimated model. After that the maximum likelihood estimate was computed with the retained observations. This study revealed that TML methods provided better model fit and accuracy of parameter estimation than GLM methods. However TML methods are not commonly known and used by general statisticians. Moreover, these methods have not been validated by other kinds of data.

It is commonly known that the distribution of LOS is usually positively skewed. Thus the residuals could violate the normality assumption when fitting the linear model without any transformation of the outcome. In the second part of this thesis, LOS was transformed by adding a constant (chosen to be 0.4) and taking the natural logarithm transformation before fitting the linear model. The selected constant was the value which provided the best residual plot. From the residual plot, the residuals departed from the line corresponding to the normality assumption for LOS equal to zero (actually less than 1 day) but the other values of LOS were located along the line. Therefore, the normality of residuals for LOS greater than zero was satisfied. Thus, the model from linear regression with log-transformed LOS in this study provided an adequate fit. However, the errors for prediction LOS were large for zero LOS. This means that zero LOS may not be accurately predicted. The advantage of this approach is that the distribution of LOS tends to be normalized by taking its logarithm and it is the predominantly used model for analyzing positive skewed data (Austin et al, 2002). Linear regression with log-transformed LOS assumes that LOS has lognormal distribution and the variance of a lognormal distribution increases when the mean of the distribution increases. Some disadvantage of this method can also be illustrated. Austin et al (2002) mentioned that if the logarithm of LOS of the patient is heteroscedasticity, then the result from lognormal model will provide the bias estimation (Austin et al, 2002). This is confirmed by Jia et al (2008) and Ai and Norton (2000).

In this study, predicted LOS from the linear model was then used for calculating the total LOS for each disease group and for each demographic factor group. This calculation provided a broader range of information result and it can indicate the severity of the problem.

4.1.1.2.2 Logistic model

Logistic regression is widely used in epidemiology studies with a binary outcome for identifying risk factors associated with the occurrence of outcome (Pfeiffer et al, 1997; Anderson, 2003). The use of logistic regression has been reviewed by many authors such as Peng et al (2002), Anderson et al (2003) and Bewick et al (2005).

LOS not only can be used as continuous outcome, but can also be considered as binary outcome, such as in studies conducted by Wen et al (1998), Spencer et al (2004) and Van den Block (2007) where they divided LOS into two groups. In our study, LOS was divided into two groups: LOS less than 7 days and LOS at least 7 days. With use of appropriate structure of the data the computational time can be substantially reduced (Panagiotakos and Pitsavos, 2004). Therefore, the data were grouped by principal diagnosis, age, gender, region and hospital size, reducing 40,498 individual cases to 1,026 records. As the data were not case-by-case data, the model could not accurately predict individual LOS. However, the logistic model has the advantage over the linear model in that no distributional assumption is needed. The transformation of binomial probability of the outcome of logistic regression is the easiest way to interpret results (Bewick et al, 2005).

Recoding a continuous outcome data into a dichotomy for analysis with logistic regression is one common approach; however performing this procedure abandons the important information (Xie et al, 2004). There are several aspects of using logistic regression that need to be considered. First, the quantitative information of the data is misplaced by using this method. Second, the explanatory variables should not be highly correlated with one another because this could cause problems with estimation (Bewick et al, 2005). Third, sample size should be large enough. A high sample size is especially needed where there is a high number of a determinant. In addition, this model assumes that all of the population in each category of determinants is homogeneous (Fienberg et al, 1985).

4.1.2 Overall findings

In the first part of the study, we used data from vital registration. Mortality from HID/AIDS and unspecified septicemia were considered separately. In the second part, we used data from National Security Health Office and focused on length of hospital stay of patients dying in hospital.

The result from the first part of the study showed that mortality from HIV/AIDS had a decreasing trend after the year 2003. This might not be due to the decreasing number of HIV/AIDS patients but may be because of the improving quality of medicine for alleviating the disease and thus prolonging the patient's life. However, some underestimation of numbers of HIV/AIDS deaths probably cannot be avoided. HIV/AIDS is a sensitive cause of death. Given that there remains a stigma with the disease in Thailand, many deaths from HIV/AIDS have been reported as having a different cause, to protect the reputation of the deceased. However, even allowing for such underestimation, the pattern of HIV/AIDS deaths still shows substantial regional variation. The results from this study confirmed the finding that geographical variations in mortality exist in southern Thailand, which is consistent with earlier research covering the whole of Thailand (Faramnuayphol, 2008).

When considering those who died from unspecified septicemia (ICD-10 = A41.9) in the hospital, a sharply increasing mortality trend appeared clearly after the year 2003. A likely reason for this result is the inaccuracy of identification by physicians of cause of death. Most unclear causes of deaths in hospitals may have an official ‘unspecified septicemia’ cause recorded by a less experienced resident rather than by the specialist physician. This study revealed this failure to record the real cause to be a specific recording problem. The variation of LOS according to demographic and geographic factors among those who died in the hospital is also worthy of study.

Therefore, in the second part of this study the variation of LOS according to demographic and geographic factors for patients dying in hospital was analyzed. LOS stay was found highest among cancer patients. Staying in the hospital is very expensive and drains hospital resources. Setting up palliative care is one of the methods for reducing LOS in the hospital for this group.

When considering the proportion of patients dying in hospital, the highest proportion were those who died from infectious diseases among males aged less than 60. This result reflects the fact that patients who died from these diseases occupied more bed days, and this caused higher health care expenditure than for other disease groups. The result of increasing LOS with increasing age in different diseases revealed that health care systems are needed to manage properly care for the elderly. This study revealed that there were differences in LOS according to hospital size and region. This may be due to differences in the availability of facilities and treatments between small, medium and large hospital. Better treatment may prolong the duration of staying in hospital before dying.

4.1.3 Implication of the study

The pattern of unspecified septicemia mortality was investigated and increasing trends were found. This reflected deficiencies in the recording of cause of death.

Ongoing training in high quality practices in identifying and recording cause of death, consistent across hospitals, is recommended. Most of the national databases in Thailand still face problems with data quality. National databases can provide useful information for setting up policy and planning for the whole country if the data are high quality. Mahapatra et al (2007) mentioned that the usefulness of vital statistics depends on their quality. Most developing countries do not have fully effective civil registration systems to provide important information on population health (Hill et al, 2007). We suggest that there is a need to improve the quality of national databases in Thailand. Not only responsible organization but also relevant resources such as personnel or facilities for storing large databases need to be developed.

4.1.4 Advantages of the study

Enormous volume of information from a country database can provide significant and useful results if appropriate statistical methods are used. This study used different statistical methods to analyse the data. The result provides substantial useful information. This study used the total number of deaths in southern Thailand, and is thus representative of that region's population and the full extent of geographical variation was covered. Population and housing census data were used as denominators to calculate the mortality rate in each region.

The time series counts of death from infectious diseases were revealed. The level of problem can be identified by the statistical method, negative binomial model, used in this study.

This study revealed variation in the length of hospital stay by demographic and geographic factors, which provided an overview of the mortality distribution. This information can assist in deciding and directing policy to reduce unnecessary length of hospital stay.

4.1.5 Limitations of the study

There are several limitations to this thesis. In the first part, information on principal diagnosis before death was not available. Misclassification of causes of death can distort the statistical results. Diseases with small numbers of deaths may provide unreliable results. This limitation may not have much effect on our study because we considered the trend across the region and assumed that each region had the same demographic characteristics of the population. Some information was not available in the vital registration database such as underlying disease, co-morbidity, religion and educational level.

In the second part of the study, information on diseases causing admission to hospital before death was not available. The causes of admitting to hospital and causes of death may not be the same. The analysis of variation of length of hospital stay should consider both cause of admitting to hospital and cause of death in order to acquire a reliable result. Co-morbidity is also another important variable that is an independent and significant predictor of length of hospitalization (Bergeron et al, 2005). If all of these main variables are included in the statistical analysis, the variation of length of hospital stay can be accurately explained and predicted. We did not include co-morbidity in the analysis due to the issue of erroneous data, such as duplicated entry, wrong ICD-10 coding and no ICD-10 code being given.

4.2 Conclusions and Recommendations

This study revealed that mortality from HIV/AIDS among persons aged 15 and over had a peak in 2003 and slightly declined after that. However, mortality inside hospitals from unspecified septicemia dramatically increased from 2003. The incidence of unspecified septicemia mortality had a peak in urban areas especially those having one or more large hospitals. The differences in incidence of mortality from infectious diseases across the region can be related to health resources such as location, hospital size, culture, religion and social context. Evidence for the above is illustrated by the second study of this thesis. Length of hospital stay was highest among those who died from cancer. The distribution of LOS varied across region and hospital sizes. In our study, the normality of residuals for LOS less than zero was not satisfied. This study suggests that LOS recording should include time of admission and time of discharge to avoid the occurring of zero days of LOS. Thus the errors from statistical modeling can be diminished with better fit. Also, in any further studies, autocorrelation should be taken into account before using longitudinal data to investigate the distribution of outcomes.

For national databases with enormous records, data should be aggregated according to each determinant category in order to make it simple and straightforward for data analysis. Normally, country data are longitudinal data, recording the same variables for many years. Analyses of these kinds of data will provide valuable results if appropriate statistical methods are used. Useful information may include effectiveness of health care systems evaluation, burden of each disease, mortality differences across regions and indications of disease epidemic. All such information is useful for setting up health policy and planning for the whole country by health policy makers.

In addition, the accuracy in recording cause of death needs to be improved otherwise the data will remain unreliable and not useful for decision making in health policy and planning. We suggest that all deaths should be medically validated or certificated, especially deaths from outside hospital, before providing a death certificate. The physicians who provide information on causes of hospital deaths for death certificates must be trained to provide the accurate cause. Responsibility in giving accurate cause of death should be implemented among relatives when they contact the local registrar, head of villages, medical doctors and relevant personnel who provide the death certificate or information for it. Good quality of death data provides invaluable information. Such information can be used to set up public health plans for the future.

Data used in this thesis were data from vital registration and hospital in-patient data.