Chapter 2

Methodology

This chapter describes the methods used in the study including study design, data collection and management, path diagram, variables and statistical methods. Graphical and statistical analyses were performed by using R program.

2.1 Study design

We performed a retrospective 10-years analysis of drowning mortality trend between 2000 and 2009 according to gender, Public Health Area (PHA), and age group.

2.2 Study groups

The study groups comprised 40,604 persons who died from drowning in Thailand during the period 2000 to 2009.

2.3 Data collection and management

Data collection

Drowning death data from death certificate were obtained from Bureau of Health Policy and Strategy, Ministry of public Health in the period from January 2000 through December 2009. These data provide information on age, gender, place of death, year and cause of death which based on the International Statistical Classification of Disease and Related Health Problem (ICD-10). The ICD-10 codes of drowning mortality are W65-W74.

Mid- year forecasted population separated by gender, age group, and province from 2000 to 2009 were obtained from the Population Projections for Thailand 2000-2030, the Institute of Population Studies at Mahidol University.

Data management

Drowning death data from the Bureau of Health Policy and Strategy, Ministry of public Health were recorded as a text file. Error checking was performed in order to find wrong codes, missing values and extreme values. All of the errors were cleaned before analyzing the data. Missing age was found for 0.01%. All of these missing values were group into aged 80 years and over according to Health Systems Research Institute and Institute for Population and Social Research, Mahidol University (2546)'s suggestions as most of missing ages are elderly people. Unknown province of death was found for 1,016 deaths or 2.44%. Thus all of these values were omitted and 40,604 deaths were used for further analysis. R program was used for graphical display and statistical analysis (R Development Core Team, 2010).

2.4 Path diagram and variables

Path diagram

Path diagram comprises the determinants and outcome. There are three determinants: gender-age group, PHA and year. The outcome is drowning death rate (per 100,000 population). The schematic diagram for this study is shown in Figure 2.1.

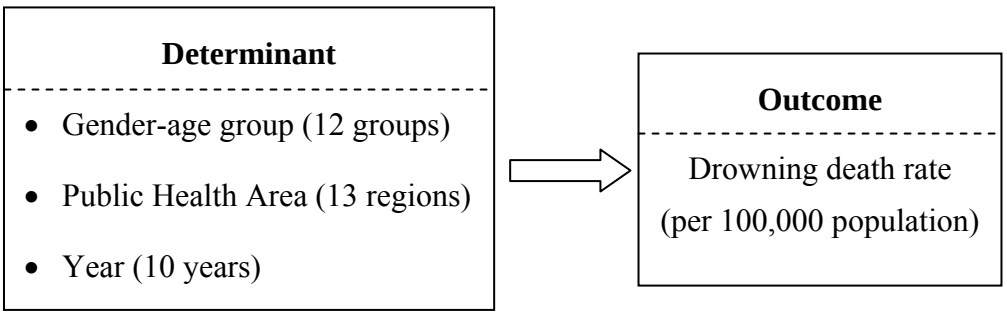

Figure 2.1: Path diagram

Variables

Death rates per 100,000 population were computed from the number of drowning deaths divided by mid-year population.

Region was classified according to Public Health Area (PHA), Ministry of public Health into 13 regions.

PHA 1 comprises 5 provinces: Samut Prakan, Nonthaburi, Pathum Thani, Phra Nakhon Si Ayutthaya and AngThong.

PHA 2 comprises 6 provinces: Lop Buri, Sing Buri, ChaiNat, Saraburi, Nakhon Nayok and Suphan Buri.

PHA 3 comprises 7 provinces: Chon Buri, Rayong, Chanthaburi, Trat, Chachoengsao, Prachin Buri and SaKaeo.

PHA 4 comprises 7 provinces: Ratchaburi, Kanchanaburi, Nakhon Pathom, Samut Sakhon, Samut Songkhram, Phetchaburi and Prachuap Khiri Khan.

PHA 5 comprises 5 provinces: Nakhon Ratchasima, Buri Ram, Surin, Chaiyaphum and Maha Sarakham.

PHA 6 comprises 7 provinces: Nong Bua Lam Phu, Khon Kaen, Udon Thani, Loei, Nong Khai, Kalasin and Sakon Nakhon.

PHA 7 comprises 7 provinces: Si Sa Ket, Ubon Ratchathani, Yasothon, Amnat Charoen, Roi Et, Nakhon Phanom and Mukdahan

PHA 8 comprises 5 provinces: Nakhon Sawan, Uthai Thani, Kamphaeng Phet, Tak and Sukhothai.

PHA 9 comprises 6 provinces: Uttaradit, Phrae, Nan, Phitsanulok, Phichit and Phetchabun.

PHA 10 comprises 6 provinces: Chiang Mai, Lamphun, Lampang, Phayao, Chiang Rai and Mae Hong Son.

PHA 11 comprises 7 provinces: Nakhon Si Thammarat, Krabi, Phangnga, Phuket, Surat Thani, Ranong and Chumphon.

PHA 12 comprises 7 provinces: Songkhla, Satun, Trang, Phatthalung, Pattani, Yala and Narathiwat.

PHA 13 is Bangkok.

Gender consists of male and female.

Age was divided into 6 groups: 0-4, 5-14, 15-29, 30-44, 45-59, and 60 years and over.

Gender and age group were combined in order to explain how drowning death rate for each gender varies with age which these two variables had significant interaction term. It was divided into 12 groups: male aged 0-4, male aged 5-14, male aged 15-29, male aged 30-44, male aged 45-59, male aged 60+, female aged 0-4, female aged 5-14, female aged15-29, female aged 30-44, female aged 45-59, and female aged 60+ years.

Variable Types

The data used for this study consist of gender and age group (6 groups), PHA (13 regions), year (2000-2009) and drowning death rate. Gender and age group were combined and formed a new variable named gender-age group (12 groups). The variable roles and data types are shown in Table 2.1.

Table 2.1: Variable types and roles

Variables	**Types (No. of categories)**	**Roles**
Drowning death rate	Continuous	Outcome
Year	Ordinal (10)	Determinant
PHA	Ordinal (13)	Determinant
Gender	Binary (2)	Determinant
Age group	Ordinal (6)	Determinant
Gender-age group	Ordinal (12)	Determinant

2.5 Statistical methods

2.5.1 Drowning death rate

Drowning death rates (y_{ijt}) was computed as the number of drowning deaths divided by the number of mid-year projected population and multiply by 100,000 population, given by

$$y_{ijt} = \frac{D_{ijt}}{P_{ijt}} \times K \qquad (1)$$

where D_{ijt} is the number of deaths in PHA (i) (i = 1, 2, 3…, 13), gender-age group (j) (j = 11,…, 16, 21,…, 26) and year (t) (t = 2000, 2001, 2002,…, 2009). P_{ijt} is the population at middle year and K is a specified constant, here equal to 100,000.

2.5.2 Multiple linear regression

Since drowning death rate was considered as a continuous outcome and the determinants comprise year, PHA, and gender-age group, multiple linear regression analysis was the appropriate method for statistical modeling. The estimated multiple linear regression model takes the form

$$y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_t . \qquad (2)$$

In this formula y_{ijt} is the drowning death rate, μ is the overall effect, α is the effect of Public Health Area (PHA), β is the effect of gender-age group, and γ is the effect of year. The model is fitted to the data using least squares, which minimizes the sum of squares of the residuals. Linear regression analysis resets on three assumptions including the association is linear, the variability of the error (in the outcome variable) is uniform and these errors are normal distributed. If these assumptions were not met,

the data may need to be transformed. Linear regression analysis can be performed for both continuous and categorical determinants. In the model, the categorical determinant is broken down into *c*-1 parameters for each determinant, where *c* is the number of categories. The omitted category is taken as the baseline or reference category.

This study, the continuous outcome variable was defined as the death rate. The rates generally have positively skewed distributions so it is conventional to transform them by taking logarithms. Drowning counts are often zero for small regions, thus to avoid the problem from taking logarithms of zero, we replaced them by a specified fixed constants by 0.5 before log-transforming. The estimated additive model for drowning death rates is taken the form

$$\ln(y_{ijt}) = \mu + \alpha_i + \beta_j + \gamma_t . \qquad (3)$$

The parameter y_{ijt} is the drowning death rate, μ is the overall effect, α is the effect of Public Health Area (PHA), β is the effect of gender-age group, and γ is the effect of years. Poisson model was also considered when the linear regression model was not fit to the data.

2.5.3 Poisson regression

Drowning death is the count data being the number of persons who died from drowning. 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,

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

where λ is the Poisson parameter, which equals both the mean and the variance of the Poisson distribution, e is the base of the natural logarithm (e = 2.71828…), and y is the number of occurrences of the event, which the probability giving by the Poisson function.

Since λ is the adjustable parameter in the Poisson model for the variation in observed count, it is natural to link λ to the values of explanatory variable of interest. The mean *(*λ) of a Poisson distribution must be greater than zero. It would be unsuitable simply to assumption that

$$\lambda = a + b_i x_i + ... + b_t x_t \ . \qquad (5)$$

The restriction of Poisson distribution is the mean must equal to variance. Poisson regression model can be fitted by using the generalized liner models (GLMs) equation with the log link function. Suppose that λ_{ijt} is a number of drowning deaths in PHA *i*, gender-age group *j*, and year *t*. Then the Poisson regression model is given by

$$\ln(\lambda_{ijt}) = \ln(P_{ijt}) + \mu + \alpha_i + \beta_j + \gamma_t \, . \qquad (6)$$

The parameter λ is the mean of λ_{ijt}, P_{ijt} is the population in PHA *i*, gender-age group *j*, and year *t*, α is the effect of PHA, β is the effect of gender-age group, and γ is the effect of year. We suppose that the effect of variables α_1, β_1, and γ_1 equal zero. The assumption of Poisson model is often violated due to the problem of over dispersion. This means that the variance is greater than mean. The alternative model which is negative binomial was then considered instead.

2.5.4 Negative binomial regression

The negative binomial is the alternative regression model for count data when Poisson regression model does not fit to the data. The distribution of observed counts *y* takes the form:

$$\Pr ob(Y = y) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{k+\lambda}\right)^{k}\left(\frac{\lambda}{k+\lambda}\right)^{y}. \qquad (7)$$

In this formula Γ is the gamma function and *k* is known as the dispersion parameter, (*k* is greater than 0). Unlike the Poisson distribution when the mean must equal the variance, the negative binomial is $\lambda+\lambda^2/k$. The negative binomial is equivalence to the Poisson distribution if *k* (the dispersion parameter) is equal to 0. Thus if *k* is equal to 0, Poisson regression model is appropriate whereas the negative binomial is appropriate if k is significantly difference from 0.

2.5.5 Goodness of Fit

A measure of discrepancy between observed and fitted values is the deviance. Negative binomial response the deviance given by,

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

The first term is identical to the binomial deviance, representing “twice a sum of observed times log of observe over fitted”. The second term, a sum of differences between observed and fitted value, is usually zero, because Negative binomial models has the property of reproducing marginal total, as noted above. For large samples, the distribution of the deviance is approximately a chi-squared with *n-p* degree of

freedom, where n is the number of observations and p is 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 denoted as

$$\chi_p^2 = \sum \left(\frac{y_i - \hat{y}_i}{\hat{y}_i} \right)^2 . \qquad (9)$$

The Pearson statistics has the same form of the Poisson and binomial data, namely a sum of squared observed counts minus expected counts over expected counts.

2.5.5 Sum Contrasts

Sum contrast (Venables and Ripley, 2002; Tongkumchum and NcNeil, 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.