Chapter 2

Methodology

This chapter includes a description of the methodology used in the study as the following components:

1. Study design and data sources
2. Data management
3. Path diagram and variables
4. Statistical model

2.1 Study design and data sources

A retrospective ill-defined mortality data analysis from year 2000 to 2009 was carried out. These data were obtained from the Bureau of Health Policy and Strategy, Ministry of Public Health, Thailand. The International Classification of Diseases tenth revision (ICD-10) of ill-defined causes of death codes are R00-R99 which defined as “symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified” when there is unavailable information on cause of death.

The projected population for Thailand from year 2000 to 2030 was obtained from the Institute of Population and Social Research, Mahidol University.

2.2 Data management

Ill-defined death data from the Bureau of Health Policy and Strategy, Ministry of Public Health were recorded as a text file. Data cleaning was performed to eliminate wrong coding and dealing with missing values. Numbers of ill-defined deaths were aggregated by gender, age group, region, year and place of death. Since age was included as a demographic determinant, it was divided into 7 groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59 and 60 years and over. The region was classified into 5 groups: Bangkok, Central, North, Northeast and South. The place of death was divided into 2 groups: in hospital and outside hospital. The projected population was used as denominator by merging with ill-defined mortality data according to gender-age group, region, year and place of death. Gender-age group were combined in order to explain how ill-defined death rate for each gender varies with age which these two variables had significant interaction term. It was divided into 14 groups: male aged 0-9, male 10-19, male 20-29, male 30-39, male 40-49, male 50-59, male 60+, female aged 0-9, female 10-19, female 20-29, female 30-39, female 40-49, female 50-59, and female 60+. R program was used for graphical display and statistical analysis (R Development Core Team, 2010 version 2.11.1).

2.3 Path diagram and variables

The path diagram of this study is shown in Figure 2.1. In this study, ill-defined mortality rate was determined by gender-age group, region, year and place of death.

Path diagram

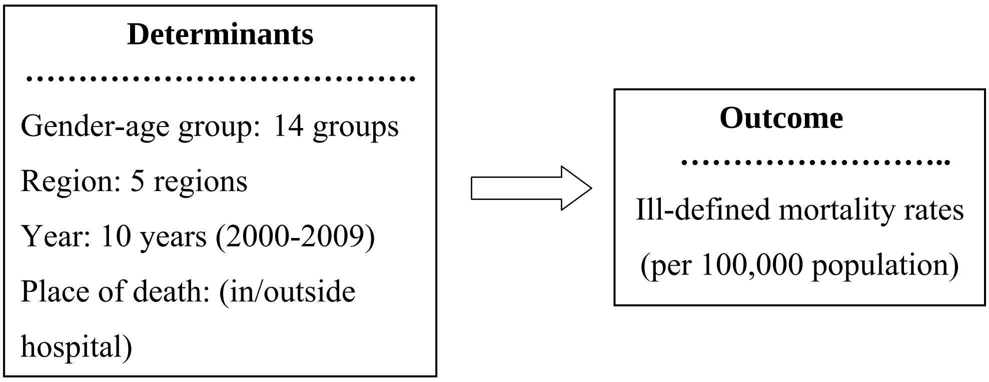
Determinants
Gender-age group: 14 groups
Region: 5 regions
Year: 10 years (2000-2009)
Place of death: (in/outside hospital)
Outcome
Ill-defined mortality rates (per 100,000 population)

Figure 2.1 Path diagram of the study

Variables

Determinants

There are four determinants: gender-age group, region, year and place of death.

The region was classified into 5 groups: Bangkok, Central, North, Northeast and South. Death rates per 100,000 population were computed from the number of ill-defined deaths divided by mid-year population.

Outcome

The outcome is ill-defined death rate from year 2000 to 2009.

2.4 Statistical Methods

2.4.1 Mortality rate

Suppose that D_{ijkm} are a random variable denoted number of ill-defined deaths in gender-age group i (i = male aged 0-9, male aged 10-19, male aged 20-29, male aged 30-39, male aged 40-49, male aged 50-59, male aged 60+, female aged 0-9, female aged 10-19, female aged 20-29, female aged 30-39, female aged 40-49, female aged 50-59, female aged 60+), region *j* (*j* = Bangkok, Central, North, Northeast and South), year *k* (*k* = 2000, 2001, 2002,…, 2009), and place of death *m* (*m* = in hospital and outside hospital) in estimated population P_{ijkm}. Thus the mortality rate can be computed by

$$y_{ijkm} = \frac{KD_{ijkm}}{P_{ijkm}} \quad (1)$$

where y_{ijkm} are ill-defined mortality rate for gender-age group *i*, region *j*, year *k*, and place of death *m*, *K* is a scaling constant such as 1,000, 10,000 or 100,000.

2.4.2 Multiple Linear Regression Analysis

Since death rate for ill-defined was considered as a continuous outcome and the determinants comprise gender-age group, region, year and place of death. Multiple linear regression analysis was fitted. The model takes the form

$$y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_m \quad (2)$$

where y_{ijkm} are the ill-defined death rates, μ is the overall effect, α_{i} are the effects of gender-age group, β_{j} are the effects of region, γ_{k} are the effects of year, and δ_m are the effects of place of death. The model is fitted to the data using least squares, which minimizes the sum of squares of the residuals. Linear regression consists of four assumptions including the association is linear, the variability of the errors (in the outcome variable) is constant and these errors are normal distributed. If these assumptions were not met, the data may need to be transformed. In this study, the death rate was transformed by taking natural logarithms. The estimated additive model for death rates takes the form

$$\ln(\mathrm{y}_{\mathrm{ijkm}}) = \mu + \alpha_i + \beta_j + \gamma_{\mathrm{k}} + \delta_m \qquad (3)$$

The parameter y_{ijkm} are the ill-defined death rates, μ is the overall effect, α_{i} are the effects of gender-age group, β_{j} are the effects of region, γ_{k} are the effects of year and δ_m are the effects of place of death. Poisson model was considered when the linear regression model was not fit to the data.

Poisson Regression

Poisson regression is appropriate for fitting models with count data, which are non-negative and integer values. The probability function for the Poisson distribution with observed counts of y is given by:

$$\text{Prob }(\mathrm{Y} = y)\ \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad (4)$$

where

e is the base of the natural logarithm (e = 2.71828…)

y is the number of occurrences of an event

λ is a positive real number, equal to the expected number of occurrences that occur during the given interval.

Poisson regression model can be fitted by using the generalized linear models (GLMs) equation with the log link function (McCullagh and Nelder, 1989). Suppose that y_{ijkm} is a random variable denoted number of ill-defined deaths in gender-age group i, region *j*, year *k* and place of death *m*. Then the Poisson regression model is takes the form:

$$ln(\lambda_{ijkm}) = ln(P_{ijkm}) + \mu + \alpha_i + \beta_j + \gamma_k + \delta_m . \qquad (5)$$

The parameter λ is the mean of λ_{ijkm}, P_{ijkm} are the population in gender-age group *i*, region *j*, year *k* and place of death *m*, α_i are the effects of gender-age group *i*, β_j are the effects of region *j*, γ_k are the effects of year *k* and δ_m are the effects of place of death *m*. We suppose that the effect of variables α_1, β_1, γ_1 and δ_1 equal zero. A problem with the Poisson regression model occurs when we encounter over-dispersion. This means that the variance is greater than mean. The alternative model which is negative binomial was then considered instead.

Negative binomial regression

The negative binomial is traditional alternative regression model for count data when over-dispersion Poisson occurred. This distribution of observed counts *y* takes the form:

$$\text{Prob (Y}= y) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{k+\lambda}\right)^k \left(\frac{\lambda}{k+\lambda}\right)^y \qquad (6)$$

where Γ is the gamma function and *k* is known as the dispersion parameter, which *k* is greater than 0. Unlike the Poisson distribution that mean must equal with variance, negative binomial can allow variance greater than mean. This can be done because variance of negative binomial is $\lambda + \lambda^2/k$. Note that negative binomial will be equivalence with the Poisson if *k* as dispersion parameter equal to 0. Thus if *k* is equal to 0 Poisson regression model is appropriate, but negative binomial is appropriate if *k* significantly difference from 0.

Goodness of fit

A measure discrepancy between observed and fitted values is the deviance. We show that for Poisson responses the deviance takes the form

$$D = 2\sum \left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i) \right\} \qquad (7)$$

The first term is identical to the binomial deviance, representing "twice a sum of observed times log of observed over fitted". The second term, a sum of differences between observed and fitted values, is usually zero, because Poisson model has the property of reproducing marginal totals, as noted above. For large samples the distribution of the deviance is approximately a chi-squared with *n-p* degrees of freedom, where *n* is the number of observations and *p* the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the model. An alternative measure of goodness of fit is Pearson's chi-squared statistic, which is defined as

$$. \qquad \chi_p^2 = \sum \left(\frac{y_i - \hat{y}_i}{\hat{y}_i} \right)^2 \qquad (8)$$

The numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. The Pearson's statistics has the same from for Poisson and binomial data, namely a sum of squared observed minus expected over expected.

In large samples the distribution of Pearson’s statistics is also approximately chi-squared with *n-p* degree of freedom. One advantage of the deviance over Pearson’s chi-squared is that it can be used to compare nested models.

2.5.5 Sum Contrasts

Sum contrast (Venables and Ripley, 2002; Tongkumchum and McNeil, 2009) was used to obtain confidence intervals for comparing means within each factor with the overall mean. An advantage of these confidence intervals is that they provide a simple criterion for classifying level of the factor into three groups according to whether each corresponding confidence intervals exceeds, crosses, or is below the overall mean.