Chapter 1

Introduction

1.1 Background

Thai deaths registration (DR) database are known to be of poor quality (Prasartkul *et al*., 2007; Rao *et al*., 2010; Vapattanawong and Prasartkul, 2011). They presented missing of province or age varied from 0.7% - 7% and 33.2% - 41.7% of ill-defined cause in 1996-2004 (Bureau of Policy and Strategy, 2010). High percentages of ill-defined lead to inaccuracy of causes of death.

Missing data can occur at the unit level or at the item level. Their mechanisms are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Failing to treat missing data caused serious problems. First, missing data can bring up potential bias in parameter estimation and weaken the generalizability of the results (Little and Rubin, 2002). Second, ignoring cases with missing data leads to decreased statistical power and increased standard errors (Dong and Peng, 2013). Most statistical procedures are designed for complete data. Before analyzed data with missing, it needs to be edited in some ways into a "complete" data set.

Missing rate of 5% or less is inconsequential but over 10% missing is worrisome. The cause of missing usually determines the direction of bias (Dong and Peng, 2013). Thus, ignoring missing data of important variables will welcome spurious results.

HIV (Human immunodeficiency virus) deaths are most commonly under-reported causes of death in many countries including Thailand (Tangcharoensathien *et al*., 2006; Khonhan, 2009; Rao *et al*., 2010), Botswana (Taffa *et al*., 2009), Brazil (Fazito *et al*., 2012; Pacheco *et al*., 2011) and South Africa (Yudkin *et al*., 2009; Birnbaum *et al*., 2011). This problem is most commonly caused by misclassification in causes of death. Under-registration or misclassification causes of death cannot be assumed uniform across the population.

1.2 Rationale of the thesis

Many studies were conducted about Thai mortality (Prasartkul and Vapattanawong, 2006; Odton *et al*., 2010a, 2010b; Vapattanawong and Prasartkul, 2011) but none of them concerned to handle missing data. Pigott (2001) suggested that model-based methods (maximum likelihood: ML and multiple imputation: MI) hold more promise for dealing with difficulties caused by missing data and appropriate for a wide range of situations than the common used ad hoc methods (complete case analysis and available case analysis or pairwise deletion). Allison (2012) found ML method is better than MI method. Thus, Part I involved missing demographic factors that was carried out with the study of “Estimation of mortality with missing data using logistic regression”.

In addition, misclassification causes of death in Thai DR data using a verbal autopsy method was conducted in several studies (Choprapawon *et al*., 2005; Rao *et al*., 2010; Pattaraarchachai *et al*., 2010; Polprasert *et al*., 2010 and Porapakkham *et al*., 2010). However, none of them estimated using modeling. Furthermore, the estimation of mortality of the 2005 VA study was mentioned by Byass (2010) that some uncertainties remained and suggested the probabilistic modeling. Logistic regression model was used to handling missing or misclassification data in several studies (Duffy *et al*., 2004; Lyles *et al*., 2011; Williams *et al*., 2005). Therefore, Part II involved under-reporting/misclassification of HIV deaths, the study of "Correcting and estimating HIV mortality in Thailand based on 2005 verbal autopsy data focusing on demographic factors, 1996-2009" was carried out using logistic regression model.

The effects of missing data and misclassification causes of death lead to inaccurate and unreliable information. Faramnuayphol *et al*. (2008) and Odton *et al*. (2010b) explored geographical variations in all-cause mortality without dealing with missing demographic data. In addition, Faramnuayphol *et al*. (2008) analyzed overall mortality using standardized mortality ratio (SMR) that is not a model based and used causes of death from the ten revision of the International classification of diseases (ICD-10 codes) (WHO, 2010) without correcting misclassification. These cannot clearly address pattern of Thai mortality. Finally, after handling missing demographic data and correcting misclassification causes of HIV deaths by imputation of demographic factors and reclassification of causes of death, all-cause and HIV mortality trends were explored in Part III.

1.3 Objectives of the thesis

1) To impute missing values of demographic data in the death registry from 1996-2009.

2) To examine under-reporting/misclassification and to estimate HIV mortality in all provinces of Thailand during 1996-2009 from a model based on the 2005 VA data.

3) To analyze patterns of all-cause and HIV Thai mortality after correcting the

DR data by imputation of demographic factors and reclassification of cause of deaths.

1.4 Literature review

Mechanisms of missing data

Missing data occur at two levels: at the unit level where not any information at all is collected from the respondent, or at the item level where only certain items were missing. The first part of this thesis concerns only the problem of missing at item levels.

For missing at the item level, the mechanisms of missing may be completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Ignoring MCAR will not introduce bias but will increase the standard errors of the sample estimates due to the reduce sample size. Thus, MCAR poses less threat to statistical inference than MAR or MNAR. MNAR occurs when the probability of missing depends on the missing value itself (Little and Rubin, 2002; Dong and Peng, 2013; Pigott, 2001).

DR data quality

In 1995, civil registration system was developed from manual paper based to electronic centralized and online system. A 13 digit identification number was assigned to each Thai citizen. The civil registration system of Ministry of Interior provided electronic death and birth data directly to the vital statistics management process of the Ministry of Public Health in 1996. These gradually improved on the reporting of deaths from 76% in the Survey of Population Change (SPC) 1985-1986 to 95% in the SPC 1995-1996. However, the quality of vital registration has been questioned (Prasartkul and Vapattanawong, 2006; Tangcharoensathien *et al*., 2006; Prasartkul *et al*., 2007; Rao *et al*., 2010; Vapattanawong and Prasartkul, 2011).

Prasartkul and Vapattanawong (2006) conducted a demographic surveillance system of the Kanchanaburi project. They reported that unregistered death was 12.5%. Furthermore, a survey by Setting Priorities using Information on Cost-Effectiveness analysis (SPICE) project carried out in 2005 found that under registration of DR was around 9.2% (Porapakkham *et al*., 2010). Subsequently, Vapattanawong and Prasartkul (2011) conducted a study of under-registration of deaths in 2005-2006 and reported the unregistered death was 8.7%. This estimation may be the minimum underestimated level of the actual rate due to the fact that not all sample data could be analyzed (11.3% of incomplete information and 10.3% of missing personnel identification number). Carmichael (2011) explored Thailand's mortality and found significant under-registration and age-selective under-enumeration.

Mathers *et al*. (2005) assessed the global status of causes of death data at the end of 2003. They found only 64 countries of 115 countries had complete data. Coverage of DR varies from less than 10% in Africa to 100% in European. Only 23 of the 115 countries had high quality data (more than 90% completeness and ill-defined causes less than 10% of registered). The percentage of ill-defined causes was reported more than 40% in Thailand and Sri Lanka. Thailand is among the 28 countries that has low quality (completeness <70% or ill-defined causes >20% of registered) of DR.

From 1996-2009, Thai DR data ill-defined cause varies from the lowest of 33.2% in 2003 to the highest of 41.7% in 1999 as shown in Table 1.1 (Bureau of Policy and Strategy, 2010). Sixty five percent of deaths occurred at home lead to high percentage of ill-defined cause. It contributed to the inaccurate causes of death because they were certified by village heads with limited knowledge and expertise in causes of death. However, some deaths in hospital that were certified by the physicians also had problems of inaccuracy. Furthermore, the DR system clearly underestimates under-five mortality rates (Tangcharoensathien *et al*., 2006; Hill *et al*., 2007; Rao *et al*., 2010; Thai Health Information Standards Development Center, 2012). Odton *et al*. (2010a) analyzed ill-defined and unknown causes of death outside the hospital in 1999-2001 and found high percentages of ill-defined causes in fatal children aged less than 5 years and adults aged 60 years and over.

Table 1.1 Percentages of ill-defined cause from 1996-2009

year	% ill-defined	year	% ill-defined
1996	35.49	2003	33.18
1997	34.52	2004	37.91
1998	37.64	2005	38.22
1999	41.67	2006	38.14
2000	40.89	2007	38.23
2001	38.03	2008	37.90
2002	37.59	2009	38.00

Faramnuayphol *et al*. (2008) examined geographical variation of DR in Thailand in 2000. They found high overall mortality in the middle part of the upper North and followed by the other parts of the North, the upper North-East and the East. Mortality in age 30-44 years most found in the middle part of the upper Northern region. Odton *et al*. (2010b) studied geographical variation in all-cause in age-specific death rates from various districts in Thailand using mortality data from 1999 to 2001. They found a higher average mortality in the Northern region and the Northern part of the North-East region whereas lower than average was found in the Southern region (except Phuket and Narathiwat). Both studies found high mortality in the Northern region but neither of them mention about deaths reported with unknown region or age.

Vapattanawong and Saplon (2011) reported that deaths outside residential areas from 1996-2009 and revealed that deaths outside residential areas were around 11%-13% and increased from 2.9% in 1996 to 12.6% in 2006. Male deaths were higher than females with the exception of females aged 75 years and over higher than males at the same age range. For age 15-49 years, male deaths were higher than female deaths about 3 times. Bangkok had the highest outside residential occurred followed by the Central, the North-East, the Northern and the Southern regions. Deaths in age 15-24 years were more likely occurred outside residential area with the significant from external causes of death. Prasartkul *et al*. (2000) conducted a study of age and sex structures in old age in Nakhon Pathom, Thailand. They found that incomplete mortality data increases with age. The completeness of aged 75-79 years in males still alive was 88.4% and 91.9% in females, and decreased in aged over 100 years to 12.5% in males and 27.3% in females. This indicated that incomplete data increasing with old age.

Verbal autopsy (VA) method

VA is a research method that helps to determine probable causes of death in cases where there was no medical record or formal medical attention given. Since DR data incompleteness, VA surveys can be used to determine individuals' causes of death. It has been widely used in several countries including Uganda, China, Brazil, Tanzania, Bangladesh, South Africa, Zimbabwe and Thailand to give more accurate information about causes of death (Prasartkul *et al.*, 2007; Lopez *et al.*, 2011; Choprapawon *et al.*, 2005; Mathers *et al*., 2005).

Choprapawon *et al*. (2005) verified causes of death between July 1999 and December 1999. The found 29% of overall agreement causes of death between routine DR and VA. The Kanchanaburi project (Prasartkul *et al*., 2007) reported 45.2% matched causes of death between VA and physicians' verification of hospital records from 1999-2003. They mentioned that percentages would be higher if external causes were included.

The latest VA study was carried out in 2005 by the SPICE project. The sample was selected from four regions of Thailand in nine provinces and 28 selected districts. The study sample was selected from the national DR database using a multistage stratified clustered approach. Thailand was stratified into four broad regions as Central, North, Northeast and South as well as Bangkok. In each of the four regions, the samples were inflated by 15%, and by 50% for Bangkok. Provinces were ordered according to numbers of registered deaths in 2005 and divided into two strata at the 50th percentile in each region.

One province was randomly selected from each stratum. The inflated regional sample was then distributed between the two provinces proportionate to the number of deaths registered in 2005. In each province, districts were ranked according to the number of registered deaths in 2005 and similarly divided into two strata at the 50th percentile. The samples in the district level were selected according to probability-proportional-to-size (PPS). In each selected district, the samples were randomly selected without replacement from all deaths registered in 2005. For Bangkok, the study sample was distributed across the inner, middle and outer concentric zones of Bangkok metropolitan area and one district was randomly selected from each zone.

For a subsample of hospital deaths, validation characteristics used medical records including VA methods. Deaths outside health facilities, the VA diagnoses were compared to DR diagnoses for the same deaths. Kappa method was used to determine reliability of registration diagnoses due to no reference diagnoses available to measure validity of those deaths. Results from the 2005 VA study comprised a set of four papers conducted by Rao *et al*. (2010), Pattaraarchachai *et al*. (2010), Polprasert *et al*. (2010) and Porapakkham *et al*. (2010).

Rao *et al*. (2010) verified causes of death in Thailand based on 9,644 cases. They found annual mortality statistics from DR systems are of limited utility because about 40% of deaths are registered with unknown or nonspecific causes. They found stroke (9.4%) was the leading causes of death among males and followed by transport accidents (8.1%), HIV/AIDS (7.9%), ischemic heart diseases (6.4%) and chronic obstructive lung diseases (5.7%). Among females, the leading causes were stroke (11.3%), diabetes (8%), ischemic heart disease (7.5%), HIV/AIDS (5.7%) and renal diseases (4%). The study has limitations in terms of method used for data collection and the generalizability of the study findings to derive final estimation for Thailand.

Pattaraarchachai *et al*. (2010) verified cause-specific mortality patterns among hospital deaths and found 35% of all deaths occurred in hospitals. About 15% of hospital deaths are registered with nonspecific diagnoses. They found relative increasing proportion of deaths in hospitals due to stroke, ischemic heart disease, transport accidents, HIV/AIDS and diabetes, respectively. Extrapolating the findings to estimate national in-hospital mortality has to do with caution because certain provincial or district-based epidemiological variations may not be adequately captured in the assessment of hospital based mortality in Thailand. Furthermore, estimates of cause-specific mortality at young ages make it very difficult to derive meaningful because of the limited numbers of sample deaths.

Polprasert *et al*. (2010) validated causes of death occurred outside hospitals. They found DR tends to under diagnosis of important causes such as diabetes, liver cancer, and tuberculosis (TB), while undercounting deaths from HIV/AIDS, liver diseases, genitourinary (essential renal), and digestive system disorders. The limitation of this study lies in the underlying uncertainty of VA methods arising from recall and/or information bias resulting in biased responses, as well as the potential for inconsistency in the application of diagnostic guidelines by physician reviewers.

Porapakkham *et al*. (2010) estimated of mortality undercount. The "capture-recapture" methods were applied to the 2005-06 SPC. They found 9.2% under registration deaths and they estimated mortality “envelope”. Proportionate mortality distributions were applied to this mortality "envelope" and ill-defined causes redistributed according to Global Burden of Disease methods to yield final estimates of mortality levels and patterns in 2005. After correction, stroke is the leading causes of death in Thailand (10.7%), followed by ischemic heart disease (7.8%) and HIV/AIDS (7.4%). Other leading causes are road traffic accidents (males) and diabetes mellitus (females). Estimated mortality is at least twice what is estimated in DR data. The limitations are completeness of mortality register remains uncertain and the potential biases arising from the sampling design of the field verification study.

HIV mortality

In several countries, HIV mortality is believed to be under-reported in death registry such as Botswana, Brazil, South Africa including Thailand. HIV DR reported deaths in Thailand only 2% and increased to 10% after correcting by VA method (Choprapawon *et al*., 2005; Tangcharoensathien *et al*., 2006). In Botswana, HIV deaths in hospital were 29%. After correcting misclassification, HIV deaths were estimated to be 48.8%-54.4% (Taffa *et al*., 2009). Birnbaum *et al*. (2011) reported HIV/AIDS accounting for 2.0-2.5% of all registered deaths in South Africa and rose from 19% to 48% after correcting. Fazito *et al*. (2012) reported 27% misclassification of AIDS deaths in Brazil, with mainly reported as ill-defined cause. Pacheco *et al*. (2011) found HIV-infected individuals in Rio de Janeiro, Brazil are under-reporting among older individuals and those with higher CD4 (cluster of differentiation 4) counts. All of these studies show under-reporting of HIV deaths due to misclassification in causes of death.

Woradet *et al*. (2012) found that factors of sex, marital status, race, occupation, residential area, complimentary care, type of patient, sexual behavior and risk of infection are significantly affected mortality among HIV/AIDS patients in the Southern region of Thailand.

Statistical methods for handling missing data and mortality estimation

Many studies ignored to handle missing data. Only 21% of studies deal with missing data. Several methods are used to handling missing data as listwise data deletion, pairwise data deletion, mean substitution, regression method, hot deck imputation, expectation maximization approach, raw maximum likelihood (ML) methods and multiple imputation (MI) (Prasithwattanasaeree and Prasithwattanasaeree, 2008-2009).

Fallah and Kharazmi (2008) introduced four alternative methods (weighting, last-group, progressive and additive methods) to handle with unknown age in cancer registry data compared to conventional method. The conventional method involved multiplying either summary measure based on known age by total number of cases of cancer of the same type in persons of the same sex divided by the number occurring in persons of known age. Four methods calculate summary statistics with four different assumptions that differ from the conventional method. Cases with unknown ages were allocated to the old age groups in all four methods. The results in all methods were different. They concluded that conventional and weighting methods are not based on acceptable assumptions. The last group method is not stable due to defining the last age group. Progressive and additive methods have more acceptable assumptions. They preferred progressive method above all methods because it can produce an age-specific curve with the expected exponential increase.

Pigott (2001) reviewed several methods for handling missing data. Many researchers used ad hoc methods (complete case analysis and available case analysis or pairwise deletion) or single-value imputation but they required assumptions about the data. It was demonstrated using the asthma study to examine the relationship of student's self-efficacy in controlling their asthma. The comparison among complete cases, ML and MI methods suggested that model-based methods (ML and MI) hold more promise for dealing with missing data but they required specialized computer program and assumptions about the nature of the missing data.

Allison (2012) used ML and MI methods to handle missing data. The five scenarios were demonstrated to dealing with missing data on the dependent variable and on predictor variables. It was found that ML method is better than MI method. The advantage of ML over the imputation is that, there is no potential conflict between an imputation model and analysis model. ML method produces the same results in the same data set whereas MI method involves random draws, when applied it to data set, it will get different parameter estimates.

Duffy *et al*. (2004) used logistic regression model to handle missing data of outcome misclassification. This method is effective to handle missing data based on the assumptions of symmetric or asymmetric misclassification. The assumptions in the context of individual study should be careful consideration.

Li *et al*. (2004) used the missing-indicator method and conditional logistic regression in match case-control study with missing exposure values. They found both methods yielded unbiased estimates. Conditional logistic regression provided a slight advantage in terms of bias and coverage probability, at the cost of slightly reduced statistical power and efficiency.

Lyles *et al*. (2011) used logistic regression to handle misclassification of binary outcome. Simulation studies illustrated the effectiveness of the ML approach. Williams *et al*. (2005) also used logistic regression to perform incomplete-data classification. Conditional density functions are estimated using Gaussian mixture model (GMM), with parameter estimation performed using both expectation maximization (EM) and Variational Baysian EM (VB-EM). They found the propose methods are superior to standard imputation procedures. When a small amount of data is available to build the GMM, in which case the VB-EM is superior the EM.

Odton *et al*. (2010a) used logistic regression model to adjust proportions of ill-defined and unknown causes of mortality that regarded to adjust missing ages and regions. However, Faramnuayphol *et al*. (2008), Odton *et al*. (2010b), and Vapattanawong and Prasartkul (2011) ignored to mention about missing data.

The Thai Working Groups on HIV/AIDS projections 2005 (2008) used Asian Epidemic Model (AEM) to estimate HIV situation in Thailand. This method provided by UNAIDS/WHO that widely used in several countries in the world. The AEM projection methods used the epidemiological and behavioral data from many sources. The data used in the model include several components such as sentinel sero-surveillance data (Bureau of Epidemiology, 2013), conscript sero-prevalence data (Royal Thai Army and Armed Forces Research Institute of Medical Sciences), sexually transmitted infection data (Bureau of AIDS, TB and STIs, MoPH), injecting drug user (IDU) data (Bangkok Metropolitan Administration and Thailand MoPH-

U.S. CDC Collaboration), men who have sex with men (MSM) data (Thailand MoPH-U.S. CDC Collaboration and MSM groups) and antiretroviral therapy data (Bureau of AIDS, TB and STIs, MoPH). Different sets of data are used to calculate the estimation of HIV prevalence for generalized and concentrated epidemics. The precise estimates depend on the quality of HIV data were used and limitations of methods. The limitations of this method need lots of data set and cannot be separated to sub-epidemic by geographic area in one dataset.

Summary of literature review

The review indicates that:

- Thai DR data are incompleteness and inaccurate causes of death. Incomplete mortality data increased in old age. Higher mortality was found in the Northeastern part whereas lower was in the Southern part of Thailand. Deaths outside residential occurred around 11%-13% and mostly occurred in Bangkok and the Central region whereas the lowest in the South.
- VA is widely used to verify causes of death in several countries where DR data are of poor quality or not available. In the 2005 VA study, stroke is the leading causes of death in Thailand in both sexes. HIV is the third rank for male and the fourth rank for female.
- HIV deaths are under-reporting in many countries including Thailand. These could be due to misclassification causes of death.
- Several methods can be used to handle missing data. Logistic regression was found to be the appropriate method to handle missing data and outcome misclassification.

These findings suggested this thesis divided into three parts. The work flow of each part is shown in Figure 1.1.

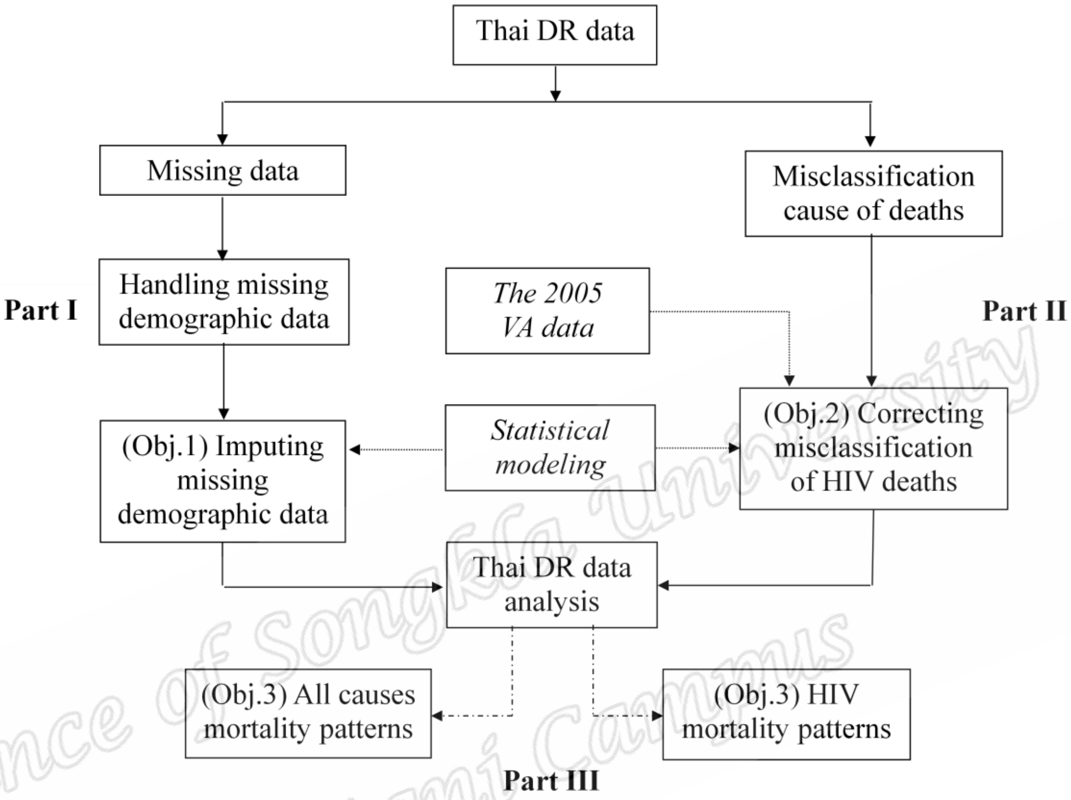
Thai DR data
Missing data
Misclassification cause of deaths
Handling missing demographic data
Part I
The 2005 VA data
Part II
Statistical modeling
(Obj.1) Imputing missing demographic data
(Obj.2) Correcting misclassification of HIV deaths
Thai DR data analysis
(Obj.3) All causes mortality patterns
(Obj.3) HIV mortality patterns
Part III

Figure 1.1 Workflow of this thesis

1.5 Path diagram and variables

In the Thai DR data, province and age were missing but not sex. The first part focuses on how to impute unknown province or age in DR data from 1996-2009. Logistic regression models were used. Sex and age-group are used to predict and fill up unknown province (Model a). Also, sex and province were used to predict for outcome of unknown age (Model b) that inflated from model (a). The determinants and outcomes of two models are shown in path diagrams of Figure 1.2.

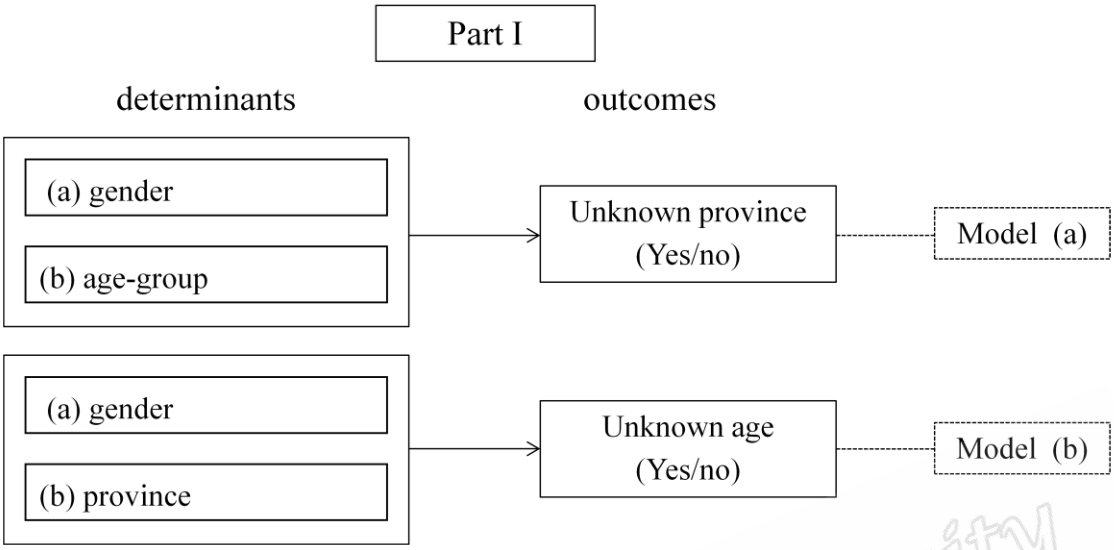

Figure 1.2 Path diagram of unknown province or age in Part I

The second part focuses on correcting the number of HIV deaths from 1996-2009 based on the 2005 VA data. Logistic regression was used to predict causes of death whether death is related to HIV and then impose the model on DR to predict HIV mortality from VA (Model B) with DR cause-location groups, sex-age groups and province. The path diagrams are shown in Figure 1.3.

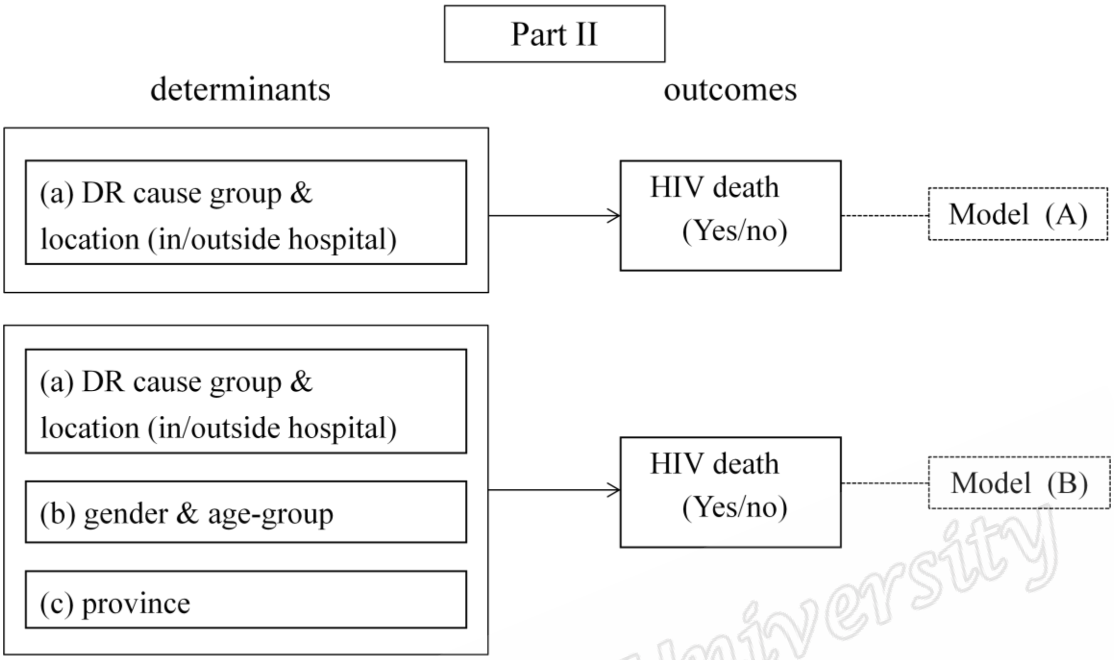

Figure 1.3 Path diagram of correcting HIV deaths in Part II

Finally, the results from the first two parts were explored for patterns of all-cause and HIV mortality from 1996-2009 by sex, age groups and province in the third part as shown in diagram of Figure 1.4.

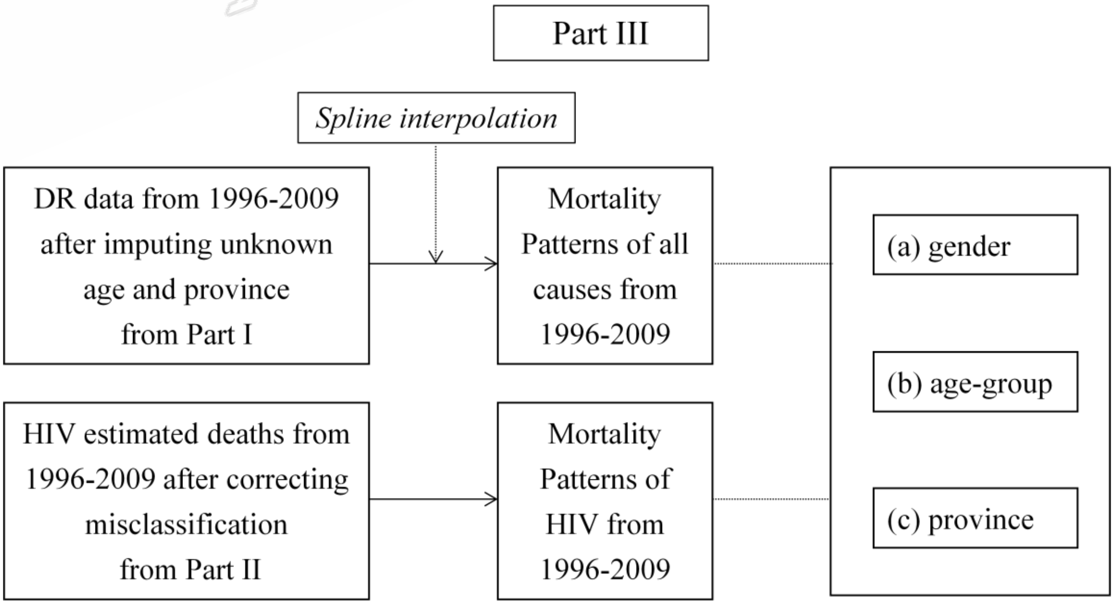
Part III
Spline interpolation
DR data from 1996-2009 after imputing unknown age and province from Part I
Mortality Patterns of all causes from 1996-2009
HIV estimated deaths from 1996-2009 after correcting misclassification from Part II
Mortality Patterns of HIV from 1996-2009
(a) gender
(b) age-group
(c) province

Figure 1.4 Diagram of Thai mortality patterns of all-cause and HIV in Part III