PREDICTING MORTALITY AMONG LABORATORY-CONFIRMED AVIAN INFLUENZA A H7N9 PATIENTS: RISK CLASSIFICATION MODEL BASED ON VARIABLE SELECTION FROM MULTIPLE IMPUTATION

by

HEEJUNG SON

(Under the Direction of Ye Shen)

ABSTRACT

The clinical risk points system makes complex statistical models practical and convenient for clinical use. This risk points system helps clinicians make their decisions for the treatment process quickly with its characteristic as a scientific tool for predicting risks of diseases or incorporating effective evidence-based approaches. To develop the clinical risk points system for data with missing observations, variable selection arises as one of the statistical problems with multiple imputation (MI). Also, we are confronted with the challenge of developing a simultaneous risk points system with multiply-imputed datasets. In our study, we suggest a multiple imputation-stepwise method (MI-Stepwise) across multiply-imputed data to yield a consistent variable selection. Simulations are conducted and we apply the methods to the Asian lineage avian influenza Asian H7N9 virus (A/H7N9) study in the China Centers for Disease Control and Prevention (China CDC) to predict death.

INDEX WORDS: clinical risk points system; risk scores; multiple logistic regression; H7N9; Stepwise methods; Rubin’s rule; multiple imputation

PREDICTING MORTALITY AMONG LABORATORY-CONFIRMED AVIAN INFLUENZA A H7N9 PATIENTS: RISK CLASSIFICATION MODEL BASED ON VARIABLE SELECTION FROM MULTIPLE IMPUTATION

by

HEEJUNG SON

B.S., Utah Valley University, 2015

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2019

PREDICTING MORTALITY AMONG LABORATORY-CONFIRMED AVIAN INFLUENZA A H7N9 PATIENTS: RISK CLASSIFICATION MODEL BASED ON VARIABLE SELECTION FROM MULTIPLE IMPUTATION

by

HEEJUNG SON

Major Professor: Ye Shen
Committee: Kevin K. Dobbin
Stephen Lynn Rathbun

Electronic Version Approved

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2019

ACKNOWLEDGEMENTS

I wish to thank Dr. Ye Shen for his encouragement and direction for this paper and appreciate my committee members' helpful comments. I am also grateful that I could have a lot of supports around me for my master's degree.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1. Background

Influenza is one of the main health issues in China as well as in the world. Influenza outbreaks have posed a major threat and caused significant concern due to how easily this disease spreads. Influenza virus can be transmitted by direct contact and aerosol transmission human-to-human, or zoonosis[1]. According to recent studies, Asian H7N9 (A/H7N9) virus is an Asian lineage avian influenza virus, first diagnosed in humans in early 2013, and since then over 1600 people have been infected from five epidemic waves [2-4].

A major clinical characteristic of the disease is that respiratory systems can experience rapid progressive pneumonia followed by respiratory failure, which leads to mortality rates above 30% [5]. Human to human transmission of the A/H7N9 is rare [6, 7]. However, the pandemic potential of avian influenza is a significant concern. In 2017, the Centers for Disease Control and Prevention (CDC) designated the A/H7N9 virus as the rapid-growing potential risk for sustained human-to-human transmission and risk factors of global public health of all influenza A viruses [8, 9]. It is not known how well the mortality predictor applies to a wide range of epidemic diseases while the high mortality rate has been observed among patients in A/H7N9. A few studies have reported the risk of death, and most of them investigated the risk factor with small sample sizes or contained few overall characteristics of the risk factor [10-15].

Also, to verify results of the risk factor, independent validation cohorts for confirmation were not obtained in previous studies.

Therefore, we study a large cohort with specific epidemiological and clinical characteristics of A/H7N9 patients identified in laboratories in Zhejiang province, southeastern China. To test the performance of the model of the risk factor to predict mortality, we investigate the risk of death in patients and develop a risk classification model that can be clinically useful in identifying and prioritizing patients with the highest mortality probability.

1.2. A/H7N9 data overview

1.2.1. Study participants and data collection

During the first case of the A/H7N9 epidemic in China in April 2013, enhanced monitoring of the A/H7N9 was implemented as part of the Chinese surveillance system. Inpatients with pneumonia or similar symptoms to influenza were classified as having suspected A/H7N9 virus infection. Once infection of the A/H7N9 was suspected, respiratory specimens were first collected, and then demographics and clinical surveys were conducted for all patients and accompanying family members using standardized forms. Epidemiological data were gathered from interviews and field observations of patients diagnosed with suspected A/H7N9 infection by local and national CDC field teams within 1 day. All medical information was reported to CDC in China, but no microbiological A/H7N9 confirmation prior to site and patient data collection was required for suspected H7N9 infected patients.

1.2.2. Derivation Cohort – A/H7N9 patients from Zhejiang province

All laboratory-identified cases of the A/H7N9 infection presented to the Information System for Disease Control and Prevention in Zhejiang province in China were classified to the derivation cohort [16]. The location of these A/H7N9 patients was geospatially mapped and followed up for their subsequent mortality. Information on demographics, exposure history, clinical symptoms, and relevant dates in disease process was collected by a standardized questionnaire. Activities related to exposure history such as visiting live poultry markets, intra-household poultry raising, occupational exposure, and direct contact with diseased or deceased poultry within two weeks of clinical onset were asked for laboratory-identified H7N9 cases. Also, prior diagnoses of chronic and/or noninfectious diseases such as hypertension, chronic pulmonary disease, diabetes, and cardiovascular disease were inquired from patients. Timelines of disease and health-care related processes for each case were arranged as follows: Dates of onset of the illness, first visit to a medical care facility, hospitalization, antivirus treatment initiation, and confirmatory laboratory test results. Clinical characteristics of the A/H7N9 infection were recorded by respiratory specialists; moreover, whether unilateral or bilateral lung infections were present in the patients was recorded.

1.2.3. Laboratory diagnostic procedures

RNA extraction was examined from throat specimens. Also, these specimens were tested using a specific real-time reverse transcription polymerase chain reaction (RT-PCR) with primers and probes specific to H7N9.

Patients, who were suspected to have the H7N9 infection but who were confirmed as negative for three consecutive days, were considered disease-free and were not tested anymore.

Other laboratory measurements included white blood count, neutrophil count percent, lymphocyte count percent, body temperature, and levels of C-protein. These measurements were collected when patients initially suspected A/H7N9 infections and was performed at multiple points throughout the course of the disease. We used the results of lab measurements taken during the first clinical visit (timing furthest from death) since we want to predict mortality.

1.3. Purpose of Study

Our study aims to build a clinical risk score point system to predict the risk of mortality when A/H7N9 is diagnosed. In other words, it may help physicians provide more effective medical treatments and direct therapies under intensive clinical monitoring to the patients receiving high-risk scores. Yet, missing values, whether it is significant values or not, would lead to bias in data analysis since variable selection are sensitive to missing values and their missing mechanisms [17]. The A/H7N9 data contains a lot of covariates with missing values. Among 19 covariates considered, 15 covariates include missing values. Especially, the missing proportion in smoking is conspicuously significant to 60.7% (185). The variables of Chronic drug use and the C-reactive protein are also remarkably indicated to 48.2% and 43.6% missing, respectively. Therefore, we will use multiple imputation under the assumption that data are Missing at Random (MAR) to have identical variable selections among multiply-imputed data. Here, we used MI-Stepwise [18] variable selection, and MI-LASSO [19] is also applicable. Based on

significant variables selected from MI-Stepwise, the risk point system was carried out to predict individual survival probabilities.

1.4. Literature review

1.4.1. Missing data

Missing values are one of the most common potential problems in data analysis. A significant amount of missing information will affect data analysis and cause issues with further analysis. The first solution to missing data, typically the default selection method of statistical packages, is list-based deletion or pairwise deletion. However, from the default packages for missing data, variables that might otherwise be significant may not be selected through statistical procedures such as forward, backward, or stepwise variable selection. Also, variables with a large proportion of missing data would be sorted out before statistical analysis. It should be considered whether or not missing variables may have a significant impact on the outcome. Usually, it is important to keep the data rather than delete it. Moreover, imputing missing information is often preferred rather than dropping all that information. Multiple imputation (MI), which was proposed by Rubin [20], is one of the commonly used methods for filling in missing values. Unlike single imputation in which one value is inserted for each missing value, multiple imputations substituted for each missing value with two or more values sampled from the conditional probability distribution of the imputed variable given ancillary variables. As a result, more than one complete dataset is created. Multiple imputation can provide unbiased statistical results given an explicitly specified imputation model [21] and provide parameter

estimates and standard errors that take into account the uncertainty due to missing data values [17].

MI is a popular method in practical use under missing completely at random (MCAR) and missing at random (MAR) mechanism; yet, MI might produce incorrect results under a missing not at random (MNAR) mechanism. MCAR is defined as the probability of missing data on a variable that does not depend on itself and any other variable in the dataset subject to analysis [22]. MAR is denoted as the probability that a datum is missing may depend on observed characteristics but not on unobserved characteristics of the subject [23]. In addition, missing data is MNAR, which is neither MCAR nor MAR, when the probability of a missing variable is related to the value of the missing datum.

Maximum likelihood estimation (MLE) utilizing expectation-maximization (EM) algorithm [24] and Bayesian estimation are two useful methods for data analysis with observed data without imputing missing values; MLE obtains statistical inferences based on the marginal distribution of observed data [25]; Bayesian estimation is based on the observed data likelihood and a prior distribution for the parameter. Monte Carlo Markov Chain simulation is to produce a sample from the joint distribution of the parameters and the missing data given the observed data [17]. While both of these two methods need sophisticated computation and different computations for different statistical models, the MI method is generally straightforward to be implemented and interpreted.

1.4.2. Variable Selection

To fit a proper model for statistical analysis, the choice of statistical selection methods and/or correct conditions of the methods need to be considered. For example, the subset selection method checks all combination of variables, and then checks models for the best fit based on significant criterion, such as adjusted R^2, the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC), Mallow's C_p, Mean Square Error, and Predicted Residual Sum of Squares (PRESS), etc. However, the subset selection method may not be the best variable selection method if many candidate predictors exist. In addition to fitting all the predictable subsets, there are more efficient selection methods including Forward Selection and Backward Elimination. Forward Selection starts to run with an empty model. It adds variables one at a time and tests the how well the model fits newly added variables. It continuously runs until the variable does not affect the suitability of the model. On the other hand, Backward Elimination starts with a full model including all candidate predictors, and then eliminates them one at a time. The Backward Elimination model is performed in reverse to the process of the Forward Selection model. Nevertheless, Forward and Backward selection methods are not guaranteed to find the best model [26]. Thus, a method of combining the Forward Selection and Backward Elimination models has been proposed: Stepwise regression, which was introduced by Efroymson [27]. Stepwise regression fits models based on prespecified criterion such as significance levels and Mallow's C_p. For each iteration of the Stepwise selection method, it adds and/or remove variables and runs until a model is returned that satisfies the given criterion. While these selection methods are more practical for datasets with large sample sizes and relatively small number of candidate variables, penalized regression can more effectively conduct variable

selection if there is a large number of candidate variables or the number of its variables is greater than the number of observations.

Variable selection methods via penalized likelihood are broadly performed these days [28]. Penalties are divided into K-Smallest Items (KSI) penalties family, which contain the least absolute shrinkage and selection operator (LASSO), the Self-adaptive penalty, and the Log-Exp-Sum penalty [28].

The LASSO, introduced by Robert Tibshirani in 1996 [29], processes regularization by minimizing the residual sum of squares with the restriction in the sum of absolute values of the coefficients. Hence, it improves regression model fitting in prediction accuracy and interpretability. Bayesian variable selection strategies are also frequently applied in many instances [30].

Considering variable selection procedures after MI, intuitively it is natural to directedly apply variable selection methods to imputed datasets one at a time. However, this could generate different selections of variables among imputed datasets, lead to unreliable parameter estimates, and make scientific conclusions challenging. For this reason, new methods of variable selection based on data from MI have been conducted; According to Heymans et al [31]., a variable selection can be applied to each imputed dataset separately, and then based on selections for each dataset we pick those common predictors for a single model under bootstrapping with automatic backward regression; Wood et al [18]. proposed a backward stepwise selection under a weighted regression applied on an integral dataset, which was attained by stacking k multiply-imputed datasets. They also proposed a MI-stepwise method, which is a stepwise variable selection method for multiply-imputed data using repeated applications of Rubin's rules [32, 33]. Chen and Wang [19] utilize MI-LASSO selection, which combines coefficient estimates for each

variable in k imputed dataset in a group LASSO penalty, and then adds or removes the whole group together. In this thesis, we will apply MI-stepwise, in which each selection step is based on the combined P-value processed by Rubin's rules, to our multiply-imputed datasets for the variable selection.

1.4.3. Clinical risk points system

Multivariable models used for estimating clinical risks have been developed for medical studies of diseases [34, 35]. These models of the risks allow us to quantify the effect of measurable risk factors on diseases. The Framingham Heart Study [36] has led to developments generating estimates of risk of coronary heart disease and help with creating the models, which can be practical for selecting appropriate treatments. For this reason, we apply the risk points system method to our A/H7N9 study for the clinical use.

There are often various risk factors associated with diseases, so it is ideal to consider all possible disease risk factors that can be measured in clinical practice. However, some verified risk factors for diseases are not always considered if it takes a lot of time to measure, needs expensive or dangerous testing procedures, and has difficulties with unquantified data [36]. In general, some risk factors can be measured accurately and be accessed easily and expeditiously. Also, restricting risk factors is an important practical way to readily generate the multivariable risk score models and to reduce noise and errors. These risk score models are often generalizable to other populations because they mainly include a limited number of clinically significant risk factors that are relatively easy to measure. While the distributions of the risk factors and the incidence rate of the outcome event, etc., are associated to the problems that influence

transportability, many of these issues can be solved with minor adjustments to the models. Details has been provided by a step-by-step tutorial following the Framingham Heart Study [46].

The Framingham Study has developed multivariable models to quantify the impact of various risk factors and to adopt a multifactorial disease process since they produced initial multivariable models for coronary heart disease in the 1960s. The first models were generated based on logistic regression and discriminant function analysis [37-39]. Models were updated using techniques of survival analysis as data were accumulated, i.e. serial assessments of the risk factors and longer follow-up for events [34, 35, 40]. In addition, the Framingham Study has generated models for specific events such as stroke [41, 42], peripheral vascular disease [43] and congestive heart failure [44] and for subsequent events, based on repeat events in persons who have a history of coronary disease [45]. The function that best predicted the likelihood of the events based on easily trackable and measurable risk factors was determined by the underlying goals in each of these models. Though these models used to estimate the risk are studied and developed for the long term by the Framingham Study, over the years they have expanded their applications to populations that differ ethnically, racially, according to risk factor prevalence or event incidence.

For our study, we aim to develop a risk point system for the risk scores using a multiple logistic regression model. The risk point system simplifies computation of $\sum \beta X$, and is derived by assigning integer points to each level of each risk factor. By summation of these integer points, we can estimate $\sum \beta X$ for a specific risk factor profile, and then a reference table providing risk estimates for each point total is produced. The points system is conducted around categories, but distinct values for the continuous risk factors can be contained.

CHAPTER 2

METHODOLOGY

2.1. Multiple imputation

Multiple imputation is generally carried out using a Bayesian approach or sequential regression imputation (SRMI). First, the Bayesian approach for imputing data under multivariate normal, log-linear, and the general location model is based on Markov Chain Monte Carlo (MCMC). This approach specifies full multivariate models for imputed variables, and then produces a posterior predictive distribution for missing data imputations which is fully conditional on observed values and unknown parameters. Yet, it is not simple to generate the joint distribution of all variables including missing values with real data. It is difficult since real data generally contain a large number of variables and with different types of distributions and consist of sophisticated data structures. If variables in the data like count data have restrictions or bounds, it could make it hard to generate the distributions, too.

SRMI is also known as multivariate imputation by chained equations (MICE) and fully conditional specification (FCS) and allows imputing multiple times on relatively complex data structures under assuming the existence of a joint distribution for variables. Each variable is successively imputed, in order from the variable with the smallest to the largest numbers of missing observations. In each step, imputation is conditional on all observed and previously imputed data. Process of this approach [46] is to first impute variables having the least amount of

missing values by specifying relevant regression models given other variables. Different regression models may be applied to different types of variables. Then these first imputed variables specified an appropriate regression model given other variables are used for the next imputation of other variables. This process operates based on regression models, conditional on all other observed or imputed variables, for each variable until all missing values are imputed, and this whole process is iterated until it converges.

Van Buuren S, Boshuizen HC, Knook DL [23] suggested what variables should be included or excluded from imputation models. Variables that will be included in the model for analysis, variables correlated with the imputed variables and variables related to the presence of the imputed variables are recommended for inclusion. Covariates will be removed if they have a large number of missing entries in observations. Similarly, Schafer [47] suggests including all possible inclusive variables for the imputation model. Then, variables, which are associated with the imputed variables and the absence of the imputed variables, will be selected to generate high-quality imputations for missing entries of a particular variable.

In our A/H7N9 data, we first include most of all variables based on demographic, clinical and laboratory characteristics for producing the high-quality imputation model. Types of variables in A/H7N9 data are continuous or categorical (binary and ordinal). We assume the data are MAR and impute the data with the FCS approach, which is a powerful and statistically valid method for creating imputations in large data sets containing both categorical and continuous variable. This can be achieved by using several R packages, which are ‘mice’ [48] and ‘mi’ [49], or the IVEware package [50]. We impute the categorical and continuous missing data using the logistic and linear regression methods, respectively. Five multiply-imputed datasets are generated.

2.2. Rubin's rules for multiple imputation inference [20]

Each independently imputed dataset is analyzed by the same method as for completed datasets, and then Rubin's rules (RR) are applied to get a combined estimate from the *D* estimates calculated from the *D* imputations. Let $\widehat{Q}_\iota$ and $\widehat{U}_\iota$, $d = 1, \ldots, D$, denote the point estimate of interest and their associated variances for a population parameter *Q*, calculated from the *D*th imputation. Then, the combined estimate of parameter *Q* from *D* imputations is the average of *D* point estimates:

$$\bar{Q} = \frac{1}{D} \sum_{i=1}^{D} \widehat{Q}_\iota.$$

The variance of $\bar{Q}$ has two components to obtain a valid standard error: the average within-imputation variance,

$$\bar{U} = \frac{1}{D} \sum_{i=1}^{D} \widehat{U}_\iota,$$

and the between-imputation variance, which describes the variability from imputation uncertainty,

$$B = \frac{1}{D-1} \sum_{i=1}^{D} (\widehat{Q}_\iota - \bar{Q})^2.$$

The combined variance related to $\bar{Q}$ is

$$T = \bar{U} + \left(1 + \frac{1}{D}\right) B.$$

When sample size is large,

$$T^{-\frac{1}{2}} (Q - \bar{Q}) \sim t_\gamma,$$

where the degrees of freedom is

$$\gamma = (D-1) \left\{ \frac{\bar{U}}{(1 + D^{-1}) B} \right\}^2.$$

Thus, a $100(1-\alpha)\%$ confidence interval of $\bar{Q}$ is

$$\bar{Q} \pm t_{\gamma,1-\frac{\alpha}{2}}\sqrt{T}.$$

Wald's test can be used for the hypothesis test:

$$H_0 : Q = Q_0,$$

where Q_0 is the null value, by comparing the test statistic, $W = \frac{(Q_0-\bar{Q})^2}{T}$, against the critical value of $F_{1,\gamma}$.

When covariates of regression models in each imputation are different, RR cannot be used to get combined coefficient estimates. After a regression model is selected, this same model is fitted on each imputation, then RR can be used. Consequently, it enables us to have a variable selection method that generates a relevant selection across all imputed datasets.

2.3. MI-Stepwise

MI-Stepwise variable selection is similar to general stepwise selection. However, they are distinguished by the process of selection; in order to add, remove or keep variables in models, the normal stepwise method depends on significance test using *P*-value and two significance levels for entering and removing variables while MI-stepwise uses a combined *P*-value. In MI-Stepwise, each imputation obtains *P*-values for a specific variable, and those *P* -values are organized together by RR under MINALYZES procedure. This combined *P* -value will be used as the determinant of actions that add, remove or keep variables. Then, the selection procedures are jointly run across all imputed datasets, and the same actions will be conducted on each variable in all imputed datasets. Wood, White, and Royston [18] depicted MI-stepwise variable

selection by repeated use of RR. Following the detailed procedures present in the MI-stepwise selection method:

Step 0: Choose α_1, α_2 for *P*-value to enter and *P*-value to remove, respectively. Specify the model with no covariates, denoted the initial model M_0. Set *t = 0.*

Step 1: Let *t = t + 1*. For each covariate *X* that is not contained in model M_{t-1}, fit *D* regressions with the model $\{M_{t-1}, X\}$ on *D* imputed datasets. Estimate the combined *P*-value for each newly added $p_a \leq \alpha_1$, and then renew the model M_t to be $\{M_{t-1}, X_a\}$; otherwise, $M_t = M_{t-1}$, and the procedure terminates.

Step 2: Refit *D* regressions with the model M_t on *D* imputed datasets and computed the combined *p*-values for covariates *X* in the model. Let X_b be the covariate with the largest combined *P*-value p_b. If $p_b > \alpha_2$, place the model M_t to be $\{M_t, X_b\}$, where the minus sign denotes removing X_b from M_t.

Step 3: Repeat step 2 until the largest combined *p*-value p_b is smaller than or equal to α_2, $p_b \leq \alpha_2$.

Step 4: Go back to step 1 and iterate *step 1* and *step 2* until the procedure terminates.

Terminating the iteration of MI-Stepwise, the combined *p*-values for all the covariates in the model should not be larger than α_2, and if covariates are added into the model, their combined *p*-values should be less than α_1. To avoid infinite iteration, the condition of $\alpha_1 \leq \alpha_2$ should be given.

2.4. Algorithm for the clinical risk points system or clinical risk scores [51]

Now we describe the general approach for generating a clinical risk points system based on regression models, such as multiple linear or logistic regression, Cox proportional hazard regression, etc. In the following steps, we describe the risk points system in the multiple logistic regression model to help us generate the points system of the mortality risk in A/H7N9. Also, since our modeling outcome data are binary (death for 1 and survival for 0), the risk points system in multiple logistic regression model can be processed.

2.4.1. Estimate the parameters of the multivariable model

Suppose the model $f(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, where *Y* is the dependent or outcome variable; Using logistic regression as an illustrating example, where $Y = 1$ denotes the presence of a particular event; Y = 0 denotes the absence of the event. The function *f(·)* is a logit link function connected to a linear combination of the risk factors $X_1, \ldots, X_p$ (X_i $i = 1, \ldots, p$, can include continuous, dichotomous, or categorical risk factors). The parameters $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients.

2.4.2. Organize the risk factors into categories and determine reference values

Suppose a risk factor is continuous, then we need to set up adjacent classes and choose reference values for each. Specifically, to determine points for each category it is important to specify a reference value for each category; thus, the mid-points approach is commonly

acceptable. If a risk factor has a bounded range of continuous values, it is obvious to determine the reference values. For example, if the range of risk factor X_1 is 0-39, we can categorize X_1 into 0-9, 10-19, 20-29, 30-39, and use 4.5, 14.5, 24.5 and 34.5 as reference values, respectively. However, there are some exceptions if extreme values or outliers exist in a risk factor. For example, if the range of risk factor X_2 is 80-210, we may use the five categories <120, 120-129, 130-139, 140-159, ≥160. It is straightforward to calculate the mid-points for the three middle risk factor categories. Otherwise, we apply the following mid-points approach: The reference value for the first category should be included in the range of 119 or less. Since there could be some extreme values in the distribution of X_2 (e.g., the minimum is 80), the mid-point between 119 and the 1th percentile of the observed systolic blood pressures is a more robust mid-point for the first category. Suppose the 1st percentile is 89, then mid-point for the first category is computed as 104. Similarly, the reference value for the last risk factor category (≥160) can be obtained using the same strategy.

If a risk factor is dichotomous, modelled as an indicator variable or consists of a set of dummy variables (e.g. each coded as 0=absent or 1=present) reflecting distinct categories of the risk factor, then the reference value is simply either 0 or 1.

Here, W_{ij} denotes the reference value (e.g. mid-points for continuous risk factors arranged into categories, or values 0 or 1 for risk factors modelled by a set of dummy variables or a single indicator) for the *j*th category of the *i*th risk factor, where $i=1,\ldots, p,$ and $j=1,\ldots, c_i,$ where c_i=the total number of categories for risk factor *i.*

2.4.3. Determine the referent risk factor profile

Subsequently, we define the appropriate category for each risk factor to serve as the base category, which will be assigned 0 points in the point scoring system. Higher scores denote higher risks in general practice. Thus, categories obtaining worse states of the risk factor will be contributed to positive points, while categories reflecting better states will be assigned negative.

Let W_{iREF} denote the reference value of the base category, for each of the *i* risk factors $i=1,\ldots, p$.

2.4.4. Determine how far each category is from the base category in regression units

Next, we calculate how far each category is from the base category W_{iREF}, in terms of regression units. In other words, we will determine the number of points for each of the categories of each risk factor and decide the following for each category *j* of each risk factor *i:*

$$Points_{ij} = \beta_i(W_{ij} - W_{iREF}),\ i=1,\ldots, p, \text{ and } j=1,\ldots, c_i.$$

Note that the base category of each risk factor will be designed 0 points from this formula.

In our study, it is problematic if we use different points for each of the categories of each risk factor across multiple datasets since we imputed missing values based on the multiple imputation. In order to generate the simultaneous points for each risk factor with multiply-imputed datasets, we combined all imputed datasets into a single one, for which we are able to apply the mid-points approach to specify a common reference.

2.4.5. Determine risks associated with point totals

In the final step of generating the point system, we will assign the estimates of risk (or probability of developing an event over the predetermined time frame) based on each point total. It requires the use of the exact model to set up the estimates of risk. The following formula is the typical model obtained for risk estimation, $\hat{p}$, along with the multiple logistic regression:

$$\hat{p} = \frac{1}{1 + exp(-\sum \boldsymbol{X\beta})}$$

Basically, the risk points system is to approximate the contribution of the risk factors in the estimate of risk, particularly, to estimate $\sum_{i=1}^{p} \beta_i X_i$, which is the component of each model shown above that relies on the specific risk factor profile.

The estimates of risk include the total number of points, which approximates $\sum \boldsymbol{X\beta}$, into the formula $\hat{p}$; the risk estimates in the risk points system is based on specific risk factor profiles. For this reason, there are some issues for the presence of an intercept term and handling of continuous risk factors.

Intercept term: Notice that we have not included a separate point allocation for an intercept for the points system. In order to approximate $\sum \boldsymbol{X\beta}$, the estimate of the initial value for the intercept β_0 should be included.

Continuous risk factors: In 2.4.2., categories for the continuous risk factors were generated and each reference value specified. In the next step, we chose a base risk factor category and assigned 0 points to them. After that, we added up all the points because we basically estimated how far a particular individual's risk factor profile is from the referent profile. Here, we note that the $\sum_{i=1}^{p} \beta_i X_i$ term obtains a particular risk profile and not the distance from the referent risk factor profile. It is important that both the referent risk factor profile and

the distance from that profile should be added to approximate the relevant $\sum_{i=1}^{p} \beta_i X_i$ for the risk estimate.

CHAPTER 3

SIMULATION STUDY

3.1. Generating data

In our study we will design the point system from the risk factors and conduct the estimates of the risk to evaluate the finite sample performance from MI-stepwise variable selection methods under two missing mechanisms: (i) MCAR; (ii) MAR

For simulation studies, all datasets consist of 14 variables and 300 observations which are sampled from a multivariate standard normal distribution, which has a mean zero and variance of one, and a compound symmetric correlation structure. 14 continuous variables (***X***'s) with a binary response variable (***Y)*** are generated in our models. ***Y*** is given by the logit function linked to the regression model below:

$$logit(E[\boldsymbol{Y}|\boldsymbol{X}]) = \boldsymbol{X}\,\boldsymbol{\beta}.$$

This logistic regression model is given as the generalized linear regression model, where predictors, 1, 5, 10, 11, consisting in the model are significant variables. The coefficients $\boldsymbol{\beta} = (\beta_1, \beta_5, \beta_{10}, \beta_{11})^T$ in the logistic regression are all set as 1.

The simulations in all scenarios are repeated 100 times each. To evaluate the performance of MI-Stepwise variable selection method under MCAR and MAR missing mechanisms, three criteria presented below would be computed:

sensitivity of selection (SEN)

$$SEN = \frac{\#\ of\ selected\ important\ variables}{\#\ of\ true\ important\ variables},$$

specificity of selection (SPE)

$$SPE = \frac{\#\ of\ removed\ unimportant\ variables}{\#\ of\ true\ unimportant\ variables},$$

and geometric mean of sensitivity and specificity (G)

$$G = \sqrt{sensitivity \times specificity}.$$

The range of *G* is between 0 to 1, and a desirable value for selecting variables correctly would be computed close to 1. According to Kubat et al. [52], this geometric criterion shows the distinctive independent property of the numbers of important and unimportant covariates. Therefore, the geometric mean of sensitivity and specificity was computed for overall performance measurement.

Mean squared errors (MSE) are used to evaluate the performance of point estimates of risk from MI-Stepwise variable selection under different missingness mechanisms and varied situations. Assuming the estimates of the risk $\hat{p}$ and the empirical estimates of the risk *p*, which we will define here as subgroup mortality rate, depends on each *k*th point total, the MSE can be estimated by

$$MSE = \frac{1}{N}\sum_k n_k(\hat{p}_k - p_k)^2,$$

where the sum is over the available observations at the *k*th point total, n_k is the number of the *k*th point total, and *N* is the total number of observations in each dataset, which is the sum of the observations of the *k*th point total $\sum_k n_k$.

3.2. Generating missing data

We assume that Y is fully observed and generated missing data in the 14 covariates of $\boldsymbol{X}$ under the missing data mechanisms. We regarded ignorable missing mechanisms for the point system of the risk: (i) missing completely at random (MCAR) and (ii) missing at random (MAR).

For MCAR, each variable was independently dropped by some missing percentages (i.e. 3%, etc.) in X_1 to X_{14} to obtain a missingness scenario with ~70% complete cases. Moreover, MAR is created by the following logistic regression model to generate the binary missing data indication R_{ij}:

$$logit\{\Pr(R_{ij} = 0 \mid X_{i(j\pm7)})\} = \alpha_0 + X_{i(j\pm7)},$$

where α_0 is given to control the average missing percentage of variable. Similarly, to yield datasets with about 70% complete cases, each variable was independently dropped by various missing percentages. Given the logistic regression model, a function of α_0 shown below describes the expected missing percentage for each variable:

$$f_i(\alpha_0) = \frac{1}{n}\sum_i \frac{exp(\alpha_0+X_i)}{1+exp(\alpha_0+X_i)}.$$

3.3. Results

We will compare the performances of the risk point system under full data, complete cases with different missing data mechanisms and multiply-imputed data from MI. In MI-Stepwise, we set up *p*-value thresholds for including and removing a variable; let $\alpha_1 = 0.05$ and $\alpha_2 = 0.06$, respectively.

We considered two simulations to compare the point system in our study: (i) full data, complete datasets from two different missing mechanisms and its multiply-imputed data; (ii) overall comparisons under different overall missing proportion of datasets.

3.3.1. Simulation one

In this simulation, we have the sample size N = 300 and number of covariates $N_p = 14$. Missing values were generated by MCAR and MAR missing mechanisms including 70% complete cases. For MCAR, we dropped 3% of each candidate variable in $X_1 - X_{14}$ including about 70% complete cases. Similarly, for MAR, we yielded about 70% complete cases using the logistic model of the expected missing percentage for each variable. The true regression coefficients β_i's were set to 1 for i=1, 5, 10, 11 and otherwise $\beta_i = 0$.

Table 3.1. Mean Sensitivity (SEN), Specificity (SPE) and their Geometric Mean (G): The Stepwise Regression on Full Cases, the CC-Stepwise on Two Different Missing Mechanisms (MCAR, MAR) including 70% Complete Cases of 300 Observations and the MI-Stepwise on Multiply-Imputed Datasets among 100 Simulations.

		SEN	**SPE**	**G**
70% complete cases	Full data			
	Stepwise	99.8	93.7	96.7
	MCAR			
	CC-Stepwise	98	96	97
	MI-Stepwise	99.8	96.6	98.2
	MAR			
	CC-Stepwise	98.8	95.4	97.1
	MI-Stepwise	99.8	96.8	98.3

CC-Stepwise, complete cases stepwise; MI-Stepwise, multiply-imputed cases stepwise.

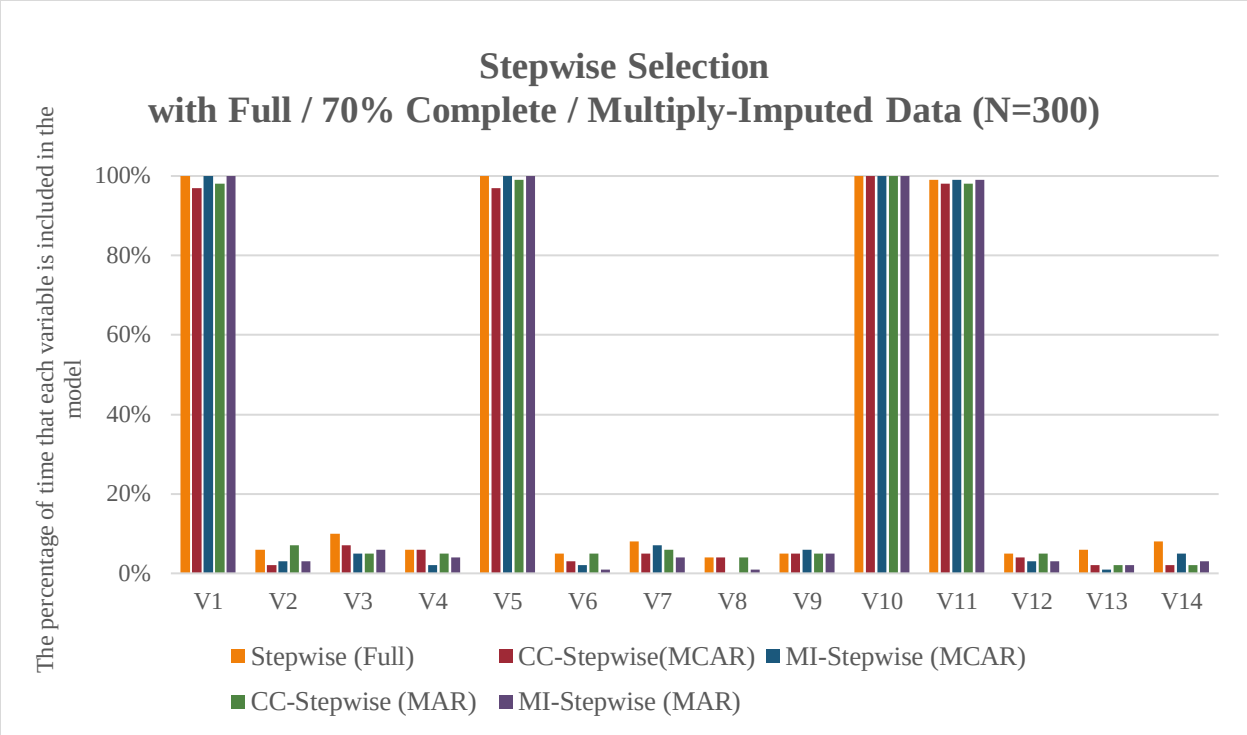

Figure 3.1. Each Variable (in percentage) Selected from Stepwise Methods including Full Data, 30% Missing Cases of 300 Observations, and Multiply-Imputed Data among 100 Simulations. All significant variable's (X_1, X_5, X_{10}, X_{11}) coefficients set to 1. Overall missing percentage of datasets are 30% of 300 observations and missing values are generated under MCAR and MAR.

Figure 3.1 shows the results of stepwise variable selection methods from full data, complete cases under MCAR and MAR missing mechanisms, and multiply-imputed cases; we denote stepwise methods as Stepwise for the full data, CC-Stepwise for the complete data and MI-Stepwise for the multiply-imputed data. From overall 30% missing datasets, all important variables are selected after multiple imputation across MI-Stepwise selection methods. After multiple imputation with MI-Stepwise variable selection, the probability of selecting important variables into the model increases in both missing mechanisms, i.e. it increases from 98% in CC-Stepwise to 99.8% in MI-Stepwise for MCAR, and from 98.8% to 99.8% for MAR. In addition, 97% in CC-Stepwise of the Geometric criterion increases to 98.2% in MI-Stepwise for MCAR,

and further improvements are observed for MAR as well. The MI-Stepwise methods on both missing mechanisms have similar SEN, SPE, and G values compared with the Stepwise method applied on the full data with no missing values. It suggests that MI-Stepwise competes well as the Stepwise method applied on full data in terms of identifying important variables. In Table 3.1. we listed results on different criteria for the performance of the variable selections.

Table 3.2. Mean Squared Errors (MSE): Evaluating the Performance of the Risk Point Systems Generated based on the Stepwise Regression on Full Cases, the CC-Stepwise under MCAR and MAR Missing Mechanisms including 70% Complete Cases of 300 Observations and the MI-Stepwise on Multiply-Imputed Datasets among 100 Simulations.

Stepwise (Full)	**CC-Stepwise (MCAR)**	**CC-Stepwise (MAR)**
0.022	0.025	0.030
	MI-Stepwise (MCAR)	**MI-Stepwise (MAR)**
	0.021	0.022

CC-Stepwise, complete cases stepwise; MI-Stepwise, multiply-imputed cases stepwise.

Table 3.2. shows that the MSE of the risk point system from MI-Stepwise on each MCAR and MAR are slightly higher than the MSE of the risk scores computed from the full datasets. However, these MSE from multiply-imputed datasets are improved (lower) under both MCAR and MAR missing mechanisms compared with the MSE from the 70% complete cases datasets which deletes incomplete observations. MSE for MCAR is dropped about 16% from 0.025 to 0.021; similarly, it drops about 25% from 0.030 to 0.022 for MAR. The results in general support the notion that we could achieve better performances in developing the risk point systems on imputed datasets than the complete case analysis under ignorable missing mechanisms.

3.3.2. Simulation two

In this simulation, we increase overall missing proportion to 50% complete cases in the same sample size to the previous simulation (N=300) and leave other major parameters unchanged. For MCAR, we generate ~50% complete cases by dropping 5% of each candidate covariate in $X_1 - X_{14}$. MAR datasets include about 50% complete cases as well. The true regression coefficients β_i's are set to 1 for i=1, 5, 10, 11 and 0 otherwise.

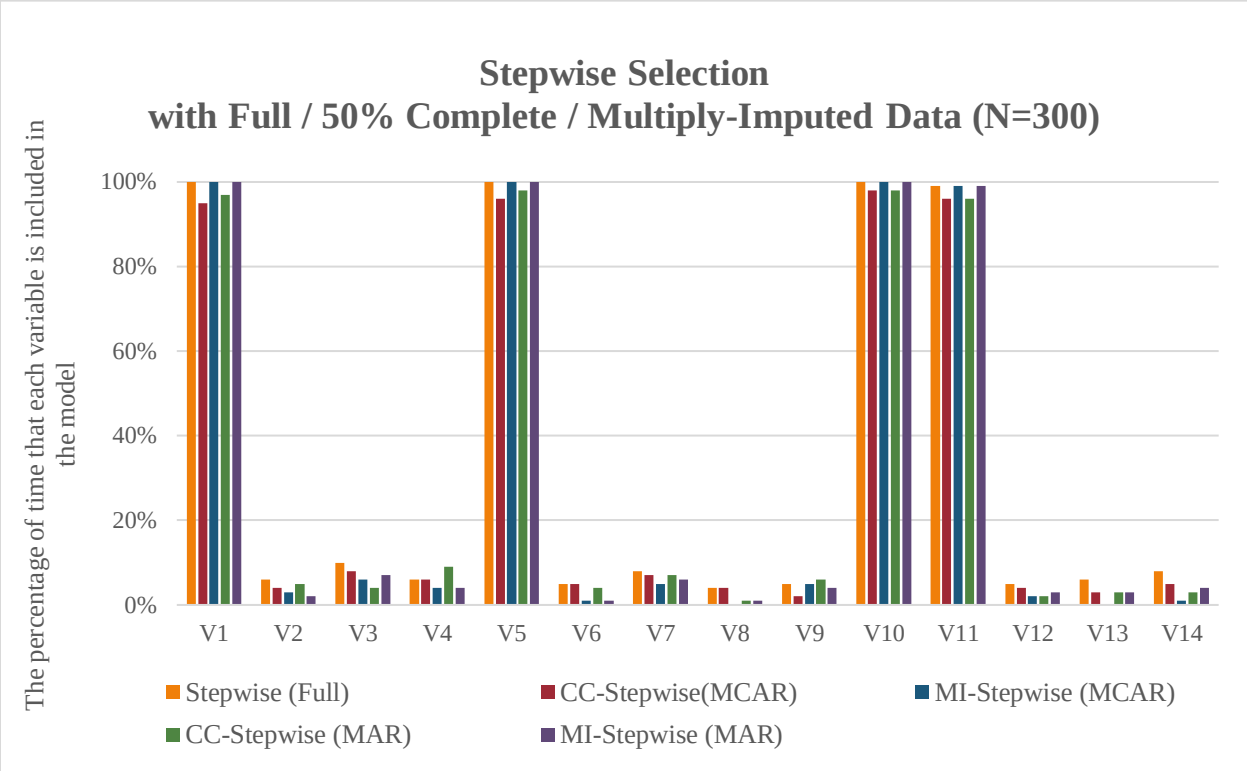

Figure 3.2. Each Variable (in percentage) Selected from Stepwise Methods including Full Data, 50% Missing Cases of 300 Observations, and Multiply-Imputed Data among 100 Simulations. All significant variable's (X_1, X_5, X_{10}, X_{11}) coefficients set to 1. Overall missing percentage of datasets are 50% of 300 observations and missing values are generated under MCAR and MAR.

Figure 3.2. shows that important variables selected by Stepwise, CC-Stepwise and MI-Stepwise reach over 90% of the total replications as well as Figure 3.1. in simulation one. Generally, after applying MI-Stepwise on multiply-imputed datasets, the correct selection for valuable covariates presents better selection than the selection from the complete case datasets without imputation. Details of the simulation results are shown in Table 3.3. Compared with results from the previous simulation, data with 50% complete cases has lower SEN and G. Otherwise, SPE from missing data scenarios are slightly higher than the results from the full data for both simulations. The SPE in our simulations have stable results from overall stepwise methods in general, with a trade-off of lower SEN in data with missingness, for which MI-Stepwise outperforms CC-Stepwise.

Table 3.3. Mean Sensitivity (SEN), Specificity (SPE) and their Geometric Mean (G): The Stepwise Regression on Full Cases, the CC-Stepwise on Two Different Missing Mechanisms (MCAR, MAR) including 50% Complete Cases of 300 Observations and the MI-Stepwise on Multiply-Imputed Datasets among 100 Simulations.

		SEN	**SPE**	**G**
50% complete cases	Full data			
	Stepwise	99.8	93.7	96.7
	MCAR			
	CC-Stepwise	96.2	95.2	95.7
	MI-Stepwise	99.8	97.3	98.5
	MAR			
	CC-Stepwise	97.2	95.6	96.4
	MI-Stepwise	99.8	96.5	98.1

CC-Stepwise, complete cases stepwise; MI-Stepwise, multiply-imputed cases stepwise.

Table 3.4. presents the results of the MSE on the risk point system generated from the overall stepwise methods with sample size N=300. The risk points system from the full data has small MSE, and the CC-Stepwise risk score has higher MSE than the risk scores from Stepwise

and MI-Stepwise. In general, the MSE based on the full data applied to Stepwise method is considered as the gold standard. The risk point system from data with the MCAR missing mechanism typically results in better MSE than that from data with the MAR missing mechanism. Compared with the previous simulation, high overall missing percentage leads to higher overall MSE, but the general comparisons among different approaches follow the same pattern observed in simulation one.

Table 3.4. Mean Squared Errors (MSE): Evaluating the Performance of the Risk Point Systems Generated based on the Stepwise Regression on Full Cases, the CC-Stepwise under MCAR and MAR Missing Mechanisms including 50% Complete Cases of 300 Observations and the MI-Stepwise on Multiply-Imputed Datasets among 100 Simulations.

		MCAR		MAR	
	Stepwise	CC-Stepwise	MI-Stepwise	CC-Stepwise	MI-Stepwise
Full data	0.022				
50% complete cases		0.030	0.021	0.033	0.021

CC-Stepwise, complete cases stepwise; MI-Stepwise, multiply-imputed cases stepwise.

CHAPTER 4

APPLICATION TO THE H7N9 DATASET

4.1 Application on H7N9

The motivating data on H7N9 in our study was provided from Zhejiang CDC in China. The study dataset included 305 laboratory-identified A/H7N9 patients from Zhejiang province between 2013 and 2018. These observations were diagnosed in 10 prefecture cities: Lishui, Taizhou, Jiaxing, Ningbo, Hangzhou, Wenzhou, Huzhou, Shaoxing, Quzhou, Jinhua. Among patients confirmed with the A/H7N9, the median age was 59 (interquartile range, 49–68), 64% were male, and 51% resided in urban residences. Antiviral treatment was given to 81% of patients, 63% had at least one type of underlying medical conditions, and 30% had 2 or more of them. Poultry exposure was common; 93% had some poultry exposure, and 69% recently visited a live poultry market. However, only 6% were poultry workers.

Overall mortality was 37.7% in our H7N9 data; 115 patients died among 305 observations. For this reason, we were motivated to identify sufficient virus exposure pathways, clinical traits and laboratory examinations and to synthesize a simple, predictive risk point system for clinical use estimating the risk of the mortality. The A/H7N9 dataset initially contained 49 candidate covariates. After variables deemed irrelevant to the study purpose were removed, the final working dataset included 19 covariates. These variables contained demographics (age, sex, smoking), clinical traits (diabetes, hypertension, cardiovascular disease,

pulmonary disease, underlying medical conditions history, admission, antiviral treatment), and laboratory examinations (white blood cells, body temperature in Celsius, chronic drug use, unilateral/bilateral lung infection, pneumonia). More details of candidate variables are shown in Table 4.1. We included dichotomous variables converted from nominal variables (i.e. categorical variables). These final covariates were analyzed containing both 6 continuous and 13 binary variables, and the data analysis for MI strategy was based on the five imputations accomplished by Olson et al [53].

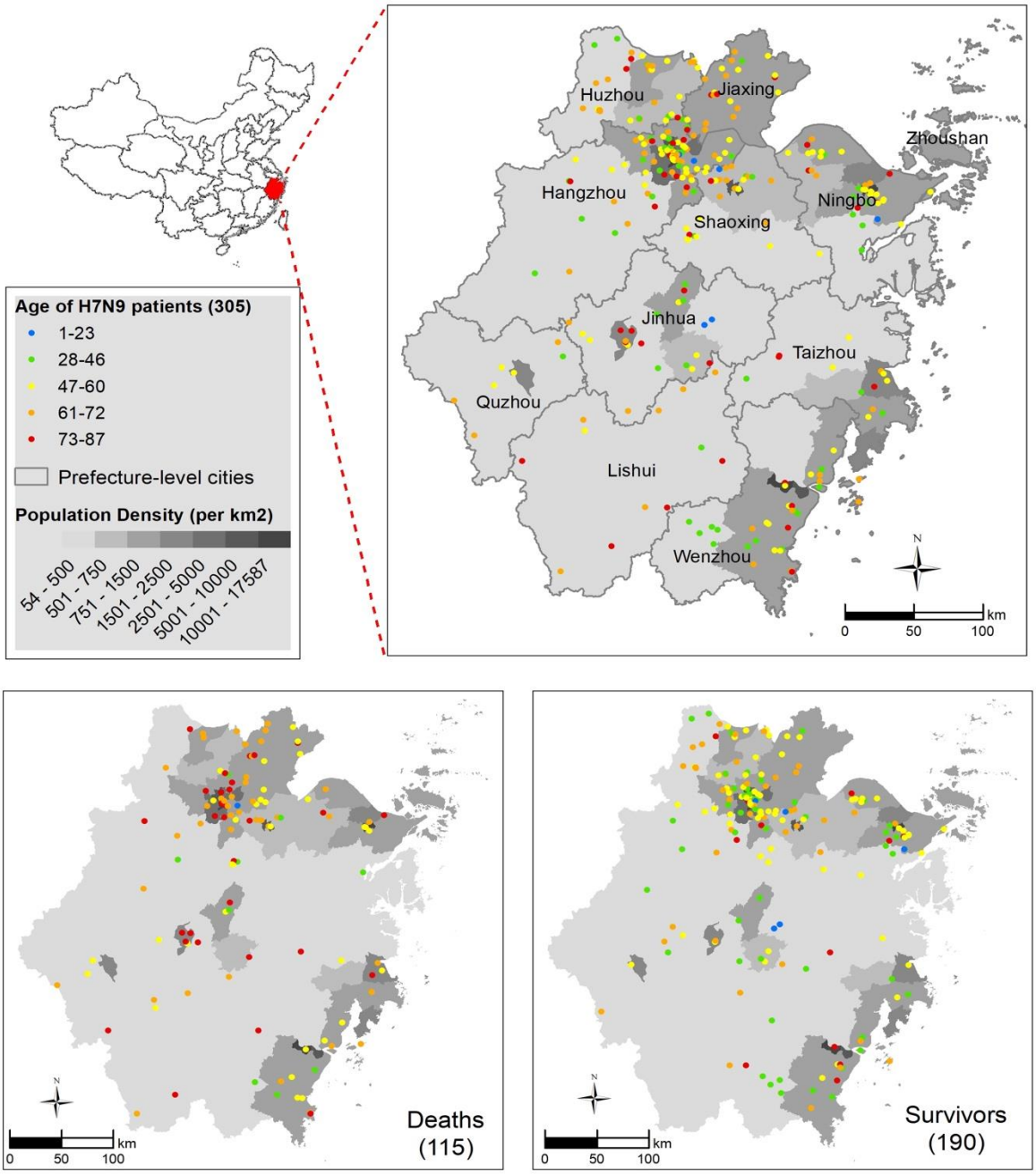

Figure 4.1. Geospatial Coordinates of Patients with Laboratory-Confirmed Diagnoses of A/H7N9 Infection from the Cohort of Zhejiang Province, Southeastern China.

Table 4.1. Demographic, Clinical, and Laboratory Characteristics of 305 Laboratory-identified A/H7N9.

Variable	No. Participants	Percent	Median (IQR)
Demographic Characteristics			
N	305	100	
Median Age, Years (IQR)			59 (49–68)
Age group, years			
<45	59	19.3	
≥45 and <55	55	18.0	
≥55 and <65	86	28.2	
≥65 and <75	64	21.0	
≥75	41	13.4	
Male	194	63.6	
Smoker			
Yes	35	11.5	
No	85	27.9	
Missing	185	60.7	
Clinical Characteristics			
Diabetes			
Yes	54	17.7	
No	230	75.4	
Missing	21	6.9	
Hypertension			
Yes	125	41.0	
No	165	54.1	
Missing	15	4.9	
Cardiovascular disease			
Yes	57	18.7	
No	213	69.8	
Missing	35	11.5	
Pulmonary disease			
Yes	18	5.9	
No	252	82.6	
Missing	35	11.5	

Presence of Underlying Medical Conditions			
Yes	193	63.3	
No	97	31.8	
Missing	15	4.9	
Admission	296	97.0	
Antiviral Treatment	248	81.3	
Laboratory Traits			
Lymphocyte Count Percent			15 (10–22)
Abnormal Lymphocyte Count Percent			
Low, <0.20	129	42.3	
Normal, 0.20 to 0.39	52	17.0	
High, >0.40	9	3.0	
Missing	115	37.7	
Neutrophil Count Percent			78 (71–85)
Neutrophil Count Percent Quartiles			
<0.70	63	20.7	
≥0.70 and <0.79	74	24.3	
≥0.79 and <0.86	63	20.7	
≥0.86	63	20.7	
Missing	42	13.8	
White Blood Cell Count, Microliter			5 (3.6–6.7)
Abnormal White Blood Cell Count			
<3.5	63	20.7	
3.5 to 10.5	208	68.2	
>10.5	19	6.2	
Missing	15	4.9	
C-Reactive Protein, Milligram per Liter			75.5 (35.6–129.3)
Abnormal C-Reactive Protein			
Normal, <10.0	12		
High, 10 to 50	45		
Severe, > 50.0	115		
Missing	133		
Temperature, Celsius			39.3 (38.9–39.8)
Fever (37.8 Celsius or above)	282	92.5	
Missing	14	4.6	

Chronic Drug Use	79	25.9
Missing	147	48.2
Unilateral Lung Infection		
Yes	255	83.6
No	2	0.7
Missing	48	15.7
Bilateral Lung Infection		
Yes	177	58.0
No	43	14.1
Missing	85	27.9
Pneumonia	232	76.1
Missing	64	21.0

IQR, interquartile range.
Percentages refer to within–characteristic column totals among participants within each clinic and in entire study. The total percentages may not be 100% since within-column percentages were rounded to the nearest integer. Column totals vary across different characteristics due to missing values for some participants.

We assumed that the general missing mechanism was MAR in the A/H7N9 dataset and fitted the multiple logistic regression model with its multiply-imputed datasets. Then we applied the MI-Stepwise variable selection method with $\alpha_1 = 0.05$ and $\alpha_2 = 0.06$ to select the A/H7N9 risk factors of the mortality. From these selected risk factors, we then created the clinical risk points system for mortality. The sufficient covariates selected from MI-Stepwise, its coefficient estimates, *P*-values, referent risk factor profiles (W_{ij}) of the clinical risk points system, 95% confidence intervals (CIs), and the risk point systems on each category of each sufficient covariate are presented in Table 4.2.

Table 4.2. Multivariable Logistic Regression Analysis of Derivation of the Clinical Risk Points System Classification in A/H7N9.

Risk Factor	Coefficient (β_i)	Reference (W_{ij})	*P* value	95% CI	Risk Points
Intercept	-6.5883		< 0.0001	(-9.4658, -3.7108)	
Age (years)	0.0574	31.5 = W_{1ref}	< 0.0001	(0.0358, 0.0790)	
<50		31.5			0
≥50 and <60		54			1
≥60 and <68		63.5			2
≥68		76			3
Neutrophil Count Percent Quartile	3.0158	0.46 = W_{2ref}	0.0450	(0.0683, 5.9633)	
<0.70		0.46			0
≥0.70 and <0.79		0.75			1
≥0.79 and <0.86		0.83			1
≥0.86		0.91			1
White Blood Cell Count	0.0586	5.0 = W_{3ref}	0.0066	(0.0164, 0.1008)	
<10.1		5.0			0
≥10.1 and <11.4		10.75			0
≥11.4 and <12.9		24.3			0
≥12.9		14.75			1

The real data analysis suggests that the mortality of the A/H7N9 is positively associated with age, neutrophil count (in percent), and white blood cell count (in microliter). These findings are dependable and consistent since, in general, many studies find that aging, chronic diseases, and immune degradation by bacterial/viral infections increase the mortality rate [54]. Moreover, extreme laboratory results in the neutrophil count test and c-protein test are highly correlated

with the risk of infections. Results from these biomarkers also suggest that severe viral infection increases in the risk of the mortality rate.

Risk Score	Risk Estimate	Mortality				
		Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
0	0.044	0.182	0.100	0.100	0.158	0.105
1	0.112	0.155	0.189	0.178	0.162	0.173
2	0.256	0.325	0.338	0.333	0.333	0.346
3	0.483	0.438	0.389	0.411	0.403	0.394
4	0.717	0.712	0.727	0.722	0.727	0.727
5	0.873	1.000	1.000	1.000	1.000	1.000

Table 4.3. The Points System of the Risk Estimate and the Empirical Risk of the Mortality in A/H7N9.

Following the risk point system approach, we investigated risk scores for each category of each selected covariate in our data. Since age, neutrophil count percent and white blood cell count covariates were continuous, we categorized them based on clinical references following quantiles in the multiply-imputed data. Overall trends of estimated coefficients are positive so that the trend of the risk scores in each variable also relevantly increases in our risk point system. The risk points assigned in ≥ 68 years old remarkably increase the risk of the death in A/H7N9 infection. In Table 4.3., the clinical risk points of the A/H7N9 ranges from 0 to 5, and their estimates of risk and the empirical risk of the mortality increase with the clinical risk points. The same pattern is observed across the imputed datasets.

Findings from our study suggest that MI-Stepwise method with $\alpha_1 = 0.05$ and $\alpha_2 = 0.06$ may be too liberal in selecting the risk factors for death resulted from A/H7N9 infection. Mortality rate among those having pulmonary disease in A/H7N9 infected cases is as high as 61%, a potential risk factor to be further explored. Furthermore, pneumonia, smoking and c-

protein may be regarded as important risk factors of the A/H7N9; pneumonia is associated with a high mortality rate (41%) in the A/H7N9 infection and is relevant to the invasive lung infection [5]; smoking is known as one of major risk factors of lung diseases; c-protein remarks infections or inflammations. Thus, further investigations with less restricted variable selections procedures are warranted.

CHAPTER 5

DISCUSSION

The clinical risk points system makes complex statistical models practical and convenient. Moreover, such systems can aid clinicians to make their decisions for the treatment process quickly with its characteristic as a scientific tool for predicting risks of diseases or incorporating effective evidence-based approaches [51]. However, missing data arise in the problem of generating the risk point system. MI has been a prevalent method for resolving the problems of the missing data since it is easy to implement and available with relevant software. However, we are confronted with the challenge of developing a simultaneous risk points system with multiply-imputed datasets. In our study, we propose to apply a MI-Stepwise method for variable selection after multiple imputation, and to combine all multiply-imputed datasets together so that we include all observations for computing a coincident risk points system that can be generalizable in application.

MI-Stepwise method is also convenient for implementation on the multiple logistic regression model, and the MI-Stepwise specifies significant variables in models as well as the Stepwise on full data. Following our simulations, dropping incomplete observations when performing stepwise selection on complete-cases results in poor sensitivity of selection, particularly if the overall missing proportion is relatively high, and/or the missing mechanism is under MAR. Meanwhile, MI-Stepwise completes variable selection effectively with better sensitivity and specificity, especially for data with tolerable proportions of overall missingness.

The MI-Stepwise method is relatively liberal for selecting significant variables into the model. If we refit the model with variables selected by MI-Stepwise, those variables are mostly significant at level α_1. This feature of the MI-Stepwise appears to be important to yield stable selection of models based on the multiply-imputed data. Overall, we conclude that MI-Stepwise methods applied to multiply-imputed data help generating a valid risk points system.

According to Chen and Wang [19], the MI-LASSO method, which adopts the concept of group LASSO, can be an alternative to the MI-Stepwise since MI-LASSO typically maintains better sensitivities, especially for small sample size data with a large number of covariates. However, MI-LASSO could be subjected to over-selecting issues when the sample size is large.

When MI is used to handle missing data, we often assume that missing information is ignorable. As such, we generate missing data under MCAR or MAR in our simulation study. This can be considered as a limitation of the current study. We also applied MI-Stepwise under different proportion of overall missingness. In terms of MSE, the points systems built on the variables selected through MI-Stepwise under all scenarios of simulation achieves good performances under ignorable missing mechanisms. Simulation scenarios of the risk points system under non-ignorable missing mechanisms can be explored in future studies.

To assess the performance of the risk points system we developed, MSE was used to compare estimated and empirical mortalities. However, it can be challenging to evaluate the risk points system this way in some scenarios. Data with large missing proportions will result in small sample sizes for the complete case analysis, such that the risk points system on the CC-Stepwise is likely to have a short range of possible risk points. Further, a mis-specified imputation procedure applied to dataset with a significant portion of its data being missing could reveal relationships between the outcome and predicting variables with bias. Consequently, the

risk points system developed from variable selection by MI-Stepwise in these scenarios may have poor performances.

In this study, we aim to develop the methodology for the computation of the risk points system on multiple logistic regression model with MI-Stepwise method for data with missingness, for which theoretical properties are still under exploration. Future studies can be extended to investigate the method for cox proportional hazards models as well. We note that assessing the effect of uncertainty in imputation on the variable selection and creating a simultaneous risk points system for multiply-imputed datasets remain as two major challenges in developing a consistent risk points system.

REFERENCES

1. Koutsakos, M., K. Kedzierska, and K. Subbarao, *Immune responses to avian influenza viruses.* Journal of Immunology, 2019. **202**(2): p. 382-391.
2. Lam, T.T.-Y., et al., *The genesis and source of the H7N9 influenza viruses causing human infections in China.* Nature, 2013. **502**(7470): p. 241-244.
3. Gao, R., et al., *Human infection with a novel avian-origin influenza A (H7N9) virus.* The New England Journal Of Medicine, 2013. **368**(20): p. 1888-1897.
4. Li, Q., et al., *Epidemiology of Human Infections with Avian Influenza A(H7N9) Virus in China.* New England Journal of Medicine, 2014. **370**(6): p. 520-532.
5. Gao, H.-N., et al., *Clinical findings in 111 cases of influenza A (H7N9) virus infection.* The New England Journal Of Medicine, 2013. **368**(24): p. 2277-2285.
6. Farooqui, A., et al., *Probable Hospital Cluster of H7N9 Influenza Infection.* The New England Journal Of Medicine, 2016. **374**(6): p. 596-598.
7. Mai-Juan, M., et al., *Influenza A(H7N9) Virus Antibody Responses in Survivors 1 Year after Infection, China, 2017.* Emerging Infectious Diseases, 2018(4): p. 663.
8. Cox, N.J., S.C. Trock, and S.A. Burke, *Pandemic preparedness and the Influenza Risk Assessment Tool (IRAT).* Current Topics In Microbiology And Immunology, 2014. **385**: p. 119-136.
9. Uyeki, T.M. and N.J. Cox, *Global concerns regarding novel influenza A (H7N9) virus infections.* The New England Journal Of Medicine, 2013. **368**(20): p. 1862-1864.

10. Ji, H., et al., *Epidemiological and Clinical Characteristics and Risk Factors for Death of Patients with Avian Influenza A H7N9 Virus Infection from Jiangsu Province, Eastern China.* PLoS ONE, 2014. **9**(3): p. 1-8.

11. Kang, M., et al., *Epidemiology of human infections with highly pathogenic avian influenza A(H7N9) virus in Guangdong, 2016 to 2017.* Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin, 2017. **22**(27).

12. Liu, S., et al., *Epidemiological, clinical and viral characteristics of fatal cases of human avian influenza A (H7N9) virus in Zhejiang Province, China*. 2013, Elsevier Science B.V., Amsterdam: Great Britain. p. 595.

13. Wang, X., et al., *Epidemiology of avian influenza A H7N9 virus in human beings across five epidemics in mainland China, 2013-17: an epidemiological study of laboratory-confirmed case series.* The Lancet. Infectious Diseases, 2017. **17**(8): p. 822-832.

14. Zheng, S., et al., *Benefit of Early Initiation of Neuraminidase Inhibitor Treatment to Hospitalized Patients With Avian Influenza A(H7N9) Virus.* Clinical Infectious Diseases, 2018. **66**(7): p. 1054-1060.

15. Shen, Z., et al., *Host immunological response and factors associated with clinical outcome in patients with the novel influenza A H7N9 infection*. 2014, Blackwell Publishing Ltd: Canada. p. O493.

16. Cheng, Q.-L., et al. *Retrospective study of risk factors for mortality in human avian influenza A(H7N9) cases in Zhejiang Province, China, March 2013 to June 2014*. 2015. Canada: Elsevier Science B. V., Amsterdam.

17. Sinharay, S., H. S. Stern, and D. Russell, *The Use of Multiple Imputation for the Analysis of Missing Data*. Vol. 6. 2001. 317-29.

18. Wood, A.M., I.R. White, and P. Royston, *How should variable selection be performed with multiply imputed data?* Statistics in medicine, 2008. **27**(17): p. 3227-3246.

19. Chen, Q. and S. Wang, *Variable selection for multiply-imputed data with application to dioxin exposure study.* Statistics in medicine, 2013. **32**(21): p. 3646-3659.

20. Rubin, D.B., *Multiple imputation for nonresponse in surveys. [electronic resource]*. Wiley classics library. 2004: Hoboken, N.J. ; Wiley-Interscience, 2004.

21. Yuan, Y. *Multiple Imputation for Missing Data: Concepts and New Development*. 2000. SAS Users Group International.

22. Roderick, J.A.L., *A Test of Missing Completely at Random for Multivariate Data with Missing Values.* Journal of the American Statistical Association, 1988. **83**(404): p. 1198.

23. Van Buuren, S., H.C. Boshuizen, and D.L. Knook, *Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis*. 1999, JOHN WILEY & SONS LTD: Great Britain. p. 681.

24. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), 1977. **39**(1): p. 1.

25. Pigott, T.D., *A Review of Methods for Missing Data*. 2001, SWETS AND ZEITLINGER: Netherlands. p. 353.

26. Miller, A.J., *Subset selection in regression*. Monographs on statistics and applied probability: 95. 2002: Boca Raton : Chapman & Hall/CRC, ©2002.
2nd ed.

27. Efroymson, M.A. *Multiple regression analysis*. 1960. Wiley, New York.

28. Geng, Z., *Variable selection via penalized likelihood*. 2014: ProQuest LLC, Ann Arbor, MI.

29. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.

30. Edward, I.G. and E.M. Robert, *APPROACHES FOR BAYESIAN VARIABLE SELECTION.* Statistica Sinica, 1997. **7**(2): p. 339.

31. Heymans, M.W., et al., *Variable selection under multiple imputation using the bootstrap in a prognostic study.* BMC medical research methodology, 2007. **7**(1): p. 33.

32. Charles, M.S., *Estimation of the Mean of a Multivariate Normal Distribution.* The Annals of Statistics, 1981. **9**(6): p. 1135.

33. Andrew, O.F. and L.C. Nicholas, *yaImpute: An R Package for kNN Imputation.* Journal of Statistical Software, 2007(10).

34. Anderson, K.M., et al., *An updated coronary risk profile. A statement for health professionals.* Circulation, 1991. **83**(1): p. 356-362.

35. Wilson, P.W.F., et al., *Prediction of Coronary Heart Disease Using Risk Factor Categories*. 1998, AMERICAN HEART ASSOCIATION INC: United States. p. 1837.

36. D'Agostino, R.B., *Likelihood Modelling: Presentation of Multivariate Data for Clinical Use: The Framingham Study Risk Score Functions.* Tutorials in Biostatistics Volume 2: Statistical Modelling of Complex Medical Data, 2004: p. 445.

37. Truett, J., J. Cornfield, and W. Kannel, *A multivariate analysis of the risk of coronary heart disease in Framingham.* Journal of Clinical Epidemiology, 1967. **20**(7): p. 511-524.

38. Cornfield, J., T. Gordon, and W.W. Smith, *Quantal response curves for experimentally uncontrolled variables.* Bull Int Stat Inst, 1961. **38**(3): p. 97-115.

39. Walker, S.H. and D.B. Duncan, *Estimation of the probability of an event as a function of several independent variables.* Biometrika, 1967. **54**(1-2): p. 167-179.

40. Kannel, W.B., D. McGee, and T. Gordon, *A general cardiovascular risk profile: the Framingham Study.* The American journal of cardiology, 1976. **38**(1): p. 46-51.

41. Wolf, P.A., et al., *Probability of stroke: a risk profile from the Framingham Study.* Stroke, 1991. **22**(3): p. 312-318.

42. D'Agostino, R.B., et al., *Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study.* Stroke, 1994. **25**(1): p. 40-43.

43. Murabito, J.M., et al., *Intermittent claudication: a risk profile from the Framingham Heart Study.* Circulation, 1997. **96**(1): p. 44-49.

44. Kannel, W.B., et al., *Profile for estimating risk of heart failure.* Archives of internal medicine, 1999. **159**(11): p. 1197-1204.

45. D'Agostino, R.B., et al., *Primary and subsequent coronary risk appraisal: new results from the Framingham study.* American heart journal, 2000. **139**(2): p. 272-281.

46. Van Buuren, S., et al., *Fully conditional specification in multivariate imputation*. 2006, TAYLOR & FRANCIS: Great Britain. p. 1049.

47. Schafer, J.L., *Analysis of incomplete multivariate data*. 1997, London
New York: London
New York : Chapman & Hall.

48. Van Buuren, S. and K. Oudshoorn, *Flexible multivariate imputation by MICE*. 1999: Leiden: TNO.

49. Yu-Sung, S., et al., *Multiple Imputation with Diagnostics (mi) in R : Opening Windows into the Black Box.* Journal of Statistical Software, 2011(2).

50. Raghunathan, T.E., et al., *A multivariate technique for multiply imputing missing values using a sequence of regression models.* Survey methodology, 2001. **27**(1): p. 85-96.

51. Sullivan, L.M., J.M. Massaro, and R.B. D'Agostino, Sr., *Presentation of multivariate data for clinical use: The Framingham Study risk score functions.* Statistics In Medicine, 2004. **23**(10): p. 1631-1660.

52. Collins, L.M., J.L. Schafer, and C.-M. Kam, *A comparison of inclusive and restrictive strategies in modern missing data procedures.* Psychological methods, 2001. **6**(4): p. 330.

53. Olson, K., et al., *Missing data in an environmental exposure study: imputation to improve survey estimation.* Organohalogen Compounds, 2006. **68**: p. 1346-1349.

54. Thompson, W.W., et al., *Mortality associated with influenza and respiratory syncytial virus in the United States.* Jama, 2003. **289**(2): p. 179-186.