Chapter 2

Methodology

This chapter describes the methodology of all three parts. The first section outlines the process of analysis in three parts. Data sources and management, statistical analysis including graphical methods in each part were separately described in the second, the third and the fourth section, respectively.

2.1 Process of analysis

This section described the process of analysis including data source, statistical methods and model application of three parts as shown in Table 2.1.

Table 2.1 Process of analysis

Topic	Part I	Part II	Part III
- Data used	- DR data from 1996-2009	- The 2005 VA data	- DR data from 1996-2009 after imputing unknown province and age from Part I - HIV estimated deaths from 1996-2009 after correcting misclassification from Part II

Table 2.1 Process of analysis (cont.)

Topic	Part I	Part II	Part III
- What to predict?	- Unknown province or age from DR data for each year from 1996-2009	- HIV deaths from the 2005 VA- assessed	-
- Model's prediction	- Logistic regression	- Logistic regression	-
- Model's assessment	- Normal quantile plots	- ROC curve	-
- Interpolation methods	-	- Triangulation with linear algebra	- Natural cubic spline functions for all-cause deaths (Part III A)
- What statistics generated from model	- Probability of unknown province in each sex and age groups from 1996-2009 - Probability of unknown age in each sex and province from 1996-2009	- Probability of HIV VA-assessed deaths in each DR cause-location, sex-age groups and provinces in the 2005 VA data	-

Table 2.1 Process of analysis (cont.)

Topic	Part I	Part II	Part III
- Data that model was generated to	- DR data from 1996-2009	- DR data from 1996-2009	-
- What statistics generated after model application	- Number of all-cause deaths in each sex and age groups from 1996-2009 after imputing unknown province and unknown age	- Proportion of HIV-estimated deaths from 1996-2009 after correcting misclassification	-
- Subsequent application	- Estimation of all-cause mortality from 1996-2009 by sex, age groups and provinces in each year	- Estimation of HIV mortality from 1996-2009 by sex, age groups and provinces in each year	- Patterns of all-cause and HIV mortality from 1996-2009 by sex, age groups and provinces

2.2 Methodology of Part I: Estimation of mortality with missing data using logistic regression (Chutinantakul *et al*., 2014a)

2.2.1 Data sources and management of Part I

DR data from death certificates were provided by the Bureau of Registration Administration, Ministry of Interior and coded as cause-of death by the Bureau of Policy and Strategy (2010), Ministry of Public Health. The data were recorded using electronic DR database since 1996. This thesis focuses on DR data from 1996-2009 that comprised of variables of interest as sex, age, region, location of death and ICD-10 codes of reported cause groups. Deaths with unknown province and age were omitted in this part because unknown province we used "age" as the determinant, conversely unknown age we used "province" as the determinant. Thus, both missing needed to be omitted.

The DR data from 1996-2009 was classified into 18 categories of five-year age groups and 90 or more (0-4, 5-9, 10-14, … , 85-89, 90+) by sex in 76 provinces of Thailand. The classification of five years interval age groups are followed the National Statistical Office, Ministry of Information and Communication Technology. The numbers of deaths reported were tabulated into cells (2x77x20 = 3,080 records) based on all combinations of the two sexes, the 77 groups comprising the provinces and an additional "unknown province" group, and 20 groups comprising the age groups and a further "unknown age" group. These two unknown values are to be imputed in Part I to allow better analysis the pattern of all-cause mortality in Part III.

2.2.2 Statistical analysis of Part I

Statistical modeling

Logistic regression models were used to predict unknown province or age in Thai DR data from 1996-2009. The goodness of fit of the models were assessed using normal quantile plots. The adjusted proportions from the models were presented as 95% confidence intervals using the standard errors based on the weighted sum contrasts.

Logistic regression

The outcome of epidemiological data is usually dichotomous such as disease or non-disease, ill or not-ill, exposed or non-exposed. The logistic regression model is well suited to analyze the binary outcome (McNeil, 1996; Woodward, 1999; Chongsuvivatwong, 2007). The method involves fitting the generalized linear model from the binomial family with logit link by maximum likelihood as described in Venables and Ripley (2002). Logistic regression formulates the logit of the probability of occurrence of the outcome as an additive linear function of the determinant factors.

When cross-tabulating numbers of deaths by sex, age-group and province, logistic regression provides an appropriate method for allocating deaths with unknown province or age (but not both) to cells in the table. The methods involved two steps. First, an additive model with factors sex and age-group were fitted to the proportion of deaths with unknown province, and these fitted proportions were used to inflate numbers in age groups by the same factor for each province. Second, this process was repeated on the inflated death counts with age group and province interchanged.

The binary outcome was defined as either (a) unknown or known province, or (b) unknown or known age. In each case, the model fitted was an additive combination of two factors that were fitted to DR data in each year.

In case of unknown province, the models are formulated as

$$logit(P_{ij}) = ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \mu + \alpha_i + \beta_j, \qquad \ldots\ldots\ldots (a)$$

where P_{ij} is the probability of death with unknown province in sex *i* and age group *j*, μ is a constant, α_i and β_j are individual parameters specifying sex *i* (*i* =1 for males and *i*=2 for females) and age group *j* (j=1,2,3,…,19), respectively.

In case of unknown age, the models are formulated as

$$logit(P_{ik}) = ln\left(\frac{P_{ik}}{1-P_{ik}}\right) = \mu + \alpha_i + \gamma_k, \qquad \ldots\ldots\ldots (b)$$

where P_{ik} is the probability of death from unknown age group in sex *i* and province *k*, α_i and γ_k are individual parameters specifying sex *i* and province *k* (k=1,2,3,…,76), respectively. The data for each model comprised all cases with known age or province, and thus only those cases with both unknown province and age were omitted. The models were fitted sequentially with the total numbers of known deaths in provinces for model (b) inflated and rounded to integers using results from model (a). The adequacy of fit of each model was assessed by using a plot of deviance residuals versus theoretical quantiles.

Sum contrasts

Venables and Ripley (2002) and Tongkumchum and McNeil (2009) described a method for computing confidence intervals using weight sum contrasts to compare population means in unbalanced designs. Kongchouy and Sampantarak (2010) presented confidence intervals for adjusted proportions using logistic regression with weighted sum contrasts. The advantage of this method is that by using appropriately weighted sum contrasts each proportion can be compared with the overall mean rather than with a specified reference group. It provides a simple criterion for classifying levels of a factor. This method was used by Odton *et al*. (2010a; 2010b) in mortality study.

The adjusted proportions of deaths with unknown province or age were presented using graphs of confidence intervals. Since it is more appropriate to compare province or age effects with their overall mean, rather than with an arbitrary province, the standard errors for the estimated parameters in the model are based not on the conventional treatment contrasts matrix where the first level is left out from the model to be the reference. We used "weighted sum contrasts" (Tongkumchum and McNeil, 2009; Kongchouy and Sampantarak, 2010). For each level of each covariate factor, 95% confidence intervals were computed using this method. The confidence intervals for the proportion from this method were classified into three levels as totally above the mean, crossing the mean and totally below the mean (Odton *et al*., 2010a; 2010b).

Estimation of mortality with missing data

The adjusted proportions from model (a) were used to correct deaths in each age group with unknown province by multiplying the reported numbers (with no missing) in sex *i* and age group *j* by $1/(1-P_{ij})$, where P_{ij} is the estimated proportion of deaths with unknown province in sex *i* and age group *j*. Thus, the estimated numbers of deaths were obtained.

In the same way, we corrected deaths in each province with unknown age group by multiplying the estimated numbers of deaths from model (b) in sex *i* and province *k* by $1/(1-P_{ik})$, where P_{ik} is the estimated proportion of deaths with unknown age group in sex *i* and province *k*. These procedures thus redistribute reported deaths with unknown province for each sex and age group to known province and redistribute reported deaths with unknown age group for each sex and province to known age groups.

2.2.3 Graphical methods for data analysis of Part I

Quantile-Quantile (Q-Q) plots

A normal quantile plots are used to check the normality assumption of the residuals from regression models.

Thematic map

A thematic map is a type of map specially designed to show a particular theme connected with a specific geographic area. Thematic map displays information about map's underlying data. It emphasizes spatial variation of one or a small number of

geographic distributions. We used it to display adjusted proportions of unknown age for comparison between geographical areas. These values were classified into three levels of their confidence intervals as totally above the mean, crossing the mean and totally below the mean.

2.3 Methodology of Part II: Correcting and estimating HIV mortality in Thailand based on 2005 verbal autopsy data focusing on demographic factors, 1996-2009 (Chutinantakul *et al*., 2014b)

2.3.1 Data sources and management of Part II

The 2005 verbal autopsy (VA) data were used as a tool to correct misclassification cause of HIV deaths. This data comprised a sample of 9,644 deaths (3,316 in-hospital and 6,328 outside-hospital) from 28 selected districts in nine provinces of four regions. These were Bangkok, Nakhon Nayok, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Suphan Buri, Chumphon and Songkhla. The VA data comprised variables of interest. These are sex, age, region, location of death, ICD-10 reported cause groups and ICD-10 VA-assessed cause groups.

Part II was confined to deaths of persons aged 5 years and older, for which HIV death is common and often misclassified (age under 5 years separate for special attention in other study and few cases of HIV deaths). Thus, this part comprised the sample of 9,495 deaths (3,212 in-hospital and 6,283 outside-hospital) from nine provinces as shown in Figure 2.1. The sample sizes were shown in bubble. Ubon Ratchathani, Suphan Buri and Chiang Rai had the largest numbers of total deaths, while Chumphon had the lowest.

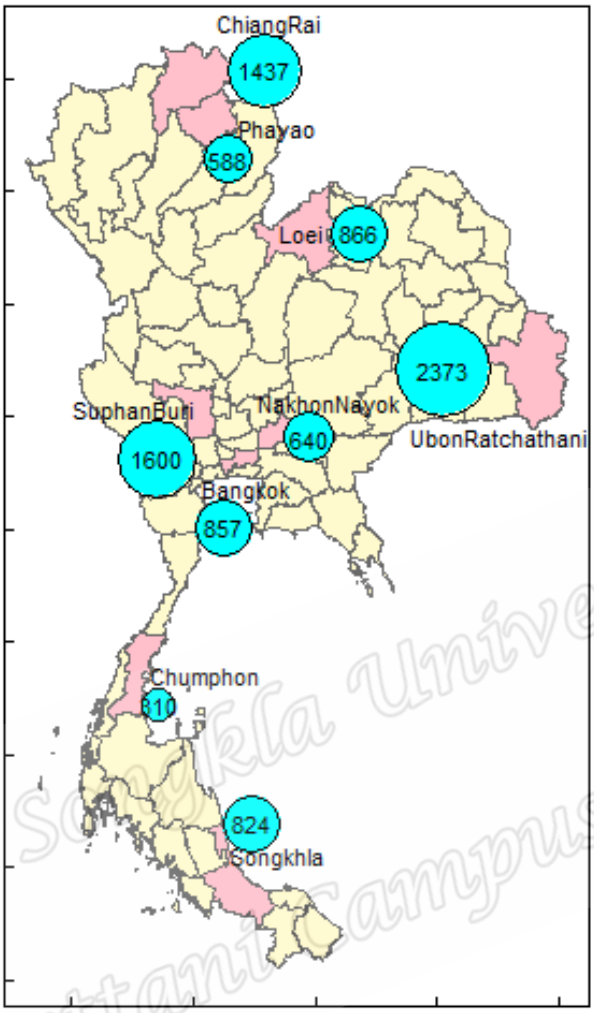

Figure 2.1 The map of nine provinces and sample sizes of death ages 5 years and older in the 2005 VA study

Accordingly, the chapter-block classifications of ICD-10 codes, consisting of blocks categorized mainly by human organs, were used to create 21 major cause groups for deaths at ages 5 years and older based on the distribution of VA-assessed deaths that obtain sufficient sample size. For statistical accuracy, groups with small counts (mainly less than 200) were combined into larger groups using medical considerations (apart from septicemia, which received special attention due to over-reporting). For HIV cause group, it comprised of 512 cases at age 5 years and older.

Table 2.2 summarizes the cause groups based on VA counts with ICD-10 codes. The proportion of all deaths represented by these categories varied from 0.8% for septicemia to 11.3% for stroke and 5.4% for HIV deaths (ICD 10 code B20-24)

Table 2.2 Cause groups based on VA counts

Cause groups	VA count	Percentages
1: TB (A15-19)	195	2.1
2: Septicemia (A40-41)	77	0.8
3: HIV (B20-24)	512	5.4
4: Other Infectious (A, B)$^{-}$	219	2.3
5: Liver Cancer (C22)	500	5.3
6: Lung Cancer^{+}(C30-39)	320	3.4
7: Other Digestive Cancer (C15-26^{-})	290	3.1
8: Other Cancer (C^{-}, D0-48)	697	7.3
9: Endocrine (E)	647	6.8
10: Mental, Nervous (F, G)	223	2.3
11: Ischemic (I20-25)	617	6.5
12: Stroke (I60-69)	1,076	11.3
13: Other CVD (I^{-})	540	5.7
14: Respiratory (J)	801	8.4
15: Digestive (K)	489	5.2
16: Genitourinary (N)	412	4.3
17: Ill-defined (R)	501	5.3
18: Transport Accident (V)	536	5.6
19: Other Injury (W, X0-59)	327	3.4
20: Suicide (X60-84)	158	1.7
21: All other	358	3.8
total	9,495	100.0

$^{+}$*Respiratory/thoracic,* $^{-}$*exclude above*

For logistic regression efficiency, the predictors were optimally grouped to obtain sufficient sample size for relatively homogeneous risk groups. Nine provinces were included in the VA study (Bangkok, Nakhon Nayok, Suphan Buri, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Chumphon and Songkhla). The effects of age in the two sexes were considered different. Sex and age were grouped together into 14 levels (with seven levels of age in years: 5-19, 20-29, 30-39, 40-49, 50-59, 60-69 and 70+). This combining variables resulted in more degrees of freedom (14-1=13) compared to leave them separated ((2-1) + (7-1) = 7). However, our data set is large, additional degrees of freedom do not cause imprecision but provided better estimate of each sex-age groups.

Similarly, misclassification of causes of death was considered differently for deaths in and outside hospitals. The number of levels in the DR cause groups and location of death factor will vary according to the number of such reported cause groups that affect the outcome cause groups. There are eight leading causes groups that affected the outcome including other group (nine major causes of death that mimic a related HIV: HIV, respiratory, septicemia, TB, other infectious, mental and nervous system, digestive, ill-defined, and the remainder, which were aggregated into a single group, and two levels of location (in and outside hospital). Thus, nine DR reported causes of death and two levels of location of death were grouped together into 18 levels.

The DR data from 1996-2009 were classified to follow the same age groups in the 2005 VA data at aged 5 years and older. Deaths with unknown age were classified to old age group because incomplete age group mostly occurred in old age (Prasartkul *et al*., 2000). In addition, unknown age is negligible (4.78% in 1996 and 2.72% in

1997) and less than 1% after 1997. We omitted unknown provinces because unknown province mostly occurred in age group 0-4 years but we focus on aged 5 years and older.

2.3.2 Statistical analysis of Part II

Statistical modeling

Logistic regression models were used to correct under-reporting/misclassification of HIV deaths based on the 2005 VA data. Receiver Operating Characteristic (ROC) curve was used to assess model's ability instead of normal quantile plots because the models were forced to give predicted number of HIV deaths in agreement with the observed number based on the cut-off point. Triangulation method with linear algebra was used to interpolate province coefficients outside the VA study provinces. These methods provided an effective way of minimizing the effects of DR data quality.

Through logistic regression, we estimated the logit of the probability P_i that a person died from HIV as a linear function of the determinant factors. The simple logistic regression model with simple cross-referencing is formulated as

$$logit(P_i) = ln\left(\frac{P_i}{1-P_i}\right) = \mu + \alpha_i, \qquad \ldots\ldots\ldots (A)$$

where P_i is the probability of death due to HIV, μ is a constant and α_i is the only parameter of DR cause-location *i*. The simple model (A) was compared with the full model (B) which includes an additive linear function of the determinant factors, which could be expressed as

$$logit(P_{ijk}) = ln\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \mu + \alpha_i + \beta_j + \gamma_k, \qquad \ldots\ldots\ldots (B)$$

where P_{ijk} is the probability of death due to HIV and α_i, β_j and γ_k are individual parameters specifying DR cause-location group *i*, sex-age group *j* and province *k*, respectively.

This method also gives confidence intervals for percentages of HIV deaths from VA-assessed in cause groups for levels of each risk factor adjusted for other risk factors, using "sum contrasts" rather than the conventional "treatment contrasts" (Tongkumchum and McNeil, 2009) and Kongchouy and Sampantarak (2010) as described in Part I. These confidence intervals are compared with bar charts of sample percentages to assess evidence of confounding.

Comparing between the simple cross-referencing and the full model suggested that the full model is appropriated and powerful. Thus, the full model was chosen to estimate HIV deaths in DR data. To do this, the nine provinces coefficients from the full model were used to interpolate the province coefficients outside the 2005 VA study using spatial triangulation method.

Model's assessment

Receiver Operating Characteristic (ROC) curve

Receiver Operating Characteristic (ROC) curve displays the discriminatory capacity of a diagnostic test or marker to discriminate between two groups of subjects. ROC curve was used to assess the logistic regression model's ability to discriminate the outcome (Chongsuvivatwong, 2007; Fan *et al*., 2006). Area under the ROC curve

(AUC) is widely recognized to measure or compare two tests or model's abilities (Grzybowski and Younger, 1997; Park *et al*., 2004 and Sakar and Midi, 2010). Williams *et al*. (2005) used area under the ROC curve to observe the performance of logistic regression method.

The ROC curve is an alternative method to assess logistic regression model's ability. The ROC analysis comes from statistical decision theory. A graphic displays the predictive accuracy of a logistic model by area under the ROC curve (AUC). AUC measures the performance of a model and represents model accuracy (Sakar and Midi, 2010; Takahashi *et al*., 2006). It shows how well a model predicts a binary outcome (Fan *et al*., 2006). The maximum value for the AUC is 1, indicates a (theoretically) perfect test (100% sensitive and 100% specific). Sensitivity is the proportion of patients *with* disease who test positive. Specificity is the proportion of patients *without* disease who test negative. An AUC value of 0.5 indicates no discriminative value (50% sensitive and 50% specific). There are several scales for AUC value interpretation but, in general ROC curves with an AUC ≤ 0.75 are not clinically useful. The cut-off points in the ROC curves present sensitivity and specificity of its model. When utility of true positive and true negative are equal and disutility false positive and false negative are also equal, the best cut-off point value provides both the highest sensitivity and the highest specificity (Grzybowski and Younger, 1997). However, some studies critic the AUC (Lobo *et al*., 2008)

The ROC curve was used to assess logistic regression model instead of using a plot of deviance residuals versus theoretical quantiles because we restricted the predicted number from our models to match the observed number from the VA data that was

fixed by the cut-off point of the ROC curve. In this case, we used VA-assessed deaths as gold standard. Denoting the predicted outcome as 1 (HIV death) if $P \geq c$ (cut off point) or 0 (other death) if $P <$ c, it plots sensitivity (proportion of positive outcomes correctly predicted by the model) against the false positive rate (proportion of all outcomes incorrectly predicted), as c varies. We choose $c = 0.3365$ that must gives 512 predicted HIV deaths, in agreement with the observed number of HIV deaths in the VA study. A cut-off point in the curve, where the predicted number of HIV deaths equals the observed value in the VA data (512 cases), was used to report sensitivity and specificity of the model. These were to be compared with results from simple cross-referencing model in equation (A). Fitting the complete logistic regression model to the 2005 VA data resulted in nine province coefficients, 14 sex-age group coefficients, and 18 DR cause-location coefficients and the estimate of HIV deaths and 95% confidence intervals.

Estimation of HIV mortality

For the remaining 67 provinces, we used a simple and easily implemented spatial "triangulation method" (Li and Heap, 2008; Yang, *et al*., 2013) to interpolate province coefficients. Estimating provinces coefficients outside the VA study are based on the latitude and longitude of their central points (Bhargava, *et al*., 2013). This was preferred to the "kriging" method because it uses fewer points than kriging, and there were insufficient sample provinces (only nine) to provide the basis for kriging (Murphy, 2013).

Triangulation interpolating method

The triangular irregular network (TIN) is a deterministic method (Li and Heap, 2008). Deterministic methods have no assessment of errors with the predicted values while stochastic methods provide an assessment of the errors associated with the predicted values. In TIN, all sampled points are joined into a series of triangles based on a Delauney's triangulation. Each triangle does not contain any of the sample points. The value of a point in each triangle is estimated by linear or cubic polynomial interpolation. TIN model is suitable for regular and irregular distributed data point. It provides better precision and efficiency of calculating the elevation than contours model (Hanjiang, *et al*., 2008).

To do this, triangles were drawn linking the nine provinces. These triangles were set as planes, like roofs on poles with heights corresponding to their model coefficients values at the vertices of the triangles. The values of province coefficients in each triangle were assigned as an average of coefficients from nearby provinces in the model. Then, we turn geometry into linear algebra. For each triangle, values *a*, *b* and *c* were obtained by solving three equations (1, 2, 3) using linear algebra based on latitude and longitude as follows.

$$a+longP_1\times b+latP_1\times c = \beta_{P1} \quad \ldots\ \ldots\ \ldots\ (1)$$

$$a+longP_2\times b+latP_2\times c = \beta_{P2} \quad \ldots\ \ldots\ \ldots\ (2)$$

$$a+longP_3\times b+latP_3\times c = \beta_{P3} \quad \ldots\ \ldots\ \ldots\ (3)$$

(Note: P = Province, β = coefficient)

The coefficient for any province *j within* a triangle could then be given by

$$\beta_{Pj} = a + longP_j \times b + latP_j \times c \quad \ldots \ldots \ldots \quad (4)$$

Coefficients for provinces outside triangles were obtained similarly by extrapolation from nearby provinces (such as three provinces in the far South). All province coefficients were thus obtained. These interpolated and extrapolated coefficients were used to estimate proportions of HIV deaths in the 2005 DR data. We used the interpolated values for the province effects and assumed that the model is valid for years before and after 2005, so that the patterns of misreporting of deaths in these years were the same as in 2005. Then, we applied the model to the DR data from 1996-2009. Finally, HIV-estimated deaths were obtained.

2.3.3 Graphical methods for data analysis of Part II

Bubble chart

A bubble chart is used to visualize a data set with two to four dimensions. The first two dimensions are visualized as coordinates, the third as color and the fourth as size. Bubble charts communicate the raw count, frequency, or proportion of some variables where the size of the bubble reflects the quantity. It is used to illustrate deaths in the 2005 VA data between DR cause groups and VA cause groups.

Bar plot with 95% confidence interval

A confidence interval is an interval from a population that contains the true value of a population parameter with specified probability (usually 95%). The confidence intervals with the line segments centered at the ends of the bars are used to compare

between crude and adjusted proportions from the logistic regression models. If the line segment covered the end of the bar, it indicated that no confounding. The bar charts presented crude proportions whereas the dots presented adjusted proportions with the line segments of 95% confidence intervals. The confidence intervals overlap the ends of the bars indicated no confounding and conversely confounding presented.

Thematic map

The thematic maps in Part II display proportions based on logistic regression model results including province coefficients. These maps were classified into three levels of province coefficients based on percentile ranks. The low-level of province coefficients are entirely below 33.33 percentiles, between 33.33 - 66.66 percentiles for medium-level, and entirely above 66.66 percentiles for high-level.

2.4 Methodology of Part III: Exploring mortality patterns of all-cause and HIV from 1996-2009

2.4.1 Data sources and management of Part III

This part used DR data from 1996-2009 after imputing unknown demographic factors from Part I and reclassification of cause of HIV deaths from Part II.

2.4.2 Statistical analysis of Part III

Spline functions

Spline functions widely use in multidisciplinary such as mathematics, engineering, demographic and public health. Cubic spline interpolation is a useful technique to

interpolate between known data points due to its stable and smooth characteristics as described by Kruger (2003). Also, McNeil *et al*. (2011) described the natural cubic spline interpolation in age-specific demographic data. The function based on this method is smooth and always non-negative. Wand (2000) compared regression spline smoothing procedures and suggested that the univariate smoothing setting with Gaussian noise and the truncated polynomial regression spline are simple and concise.

Natural cubic spline function was used to smooth the mortality patterns after manipulated missing data. Cubic spline interpolation is a powerful data analysis tool. A general approach is that the user enters a sequence of points, and a curve is constructed whose shape closely follows this sequence. The points are called *control point*. A curve passes through each control point is called *interpolating curve*; a curve that passes near the control points but not necessarily through them is called an *approximating curve*. Their goal is to get an interpolation formula that is continuous in both the first and second derivatives, both within the intervals and at the interpolate knots. This will give a smoother interpolating function. For simplicity and conciseness in smoothed age distributions of deaths, cubic spline interpolation is an efficient strategy for smoothing cumulative death counts (Wand, 2000).

After allocation of deaths with unknown province or age, we can create such data by fitting *natural cubic spline functions* to province mortality of all-cause grouped by sex and age groups, using a method developed by McNeil *et al*. (2011). Natural cubic splines were used to interpolate age-specific demographic data to ensure that relevant boundary conditions on second derivatives are satisfied. This method can be used to

smoothly interpolate non-negative mortality data because age-specific demographic data functions are necessarily non-negative.

A cubic spline with *n* knots $x_1 < x_2 < \ldots < x_n$ is any function *s(x)* with continuous second derivatives comprising piecewise cubic polynomials between and beyond the knots. Denoting by x_+ the function taking the value *x* for *x* > 0 and 0 elsewhere, s*(x)* may be written as

$$s(x) = d_0 + d_1x + d_2x^2 + d_3x^3 + \sum_{i=0}^{n} c_i\,(x - x_i)^3{}_+ \ldots\ldots\ldots (5)$$

Linear function for values of *x* outside the knots was satisfied for the additional requirement of natural cubic spline. All functions with specified values at the knots of the natural cubic spline minimizes the integral of its squared second derivative over the interval (x_1, x_n).

Since *s*(*x*) is linear for *x* <x_1 if d_2 and d_3 are both 0, this requires that the cubic and quadratic terms in *s*(*x*) must also disappear for *x* < x_n, so to be a natural spline the *n*+4 coefficients in the cubic spline must satisfy the two sets of the following equations as

$$d_2 = 0,\ \sum_{i=1}^{n} c_i = 0, \qquad \ldots\ldots\ldots (6)$$

$$d_3 = 0,\ \sum_{i=1}^{n} x_i c_i = 0. \qquad \ldots\ldots\ldots (7)$$

The four properties of cubic splines (McKinley and Levine, 2012) that will need to conform to the following stipulations.

1. The piecewise function *s*(x) will interpolate all data points.
2. *s*(x) will be continuous on the interval $[x_1, x_n]$

3. s'(x) will be continuous on the interval $[x_1,x_n]$
4. s''(x) will be continuous on the interval $[x_1,x_n]$

2.4.3 Graphical methods for data analysis of Part III

Line chart

Line charts are used to illustrate number of deaths in all-cause by age groups in both sexes with spline interpolation.

Area chart

An area chart displays graphically quantitative data. It is based on the line chart. The area between axis and line are commonly emphasized with colors, textures and hatchings. They are used to represent cumulated totals using numbers or percentages (stacked area charts in this case) over time. It is like the plot chart except that the area below the plotted line is filled in with color to indicate volume. Area graphs were used to display the number of HIV- estimated deaths by age groups.

Thematic map

The thematic maps in this part display values based on logistic regression model results after applied to DR data including proportions of HIV-estimated deaths for comparison between geographical areas. HIV-estimated deaths from 1996-2009 vary over the period (minimum of 2.45% in 2005 to maximum of 16.49% in 2002). To avoid no level in any years with fixing level of percentages, percentile rank was used to classify the percentages of HIV-estimated deaths. Furthermore, percentile rank clearly illustrated percentages of HIV-estimated deaths in each province that would be

changed over the period by itself. These percentages of HIV-estimated deaths were classified into three levels based on percentile ranks as described in Part II.

All data analysis in three parts were undertaken using R software (R Development Core Team, 2012) including graphical displays (Murrell, 2006).