The Hong Kong Polytechnic University

School of Nursing

Alternative screening method for potential airborne disease —using Severe Acute Respiratory Syndrome data as an example

Tai Ling Yin Winnie

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Nov 2010

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduced no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement had been made in the text.

Tai Ling Yin Winnie

Abstract

Introduction

Infectious diseases are alarming. They are contagious and detrimental as the illnesses concern the human respiratory systems which are the means that the infections are transmitted swiftly from one person to another within seconds. Nowadays, close contacts among people may result in chaotic situations attribute to the potential wide spread of infections when traveling abroad is becoming more common and frequent. The disastrous incident of Severe Acute Respiratory Syndrome ("SARS" thereafter) in 2003 caused more than eight thousands people worldwide infected with an initial of only ten infected people who once stayed in the same hotel, and those ten people were suspected to have shared the single infected origin. This case is the best illustration of how powerful a deadly infection can be. It awakened most countries to re-examine their existing public health delivery systems and to figure out what instant and remedial actions should be taken when infectious disease starts to spread, for instance, to determine the efforts needed for containment and mitigation. Historical incidents have already proved that prompt detections and comprehensive surveillances are vital to the eradication of infectious diseases, but the most appropriate way for effective and efficient screening has not yet been confirmed as each of them carries its own advantages and disadvantages. In order to ensure the isolation precaution starts at the right time for effective containment, a sensitive and reliable screening approach to trigger off the whole process is crucial.

Purpose

To explore the feasibility of using data mining as an effective screening

method to predict the occurrence of airborne disease based on the pre-hospitalized clinical presentations using the data of SARS in Hong Kong as an example.

Method

This study is an observational retrospective case record review study. All patients aged 18 or above, attending the Accident & Emergency Department (AED) of a major hospital during the period from 1st February 2003 to 30th June 2003 with provisional diagnosis of SARS, were recruited. Data collected for analysis included patient particulars, clinical presentations, co-morbidities, and laboratory results for confirmation of SARS based on the World Health Organisation (WHO) guidelines. There are four stages in this study. The first stage is the preparation of a comprehensive database for further analysis, followed by an evaluation of the existing prediction rules reported by others in stage two. The third stage is the attribute identification stage and the last stage is the model testing.

Results

A total of 549 adult case records were examined. Eighty percents of them were randomly selected to form the training dataset and the remaining cases were used as testing dataset. The testing data was fitted into the existing prediction model reported by Chen et al. (2004), Wang et al. (2004) and Leung et al. (2004) that all the studies were carried out in the most similar situation or inclusion criteria as the current study. The testing data was then classified into SARS and non-SARS based on each prediction rule and counterchecked with the laboratory diagnostic results. The sensitivity and specificity of each prediction rule were calculated and

compared with the quoted value. The poor agreement of the calculated sensitivity (ranged from 0.17 to 0.95) and the specificity (ranged from 0 to 0.67) with the quoted values showed a strong need to have a new prediction model with better prediction power. Data mining technique was employed to see if it can be an alternative prediction method for airborne disease. Association rule mining could not find any sequential /affinity relationship between the clinical variables and the disease status. Classification rule mining showed that malaise, sore throat, fever and shortness of breath were critical clinical predictors where clustering method identified chills, malaise, sore throat and shortness of breath as critical clinical predictors. The testing data was fitted into the mined rules again and another set of calculated sensitivity (0.86) and specificity (0.71) values were obtained for comparison. The results were further tested under different circumstances and similar findings were obtained.

Conclusion

Data mining can be a better and an efficient option with higher specificity and sensitivity for predicting airborne disease in AED in the future.

Publication arising from the thesis

Tai, W.L.Y., Lau, A.S.M., & Chung, J.W.Y. (2008). Triaging: Clinical presentation among SARS patients. Paper presented at *the 1st NUS-UH Conference: Advanced Practice Nursing in Multicultural Environments*, Singapore.

Tai, W.L.Y., Chung, J.W.Y., Lau, A,S,M., & Wong, T.K.S. (2009). Comparison of predicting score of clinical presentation among SARS patients at accident and emergency department. Poster presented at *ConTIC Saude 2009 – Congress on Technology and Humanization in Health Communication*, Ribeirao Preto, Sau Paulo, Brazil.

Acknowledgements

I wish to express my deepest gratitude to Prof Thomas Wong Kwok Shing and Prof Joanne Chung Wai Yee for their invaluable advice and guidance throughout the whole process of this study. I have deep appreciation for their kind support on solving problems that I encountered. I also wish to take this opportunity to extend my sincere thanks to Dr Adela Lau for her guidance in learning the data mining technique in my study.

Last but not least, I would like to thank all my family members, friends and colleagues for their support and encouragement throughout the study.

Table of contents

Title page……………………………………………………………….……...…..i
Certificate of originality……………………………………………….…….ii
Abstract…………………………………………………………………..iii
Publications arising from the thesis……………………….....………….vi
Acknowledgments…………………………………………………….vii
Table of contents………………………………...…………………viii
List of figures..…………………..……………...………………..….xv
List of tables..………………..…………………………….…..………xviii

CHAPTER 1 INTRODUCTION…………..…………………..………….…..1
1.1 Introduction ………………….…………………………………..…..………1
1.2 Aim of the study ………………..………….……..……………….…….…5
1.3 Research questions ………………….……..……….………………..………5
1.4 Hypotheses…………………..…….…………..…………..…………………..7
1.5 Definition of terms………………….….………..……….………….……..……8
1.6 De-limitation……………...…………….……..……..………11
1.7 Significance of the study………………..………………………….…. 12
1.8 Organization of thesis……………….……………………………...12

CHAPTER 2 CONTROL OF INFECTIOUS DISEASE……15

2.1 Introduction ……15

2.2 Impact of infectious diseases ……15

2.3 Progression of an infectious disease ……18

2.4 Control of infectious disease through detection ……23

2.5 Importance of detection and surveillance in controlling respiratory infectious disease ……28

2.6 Lessons learnt ……31

2.7 Future preparation ……37

CHAPTER 3 DIAGNOSING AND SCREENING FOR RESPIRATORY INFECTIOUS DISEASE ……39

3.1 Introduction……39

3.2 Methods for diagnosing respiratory infectious disease…… ……39

3.2.1 Detailed medical history taking……40

3.2.2 Comprehensive physical examinations…….. ……..41

3.2.3 Laboratory testing methods……42

3.2.3.1 Direct microscopy methods

3.2.3.2 Pathogen-specific macromolecules testing methods

3.2.3.3 Cultural testing methods

3.2.3.4 Serological testing methods

3.2.3.5 Factors to be considered in choosing different laboratory testing methods

3.3 Difference between diagnosing and screening diseases……..55

3.4 The need to screen for SARS at Emergency Departments……..59

3.5 Conclusions……..…62

CHAPTER 4 PREDICTION METHODS……65

4.1 Introduction**……**……**……..…...**65

4.2 Essential concepts for prediction in health care settings……..65

4.3 Common prediction methods used in health care settings……68

4.3.1 Dichotomous tests……**....**69

4.3.2 Odds-likelihood calculation based on Baye's Theorem…70

4.3.3 Presentation of results using Receiver Operating Characteristic (ROC) Curve ……..….71

4.3.4 Logistic regression……73

4.3.5 Bootstrap calculation……..**..**74

4.4 Use of data mining as an alternative……..75

4.5 Data mining techniques……79

4.5.1 Mining Association Rules in Large Databases……..80

4.5.2 Classifications and Prediction……**...**…85

4.5.3 Clustering techniques…….……90

4.5.4 Comparison of the three data mining techniques……92

4.6 Reviewing existing prediction rules……95

4.7 Conclusions……**.**100

CHAPTER 5 CONCEPTUAL FRAMEWORK……..103

5.1 Introduction……**.**103

5.2 Breaking the Chain of Infection…….…**.**103

5.2.1 Reservoir and susceptible host……104

5.2.2 Portal of exit and entry ……..……..…106

5.2.3 Appropriate intervention……107

5.3 Incorporating the use of data mining techniques into the

management of emerging infectious diseases……109
5.4 Conclusions……114

CHAPTER 6 STAGE 1 – DATA PREPROCESSING……..115
6.1 Introduction……115
6.2 Research objective and questions……..115
6.3 Methods……...116
6.3.1 Design……...116
6.3.2 Inclusion criteria……117
6.3.3 Ethical considerations……117
6.3.4 Procedures……118
6.3.5 Data management……118
6.3.6 Deliverables……119
6.4 Results……120
6.4.1 Demographic profile……121
6.4.2 Socio-economical profiles……122
6.4.3 Geographical profiles……123
6.4.4 Co-morbidity profiles……124
6.4.5 Laboratory result profiles……125
6.4.5.1 Renal function tests
6.4.5.2 Liver function tests
6.4.5.3 Complete blood profiles
6.4.5.4 Clotting profiles
6.4.5.5 Infection index
6.4.5.6 Other blood results
6.5 Discussion……141
6.5.1 Demographic profiles……142

6.5.2 Socio-economical profiles……..……144
6.5.3 Geographical profiles…….……146
6.5.4 Co-morbidity profiles……..…..147
6.5.5 Laboratory result profiles……148
6.6 Conclusion……..……153

CHAPTER 7 STAGE 2 - EVALUATION OF EXISTING PREDICTION RULES……154

7.1 Introduction……154
7.2 Research objective……155
7.3 Research questions…….……..155
7.4 Methods……..156
7.4.1 Procedures……...156
7.4.2 Data management……158
7.5 Results……...158
7.5.1 Profile difference between training and testing data……158
7.5.2 Data fitting……...…159
7.6 Discussion……162
7.6.1 Profile difference between training and testing data…..162
7.6.2 Data fitting……...…162
7.7 Conclusions……..166

CHAPTER 8 STAGE 3 - ATTRIBUTE IDENTIFICATION……168

8.1 Introduction……168
8.2 Research objectives……..……168
8.3 Research questions……...……169
8.4 Methods……..……169

8.4.1 Procedures……...…170
8.4.2 Data management……171
8.4.3 Deliverables…….……171
8.5 Results……172
8.5.1 General clinical predictors for adults…….....……..172
8.5.1.1 Use of association rule mining
8.5.1.2 Use of classification techniques
8.5.1.3 Use of clustering techniques
8.5.2 Different clinical predictors for different situations……189
8.5.2.1 Clinical predictors for geographical factors
8.5.2.1.1 Use of association rule mining
8.5.2.1.2 Use of classification
8.5.2.1.3 Use of clustering
8.5.2.2 Clinical predictors for different genders
8.5.2.2.1 Use of association rule mining
8.5.2.2.2 Use of classification
8.5.2.2.3 Use of clustering
8.5.2.3 Clinical predictors for young adults versus elderly patients
8.5.2.3.1 Use of association rule mining
8.5.2.3.2 Use of classification
8.5.2.3.3 Use of clustering
8.5.2.4 Clinical predictors for different co-morbidities
8.5.2.4.1 Use of association rule mining
8.5.2.4.2 Use of classification
8.5.2.4.3 Use of clustering
8.6 Discussion……..194

8.6.1 General clinical predictors for adults……194
8.6.2 Different clinical predictors for different situations……196
8.6.2.1 Clinical predictors for geographical factors
8.6.2.2 Clinical predictors for different gender
8.6.2.3 Clinical predictors for young adults versus elderly
8.6.2.4 Clinical predictors for different co-morbidities
8.6.3 Different data mining techniques……200
8.7 Conclusion……..……201

CHAPTER 9 STAGE 4 - MODEL FITTING AND TESTING ……..202
9.1 Introduction……202
9.2 Research objective…….……202
9.3 Research questions……..……..203
9.4 Method……203
9.4.1 Procedure……203
9.4.2 Data management…….204
9.5 Results…….204
9.5.1 General clinical predictors for adults ……204
*9.5.2 Differences in clinical prediction for different situations*204
9.6 Discussion……209
9.6.1 General clinical predictors for adults……209
9.6.2 Using of data mining technique…….….210
9.7 Conclusion…….…..214

References……..218

List of figures

Figure 1 Natural history of infectious disease timeline……..…….18

Figure 2 Distribution of test results (McGee, 2010)……67

Figure 3 A typical receiver operating characteristic (ROC) curve (Williams, Hand, & Tarnopolsky, 1982) ……..…72

Figure 4 Concepts of screening and diagnosis in relation to infectious diseases……...……...108

Figure 5 Overview of the steps in data mining process to develop a predictive model relative to patients who are suffering from SARS in Emergency Departments……110

Figure 6 A snowflake schema of the database design……..112

Figure 7 Theoretical framework……..113

Figure 7 Geographical distribution of the subjects diagnosed with SARS……...…..…..124

Figure 8 Distribution of selected renal function test results of disease groups (N=528) ……..….. 127

Figure 9a Distribution of selected renal function test results of disease groups for those without comorbidity (N=356) ……..128

Figure 9b Distribution of selected renal function test results of disease groups for those with comorbidity (N=172) ……..…… 128

Figure 10 Distribution of selected liver function test results of disease groups for all subjects (N=525) ……130

Figure10a Distribution of selected liver function test results of disease groups for those without comorbidity (N=354) ……..……131

Figure 10b Distribution of selected liver function test results of disease groups for those with comorbidity (N=171) ……..…131

Figure 11 Distribution of selected complete blood profile results of disease groups for all subjects (N=538) ……133
Figure 11a Distribution of selected complete blood profiles results of disease groups without comorbidity (N=362) ……133
Figure 11b Distribution of selected complete blood profiles results of disease groups with comorbidity (N=176) ……134
Figure 9 Distribution of selected clotting profile of disease groups for all patients (N=477) 135
Figure12a Distribution of selected clotting profile results of disease groups for those without comorbidity (N=301) ……136
Figure12b Distribution of selected clotting profile results of disease groups for those with comorbidity (N=176) ……136
Figure 103 Distribution of selected infection index of disease groups (N=549) ……137
Figure13a Distribution of selected infection index results of disease groups for those without comorbidity (N=377) ……138
Figure13b Distribution of selected clotting profile results of disease groups for those with comorbidity (N=172) ……138
Figure 14 Distribution of selected blood results by disease groups (N=549) ……140
Figure14a Distribution of selected blood test results of disease groups for those without comorbidity (N=377) ……140
Figure14b Distribution of selected blood test results of disease groups for those with comorbidity (N=172) ……141
Figure 115 Raw predictive scores of existing prediction rules (N=110) 160
Figure 126 Example of association rule mining.....177
Figure 17 Example of classification rule mining.....178

Figure 18 Example of the results mined by decision tree induction....180
Figure 19 Results showing the clinical predictors with different significances (number of clusters =3) 183
Figure 20 Results showing clinical predictors with different significance shown (number of clusters =8) 185
Figure 21 Results showing two-step clustering technique.....186
Figure 22 Data cube constructed for mining geographical difference predictions.....189
Figure 23 Data cube constructed for mining gender difference predictions.....191
Figure 24 Data cube constructed for mining age difference predictions.....192
Figure 13 Data cube constructed for mining comorbidity difference predictions.....193
Figure 26 Distribution of raw predictive scores of various prediction rules (N=110) ……205
Figure 26a Distribution of raw predictive scores of various prediction rules by disease group without comorbidity (N=73) ……206
Figure 26b Distribution of raw predictive scores of various prediction rules by disease group with comorbidity (N=37) ……206
Figure 27 Comparison of raw predictive scores of existing prediction rules (N=110) 207

List of tables

Table 1 Comparison of the three commonly used data mining techniques……93

Table 2 Summary of three SARS prediction studies……99

Table 3 Frequency distribution of smoking and drinking habits of the subjects (N=549) ……. 121

Table 4 Types of occupation distribution of the subjects (N=549)…123

Table 5 Profile difference between training and testing data (N=549)159

Table 6 Quoted and calculated values of sensitivity, specificity and predictive values for different prediction models for predicting the occurrence of SARS..... 161

Table 7 Summary of clinical presentation among different studies 176

Table 8 Summary of significant clinical predictors by different data mining techniques…… 187

Table 9 Percentage of correct classification of testing dataset under different situations..... 205

Table 10 Calculated values of sensitivity, specificity and predictive values ……..208

CHAPTER 1

INTRODUCTION

1.1 *Introduction*

Infectious diseases are alarming as they not only affect one's health but also affect a society financially and socially. The transmission of respiratory infection from one person to another occurs within seconds, and the number of sufferers increases immensely in a short period of time thereby. Easy and efficient modern transportation contributes to the spread of infectious diseases over great distances, leading to potential pandemics. The widespread of Severe Acute Respiratory Syndrome (SARS hereafter) in 2003 is an excellent example of modern pandemic. It demonstrated that an initial of a single tainted person infected ten people who stayed in the same hotel in Hong Kong, and finally more than eight thousands people were infected worldwide (Sampathkumar,

Temesgen, Smith, & Thompson, 2003).

Recent medical case is a 59-year-old woman suffering from influenza A (H5N1) activated the serious response level of the HKSAR government's Preparedness Plan for an influenza pandemic. The patient developed a runny nose, fever and cough in three days after visiting mainland China. She was not admitted to hospital at first when she sought medical advice in the Accident and Emergency Department of the Tuen Mun Hospital on 12 November 2010. Two days later, nevertheless, she was admitted to the hospital and diagnosed with pneumonia when she was found with continuous high fever as well as blood-stained sputum. Laboratory results revealed that she had contracted H5N1 influenza even though she did not visit any farms nor had any contact with live poultry during her stay in China. Due to the delay of intervention, her family members and five other patients staying in the same hospital cubicle were asked to receive further testing and were kept in quarantine at the Princess Margaret Hospital (Centre for Health Protection, 2010).

For patients presenting non-specific symptoms, even with prompt medical attentions, it is sometimes difficult for physicians to diagnose highly contagious diseases that require isolations. Due to the emergence of infectious diseases, like swine influenza and bird flu, all nations have been obliged to improve their approaches in public health to keep diseases under control right from the onset, that is, a concerted effort for containment and mitigation. Containment is an effective method to cut down the number of infected individuals and it requires state intervention along with cooperative individuals (Centers for Disease Control and Prevention, 2006c). The actual impact of a contagious disease, for example, a potential pandemic which may be caused by a highly pathogenic influenza virus, has to be evaluated using different criteria for minimizing the total amount of people being infected. Individual's awareness and state intervention play important roles. To ensure that the isolation procedures start at the right time for effective containment, a sensitive and reliable screening tool is essential for

triggering the process.

Although Hong Kong's health care service is among the top in the world's vital statistics (as evidenced by the low infant mortality rate, low prenatal mortality rate, low maternal mortality ratio and high life expectancy in Hong Kong) ,the question still remains on whether Hong Kong has a good protocol in place for minimizing the impact of an infectious disease.

In 2003, during the SARS outbreak, Hong Kong had the highest rate of infection among health care professionals and the second greatest number of deaths in the world (World Health Organization, 2004). How can we improve the implementation of a successful health care delivery system during the initial stage of an epidemic? Is the current health care system, the triaging process at the Emergency Department in particular, effective enough to differentiate and identify potential outbreaks of respiratory diseases in the future?

1.2 Aim of the study

This study aims to:

- Evaluate the existing predictive rules' effectiveness in predicting respiratory infectious diseases with the available SARS data.
- Identify the critical clinical variables of SARS patients by data mining based on the comprehensive database from available SARS data.
- Evaluate whether data mining can act as a screening tool based on their signs and symptoms to identify potential patients to contract SARS when attending Accident and Emergency Department (AED).

1.3 Research Questions

The research questions of this study include the following:

- Can the existing prediction rules developed by other researchers correctly forecast the occurrence of SARS with our data?

- What are the critical clinical variables of SARS by data mining based on the comprehensive database?
- Are there any associations among the clinical variables and the outcomes of SARS?
- What prediction rules can be identified from the SARS database?
- Can any of these prediction rules be highly effective and precise based on the clinical variables?
- Do the critical clinical variables show any gender difference?
- Are the critical clinical variables age-specific?
- Are there any differences between the critical clinical variables of the residents of the Amoy Garden housing estate (one of the identified estate that had a large group of SARS patients) and non-Amoy residents?
- Are there any variations found among the critical clinical variables between patients with or without morbidity history?
- Are there any differences of the predictive models in this study compared to others?

1.4 Hypotheses

The null hypotheses of this study are:

1. The existing prediction rules from other researchers can best predict the occurrence of SARS.
2. The critical clinical variables mined for males and females show no significant difference.
3. The critical clinical variables mined for different age groups indicate no significant difference.
4. The critical clinical variables mined for different geographic presentations illustrate no significant difference.
5. The critical clinical variables mined for patients with morbidities denote no significant difference when compared with those without chronic illnesses.

The alternate hypotheses are:

1. There are some other prediction rules than the existing ones

reported by other researchers with higher predictive value in the occurrence of SARS.

2. There is a significant difference in critical clinical variables mined for males and females.
3. The critical clinical variables mined for different age groups showed a significant difference.
4. The critical clinical variables mined for different geographic presentation illustrate a significant difference
5. A significant difference is noted from the critical clinical variables mined for patients with morbidity than those without chronic illness.

1.5 Definition of terms

Predictive values:

Theoretical use of empirical evidence to forecast how the clinical

variables of SARS will behave in new settings and with different individuals (Portney & Watkins, 2009).

Positive predictive values:

The values represent the percentage of patients with a positive test result who actually have the condition.

Negative predictive values:

The values represent the percentage of patients with a negative test result who do not have the condition.

SARS operational case definition (World Health Organization, 2003a)

(A) Clinical case definition of SARS:

A person with a history of fever ($\geqq 38^0$C) **and** one or more symptoms of lower respiratory tract illness (cough, difficulty breathing, shortness of breath) **and** radiographic evidence of lung

infiltrates consistent with pneumonia or respiratory distress syndrome (RDS) **or** autopsy findings consistent with the pathology of pneumonia or RDS without an identifiable cause **and** no alternative diagnosis to fully explain the illness.

(B) Laboratory case definition of SARS (World Health Organization, 2003b)

A person with symptoms and signs that are clinically suggestive of SARS **and** with positive findings for SARS-CoV based on one or more of the following diagnostic criteria:

(i) Polymerase chain reaction (PCR) positive for SARS-CoV

PCR positive using a validated method from at least two different clinical specimens (e.g. nasopharyngeal and stool) **or** the same clinical specimen collected on two or more occasions during the course of the illness (e.g. sequential nasopharyngeal aspirates) **or** two different assays or repeat PCR using a new RNA extract from the original clinical sample on each occasion of testing.

(ii) Seroconversion by enzyme linked immunoassays (ELISA) or indirect fluorescent antibodies (IFA)

Negative antibody test on acute serum followed by positive antibody test on convalescent phase serum tested in parallel **or** a fourfold or greater rise in antibody titres between the acute and convalescent phase sera tested in parallel.

(iii) Virus isolation

Isolation in cell culture of SARS-CoV from any specimen **and** PCR confirmation using any method.

1.6 De-limitation

This is a retrospective study based on reviewing case notes; some key points may have been omitted at the very beginning during the emergent stage of the disease. At the time, nobody knew which piece of information is important and when many patients occupied one hospital

with a high death rate, sometimes detailed documentation might not be secured. However, this kind of phenomenon is common when no previous data is available.

1.7 Significance of the study

Based on the lessons learnt from the SARS outbreak in 2003, this study introduces an efficient and reliable method to screen and identify infectious diseases.

1.8 Organization of thesis

This thesis is comprised of nine chapters. Chapter One reviews the background of the study. Chapter Two focuses on the importance of detection and surveillance on the eradication of historical infectious

diseases, followed by a comparison of the pros and cons of using different screening methods for identification of infectious diseases. Chapter Three examines the existing diagnostic and screening methods for respiratory infectious diseases and the incentives to employ correct screening tools in Emergency Departments. Chapter Four presents the essential concepts and different kinds of prediction methods commonly used in health care environments, followed by introducing data mining as an alternative method for prediction and a brief overview of the three most commonly used data mining techniques. A literature review on the existing prediction rules is included in this chapter. Chapter Five presents the conceptual framework of the study. Chapter Six is the details of the Stage One of this study including the rationale, method, design, procedure, and results on how to construct a comprehensive database from the data available in this study and displays the compute descriptive statistics of the data obtained. Chapter Seven describes the Second Stage of the study. It evaluates the existing prediction rules using the constructed database from Stage One. Chapter Eight focuses on the

construction of a new prediction rule using the data mining method and compares the sensitivity and specificity of the new prediction rules with the existing ones. Chapter Nine compares the effectiveness of the newly identified prediction rules with those from others and a conclusion is given.

CHAPTER 2

CONTROL OF INFECTIOUS DISEASE

2.1 Introduction

This chapter describes the progression of an infectious disease, the impact of the infectious disease, and the most appropriate time for intervention to achieve optimal results. The historical importance of detection and surveillance to control infectious diseases is outlined. The reasons for the widespread of swine flu all over the world are provided later in this chapter, in spite of discussions on the years of experience in handling infectious diseases.

2.2 Impact of infectious diseases

An infectious disease is an illness clinically evidenced by the presence of pathogenic microbial agents, which cause diseases in animals and/or plants. Infectious pathologies are also called communicable diseases or transmissible diseases due to their potential transmission from one person to another by a replicating agent (Hornby, 2005). There are many different kinds of human diseases that involve different causative agents such as bacteria, viruses, protozoa and fungi.

The term "infectious disease" was first mentioned in 1940 (Burnet & White, 1972). Archaeologists found malaria antigens in skin and lung samples of Egyptian mummies with enlarged spleens in 3204 BC (Khan, 2008). Though we have made significant advances in medical sciences in the past 5,000 years, malaria is still killing more than a million of people every year worldwide (World Health Organization, 2010b), while HIV or AIDS is causing nearly 2.7 million of deaths annually and tuberculosis contributes to 1.7 million of deaths. It can be seen that pathogens are continuously developing means or strategies to resist the treatment

available at our disposal. The process of drug resistance takes place at a much faster rate than our ability to produce new drugs or new intervention to halt the disease process. Hence, the ultimate goal of combating an infectious disease is to stop it from spreading.

Infectious diseases not only attack the developing countries and cause enormous number of deaths, but also affect developed countries by destroying the entire community and causes devastating loss to economy. Annually, there are approximately 14 million deaths caused by different kinds of infectious diseases worldwide (World Health Organization, 2010b). Apart from the fatal consequences, hundreds of millions of people are left disabled or orphaned because of the spread of infectious diseases (World Health Organization, 2010a). Case numbers are a poor indication of the burden of an infectious disease. Infectious diseases such as poliomyelitis have a low mortality rate but inflict heavy loss of healthy years of life.

There is no official data on the global prevalence rate of infectious diseases owing to the various methods used in different countries in obtaining the numbers. In the section of "Selected infectious diseases" in the *Global Health Indicators* (World Health Organization, 2010b), we can find that majority of the infectious diseases are transmitted either by air-borne or by droplet infection. Therefore, focusing on combating infectious diseases related to air-borne or droplet infection is the right approach.

2.3 Progression of an infectious disease

One has to get familiar with the classic epidemiologic triad of host, agent and environment model of disease causation in order to understand better the emergence of an infectious disease. In this model, it shows that the development of disease depends on factors that determine the probability of contact between an infectious agent and a susceptible host in an

environment that supports transmission of the agent from a source to the host. The spread of an infectious disease follows the chain of infection of three components. Transmission of disease occurs when the agent leaves its reservoir or host through a portal of exit conveyed by some mode of transmission and a portal of entry to infect a susceptible host. This relationship among these three factors can be illustrated by the stage of susceptibility in Figure 1.

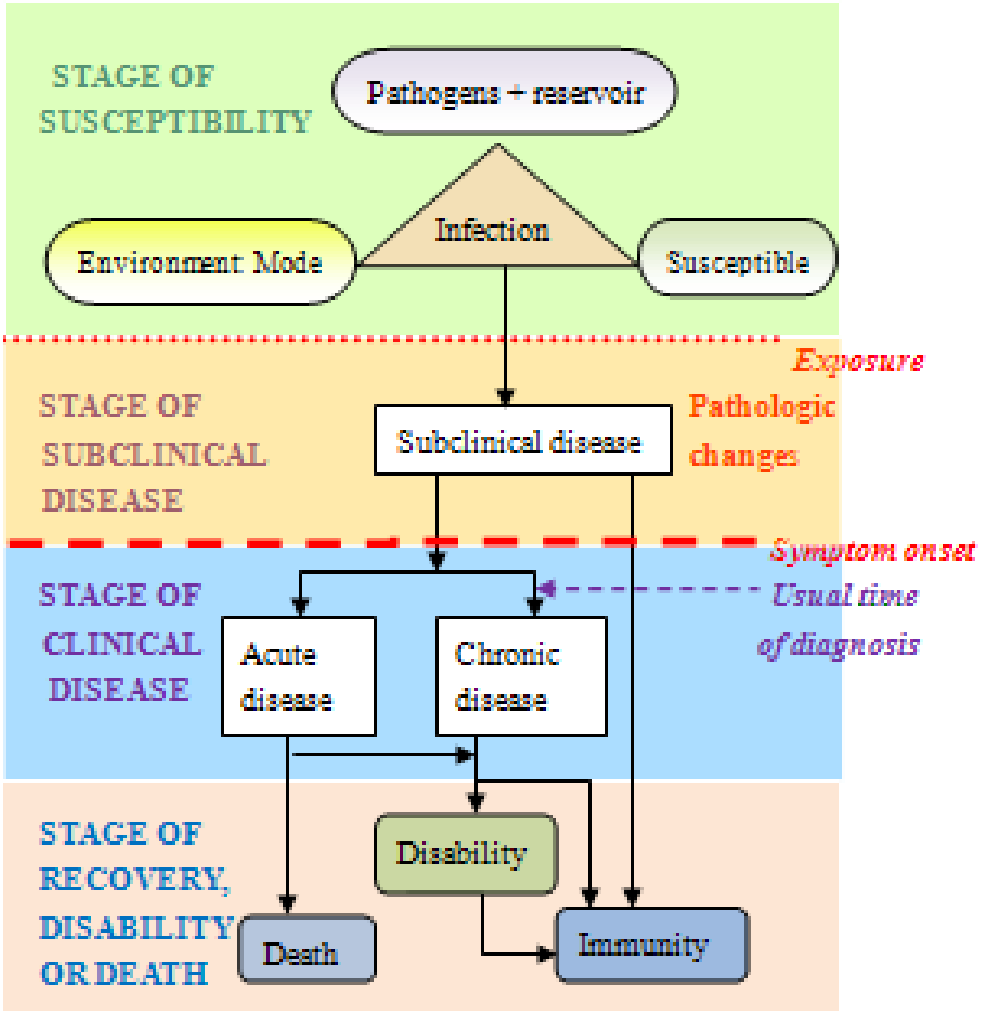

Figure 1 Natural history of infectious disease timeline

(Centers for Disease Control and Prevention, 2006c)

The natural progression of an infectious disease process begins with the exposure to or accumulation of factors sufficient for the disease process to begin in a susceptible host; for infectious disease, the exposure is a microorganism. Most diseases have their own natural histories, their

own characteristics of progression, time frame and specific manifestations varying from individual to individual and can be influenced by treatments or preventive measures (Centers for Disease Control and Prevention, 2006b).

After triggering the disease process, the pathological changes can occur without the host being aware of; this is known as the subclinical disease stage (see Figure 1), and may also be termed the incubation period for infectious diseases (or latency period for chronic diseases) and it extends from the time of exposure to the onset of disease symptoms. During this stage, the disease is asymptomatic but some pathologic changes may be detectable with laboratory, radiographic or other screening methods. Most screening programs try to identify the disease process during this phase of its natural history. It is believed that intervention at this early stage is likely to be comparatively more effective than having the medical treatment after the disease has progressed and become symptomatic.

The onset of symptoms marks the transition from the subclinical disease stage to the clinical disease stage. During this stage, the majority of diseases can be diagnosed. Nevertheless, the possible outcomes include recovery with immunity, or disability or even death depending on the individual's response.

As the infectivity, pathogenicity and virulence of different causative agents vary, not all diseases are/can be diagnosed by clinicians. Many cases may go undetected and never progress to the clinical stage so it allows the carriers spread unsuspectingly.

Different balances and interactions of the three components: host, agent/pathogen and environment cause different diseases. Therefore, appropriate, practical and effective public measures are necessary to control and prevent the spread of the diseases and should be taken according to the assessment of these three components and their

interactions (Centers for Disease Control and Prevention, 2006b). There are plenty examples of massive and deadly pandemics resulting from ineffective public measures and these are to be discussed in the following sections.

2.4 Control of infectious disease through detection

The bubonic plague (also known as the "Black Death") as exploded in Europe in 1350 was caused by the Gram-negative bacterium called *Yersinia pestis* (Achtman et al., 2004). The plague spread so rapidly that before any physicians or government authorities had time to look for its origin, it had already killed hundreds of millions of people. Scholars nowadays postulated that the plague spread through flea-infected black rats on the merchant ships, which sailed between the trade routes connecting Asia and Europe. The conditions of war, poor sanity, drought, and malnutrition are also considered to be the contributing factors (Arita

et al., 2004; Heymann, 2005; Kutyrev, 2008). It affected humans drastically and it killed approximately half of Europe's population at that time (Morelli et al., 2010).

The eruption of the plague irrevocably changed the social structure and the emotional state of the people in that period, the uncertainty of daily survival posed a drastic effect on the general mood of morbidity and influencing people to "live for the moment" (Halsall, 1996). The plague returned from time to time killing more and more people each time until it left Europe in the 19th century. The spread of the plague finally died down when survivors became immune and more people knew how to protect themselves against it. Of course, the development of antibiotics in 1890 played a vital role in controlling the disease.

The cholera outbreak in the Golden Square of London in 1850 resulted in more than 600 deaths within a fortnight. The disease came to an end when John Snow investigated the cause of it. He believed that

contaminated water was the source of the infection, and it was a ludicrous idea at the time and most people ~~who~~ considered him a fool. He finally proved that the water pump was the source of the epidemic and took the significant preventive measure by removing the pump handle~~. And~~ , thus breaking the chains of the infectious disease (Berkelman, Byran, Osterholm, LeDuc, & Hughes, 1994; Centers for Disease Control and Prevention, 2006c; Colwell, 1996; Tablan et al., 2004).

Rainer (2004) briefly summarized the clinical progress of SARS into 3 phases. Phase 1 included initial presentation of signs and symptoms which was the viral replication phase. It usually resolves with antiviral and immunomodulatory therapy. Patients in this stage can either fully recover or will migrate to Phase 2 with unclear mechanism. Phase 2 begins about 8 days after initial onset of signs and symptoms, including immune mediated lung injury and/or responds to steroids and oxygen therapy in most cases. More than 20% of the patients in Phase 2

deteriorate physically. They may then need intensive care with a high mortality rate in Phase 3. Hence, if we can detect or identify patients promptly from their presenting symptoms and start necessary intervention (such as antiviral and immunemodulatory therapy), the number of patients proceeding to Phase 3 will certainly decrease and hence reduce the mortality rate.

A more recent infectious disease outbreak was caused by *Escherichia Coli O157:H7* in September 2006. The bacteria in the organic bagged fresh spinach was discovered in 28 states in the USA causing sickness in more than 200 people with 3 deaths within weeks (International Food Safety Network, 2006; Surak, 2007). Since the spinach was also distributed to Canada and Mexico, several cases were also reported in these countries. Subsequent outbreaks in the same year in the USA were all related to contaminated vegetables imported from the same supplier. Information gathered from two surveillance systems, PulseNet (part of the Association of Public Health Laboratories, coordinated with Centres

for Disease Control and Prevention (CDC)) and OutbreakNet (state public health officers who investigate food borne infection outbreaks), collaborated to find the cause of the outbreak. PulseNet detected clusters of outbreaks in Oregon and Wisconsin and shared this information with OutbreakNet drawing a conclusion that the fresh spinach was infected (Centers for Disease Control and Prevention, 2006d).

All these examples show us that if early detection of the pathology was available and we had a better knowledge of how the disease was transmitted to one another, the effect and the mortality rate caused by the disease could have certainly decreased and the impact caused by the disease would have reduced to the least. Similarly, if the scientists and microbiologists managed to have an earlier detection of an infectious disease or SARS and knew how the disease spread promptly, the outcome and the impact would have been totally different.

2.5 Importance of detection and surveillance in controlling respiratory infectious disease

Epidemiologic investigation indicates that prompt detection and appropriate intervention can help restraining the progressions of diseases. It is particularly true if the attack of infectious disease concerns human respiratory system. We can easily recall that there are several unforgettable pandemics related to respiratory infectious diseases (Hilleman, 2002). Flu spreads around the world in seasonal epidemics. Three influenza pandemics, each caused by a new strain of virus in human, occurred in the 20th century and killed millions of people. The most significant and serious one happened in 1918, which recorded a clinical attack rate exceeding 40%. Around 40 million deaths were reported during this Spanish flu attack (The University of Hong Kong, 2005). The symptoms found in 1918 were very unusual that influenza was initially misdiagnosed as dengue, cholera, or typhoid. Most victims were healthy young adults contrastive to the traditional influenza

outbreaks which attack infants, the elderly or patients with poor immunity (Centers for Disease Control and Prevention, 2006a). Scholars believe that, this influenza pandemic was not directly related to World War I (1914-1918) but rather the close troop quarters and massive troop movements hastened the pandemic and probably increased the chances of transmission and augmented mutation. The pandemic almost disappeared in November 1918. Different schools of thought emerged to explain the reasons for such rapid decline. One theory claimed that the doctors simply knew how to better treat and prevent the dreadful complication of the influenza, that is, pneumonia; another theory stated that the 1918 virus mutated extremely fast to a less lethal strain which was comparatively common in influenza viruses; and some considered that the virus become less lethal since more and more people survived from the pandemic and became immune (Taubnberger & Morens, 2006).

The second influential case was the Asian flu outbreak in 1957, which originated from the mutation of the influenza virus A H2N2 subtype in

wild ducks in China. In this incident, the death toll in the United States alone was approximately 70,000. There were two million people died worldwide, particularly the elderly who were the most vulnerable (Centers for Disease Control and Prevention, 2006a). This Asian flu was considered mild compared to the Spanish flu of 1918, and it attributed to the advanced scientific technology that the pathogen could be identified quickly and vaccines were promptly produced 3 months after the outbreak (Henderson, Courtney, Inglesby, Toner & Nuzzo, 2009).

The third case was the "Hong Kong flu" originating from the influenza virus A H3N2 subtype which was first detected in Hong Kong in early 1968 and spread to the USA (causing about 34,000 deaths) and it was estimated to have affected one million people worldwide (Centers for Disease Control and Prevention, 2006a). This pandemic was considered to be the mildest pandemic of the 20th century. The reason behind this could be that the virus itself was a variant strain of the 1957 Asian flu virus H2N2, mutated via antigenic shift. Moreover, this pandemic had its

peak during school holidays which greatly reduced the infectious rate. This strain of virus keeps on mutating and is now considered as seasonal influenza and kills more than 30,000 people in the USA each year (U. S. Department of Health & Human Service, 2010).

2.6 Lessons learnt

Scientists and epidemiologists (Taubnberger & Morens, 2006; Oshitani, Kamigaki & Suzuki 2008; Moreno, Rhodes & Chiche, 2009) summarize some of the characteristics and challenges of a flu pandemic from the past experience. First, when a pandemic emerges, it has the potential to spread rapidly worldwide so it has to be assumed that the entire world population is susceptible to infection. Second, most people in the world will have little or even no immunity to this emerging virus which is a threat that a substantial percentage of the world's population may require some forms of medical care when the pandemic emerges. Health care

systems will then be under tremendous strain when copping with this sudden increase of demands in the availability of medical staff, facilities, equipment and hospital beds to cater for a large number of patients. Supplying antiviral drugs and vaccines will cause financial pressure to the government as well as social unrest when decisions have to be made in prioritization of treatment amongst different patient groups. Third, social and economic disruption is unavoidable as people need to take care of sick family members and they fear that significant worker absenteeism will be resulted. All the traveling among countries will diminish and the closing of schools and businesses, and cancellations of events would further impact various communities (U. S. Department of Health & Human Service, 2010).

Even after so many large pandemic attacks, the health care professionals especially the epidemiologists cannot prevent them from happening again. Epidemics have kept returning back in reality. The SARS outbreak provided an opportunity for the public health professionals worldwide to

realize that public health systems were outdated and in lack of unified command, and management systems ~~that~~ failed to cope with the crisis. The existing health care system was flawed not only in the rapid mobilization of resources and manpower during the emergency but also in the sensitive to detect any clues in finding the major cause of transmission. As all countries are concerned about future influenza pandemics, better approaches are urgently needed to improve the current systems for enabling frontline health care workers to trigger off the system alarm. This is the reason that we need to revisit the course of SARS again to see how the screening and implementation of preventative measures for infectious diseases can be improved.

Symbolized by the image of a masked face, SARS was first recognized in March 2003 in Hong Kong but probably had its origins in the Guangdong Province, China emerging earlier in November 2002. Between March and July 2003, over 8000 probable cases of SARS were reported from all around the world including North America, Europe and

other Asian countries. The global outbreak can be traced to a man who spent a night in a hotel on 21 February 2003 in Hong Kong. Scientists are still baffled about how Dr Liu, a 64-year-old physician from China's Guangdong province, could have transferred SARS to at least 16 other guests on the same floor during his brief stay. SARS travelled more widely, swiftly, and lethally than any other recent new disease. A novel coronavirus, known as the 'SARS coronavirus' (SARS CoV), was identified as the cause of SARS and preliminary animal studies isolated the SARS CoV in wild animals native to Guangdong and other parts of China (Peiris, Lai et al., 2003).

Near the end of July 2003, the total cases reported were 8096 in 30 countries, 774 of which resulted in death (Centers for Disease Control and Prevention, 2003b). There were 349 deaths from 5327 cases of SARS in China. Clusters of patients with pneumonia were noted in Hong Kong, 299 deaths arose from 1755 cases. The most heartbreaking fact was that one fifth of the infected cases concerned the frontline health

care workers. SARS struck fear to the public across the globe triggering drastic measure, mass quarantine in hospital wards enforced by armed guards, infectious passengers removed from planes, and businesses and schools closed. As the epidemic grew, China threatened to execute any SARS patient who violated quarantine. The power of this disease was so potent that it rocked the stability of a society economically and socially. Effective prevention measures such as vaccines were not available for controlling the spread of the infection. The most frustrating fact was that the disease did not respond to empirical antimicrobial treatment for acute community-acquired, typical or atypical, pneumonia.

In March 2009, another influenza-like illness pandemic occurred. The cause was unknown at the very beginning. The disease spread rapidly and more than 3,700 confirmed cases were reported worldwide within one month. The World Health Organization (WHO) declared the first influenza pandemic of the 21st century on 11 June 2009. It was commonly referred to as “swine flu” as it was mutated from four known

strains of influenza A virus subtype H1N1 with one endemic found in humans, one found in birds and two found in pigs. The pandemic was officially adjourned on 10 August 2010 by the WHO Director as out-of-season outbreaks were no longer being reported in either the northern or the southern hemispheres (Harti, 10 August 2010). According to the latest statistics from the WHO, the pandemic killed more than 18,000 people worldwide which accounts for approximately 4% of the annual influenza deaths (Trifonov, Khiabanian, & Rabadan, 2009; World Health Organization, 2009). The WHO was accused of being too cautious in its early pandemic announcement as the death rate was only 0.04% which is very low compared to other pandemics (World Health Organization, 2009). However, there were also voices supporting the WHO as the early announcement that could be seen to curb the progression of the disease, and a later announcement could have resulted in a higher infection rate leading to unthinkable situations.

2.7 Future preparation

The relatively short duration and the low death rate of the recent H1N1 (swine) flu pandemic is the evidence of effective and prompt solutions found to break the chain of infection. Containment strategies like isolation of index cases and treatment, tracing of contacts, closure of schools and avoiding large-scale community activities involving hundreds or even thousands of people, certainly help tremendously in preventing the disease propagating. Constant surveillance as well as contact tracing are all vital steps that must be done whenever an infectious disease is suspected.

Infectious disease undergoes constant changes to adapt to the ever-changing extrinsic factors. The occurrence of any infection rises and falls with changes in the immunity of the host population and through changes in the virulence of the pathogens. Endemic infections always present in a community and the number of infected cases varies over

time and sometimes it results in superimposed conditions which are considered to be normal as well. This situation is true for influenza infections that there are always people contracting the disease over the entire year with more cases over specific periods for seasonal influenza (March and December). Bearing this in mind, we need to know that the goal is to develop a good screening tool when a specific condition is likely to be the start of a pandemic.

CHAPTER 3

DIAGNOSING AND SCREENING FOR RESPIRATORY INFECTIOUS DISEASE

3.1 Introduction

This chapter firstly focuses on different types of diagnostic and screening methods which are being currently used to identify infectious diseases. What follows is a review of their applicability in predicting or identifying an emergent disease in terms of effectiveness and efficiency.

3.2 Methods for diagnosing respiratory infectious disease

Several methods can be used to diagnose or identify infectious diseases. They are usually initiated by checking the detailed medical history as

well as comprehensive physical examinations. When a clinician makes a decision for a diagnosis, it is likely to be derived from a physical examination (Bachmann, Kolk, Koller, Steurer, & ter Riet, 2003; Backmann, Haberzeth, Steurer, & ter Riet, 2004), patient history (Buchsbaum, Buchanan, Centor, Schnoll, & Lawton, 1991), or a combination of both (Hawker, Jamal, Ridout, & Chase, 2002; Wells et al., 1997), or a combination of patient history, physical examination and patient beliefs (Childs et al., 2004; Flynn et al., 2002). In addition, the identification of the infectious agents isolated from a patient's specimen, microscopy, culture and serology testing are undertaken (Liao et al., 2007; Saijo et al., 2005; Yu et al., 2005).

3.2.1 *Detailed medical history taking*

Similar to many other diseases, correct screening or diagnosing methods for an infectious disease depends on detailed medical health records, epidemiological data, together with the client's traveling history, occupation, and contact history. These are the primary data for managing

a patient's condition and the spread of the disease. Infectious diseases vary in different conditions (for example geographically, seasonally, and epidemically), therefore, it is sometimes very difficult to confirm the diagnosis based on only clinical manifestations such as diarrhoea or respiratory symptoms.

3.2.2 *Comprehensive physical examinations*

Other than medical health history and physical examination, it is now generally agreed that Koch's postulate, proposed by Robert Koch in 1890 (Murray, Rosenthal, & Pfaller, 2005), also needs to be satisfied to prove a given disease "infectious". The infective agent must be identified only in patients who develop the disease but not in healthy patients who act as controls. The microorganism must be isolated from a diseased organism and be grown in pure culture in which the microorganism can be re-isolated from the inoculated, diseased experimental host and identified as being identical to the origin specific causative agent.

Several methods are commonly used in the laboratory for diagnosing infectious diseases namely direct microscopy (sometimes followed by staining), cultural techniques and immunological methods.

3.2.3 *Laboratory testing methods*

Other than medical health history and physical examination, it is now generally agreed that Koch's postulate, proposed by Robert Koch in 1890 (Murray, Rosenthal, & Pfaller, 2005), also needs to be satisfied to prove a given disease "infectious". The infective agent must be identified only in patients who develop the disease but not in healthy patients who act as controls. The microorganism must be isolated from a diseased organism and be grown in pure culture in which the microorganism can be re-isolated from the inoculated, diseased experimental host and identified as being identical to the origin specific causative agent.

Several methods are commonly used in the laboratory for diagnosing infectious diseases namely direct microscopy (sometimes followed by

staining), cultural techniques and immunological methods.

3.2.3.1 Direct microscopy methods

This method is simple and direct, and provides useful clinical information within 4 hours after receiving the specimen. The different microscopy methods include bright/dark-field microscopy, phase-contrast microscopy, fluorescence microscopy, and electron microscopy (Elliott, Worthington, Osman, & Gill, 2007; Virella, 1997).

3.2.3.1.1 Direct examination using microscope

Direct microscopy is usually used to detect parasites by visualization or by assessing the type and number of inflammatory cells. Various kinds of microscopy techniques can be used to identify the infective agent.

In bright-field (light) microscopy, light passes through the specimen directly and magnifies the image without using specific condensers. It is the most common way of identifying microbes but its resolution is not high enough to visualize viruses and the specimen must be stained to overcome the similar refractive indices of the organisms and background.

Dark-field microscopy employs a condenser to prevent transmitted light from directly illuminating the specimen and only oblique and scattered light reaches the specimen and passes through the lens system. This greatly improves the resolving power and makes it possible for extremely thin microbes to be detected. However, the internal structure of the organisms cannot be visualized clearly as the light source is not strong enough. This method needs to be performed by experienced personnel. However, they

may not be able to run numerous tests during the spread of pandemic when time is limited.

In phase-contrast microscopy, parallel beams of light pass through a transparent specimen. The light waves are shifted in phase by different compositions in the specimen and by manipulating the annular rings in the condenser a three-dimensional image is created and the internal details of microbes can be analyzed. The specimen does not require staining. It is possible to study the cell cycle but experienced personnel need to perform the tests again, thus difficulties of having enough experienced laboratory personnel could arise during a pandemic.

Electron microscopy makes use of the magnetic coils instead of lenses to direct beams of electrons from a tungsten filament through a specimen and onto a screen.

The advantage is that the image obtained can be recorded on a photographic plate or computer screen and it can be used to observe a wide range of specimens, not only cells. The drawback is that it is very expensive to build and maintain the standard of the laboratory, and it is very sensitive to vibrations and external magnetic fields. Hence, it needs to be housed underground with magnetic field cancellation systems to compensate for beam fluctuations.

3.2.3.1.2 Direct microscopy with staining

Fluorescence microscopy involves firstly, staining of the organism with a fluorescent dye and then viewing it using a fluorescence microscope in which the emitted light is detected and forms the image. This method is relatively simple but the price of the reagents and the requirement of experienced personnel are the limiting factors for an

extensive use during a pandemic.

Gram staining is the most common method being used routinely to run identification test in diagnostic microbiology and it allows rapid classification of bacteria into five simple categories: Gram-positive cocci, Gram-negative cocci, Gram-positive bacilli, Gram-negative bacilli or Fungi. It also helps to identify the presence of inflammatory cells and to delineate confusion caused by commensals.

Acid-fast staining or Ziehl-Neelsen staining are usually used to identify different infections such as mycobacterial infections.

3.2.3.2 Pathogen-specific macromolecules testing methods

This method concentrates mainly on detecting bacterial antigens and nucleic acid sequences. For the detection of bacterial antigens, it primarily utilizes the specific relationship properties of antigens and antibodies. It involves the use of enzyme immunoassays (EIA), complement fixations, hemagglutination inhibitions or particle agglutinations. The antigens bind to the specific antibodies which are immobilized in a solid phase such as a microtiter plate. This method can be extremely simple and rapid to perform.

For the detection of nucleic acid sequences, the most common methods are the nucleic acid probe test and the polymerase chain reaction (PCR). The nucleic acid probe test involves using a probe to detect and measure chemiluminescence emitted by the

hybrid formation of complementary deoxyribonucleic acid (cDNA) from a specific ribosomal gene sequence. This method is especially expensive due to the prices of the reagents and the instrument.

The PCR method involves the use of two short DNA primers specifically for the flanking regions of the DNA segment one wishes to identify with. Repeated heating and cooling together with the *Thermus aquaticus* (Taq) polymerase form the hybridization of millions of copies of the chosen section of targeted DNA. Thus, it is a very powerful and sensitive test to confirm whether a microorganism is in a patient's spectrum.

Environmental contamination in the specimen leads to false-positive results which is a major concern when deciding on the choice of laboratory test. Moreover, the high cost of the equipment and reagents and the need for technical expertise are

the major concerns during diagnosing processes.

3.2.3.3 Cultural testing methods

Culture is still considered as the gold standard and the most reliable method to confirm diagnosis and to optimize treatment by the performance of antibiotic susceptibility tests. Different combinations of culture media such as sheep blood agar, chocolate agar, MacConkey agar and different culture conditions such as aerobic or anaerobic, are utilized for different specimens to determine the causative agents. Bacteria can often be identified presumptively by their growth of characteristics on different media and the results of biochemical tests.

Viral culture, however, requires more time to replicate in order to be detectable as viruses must be cultured on living cells. The culture can be primary freshly isolated cells which must be used

within 2-3 weeks, or diploid cultures which are prepared with cells that can be sub-cultured twenty to thirty times, or heteroploid cultures which are prepared from immortalized cells. The method of detecting the viral growth depends on the observation of cytopathogenic effects or indirectly through the technique of hemadsorption. Additional tests such as morphological examination, neutralization, hemadsorption inhibition, cytopathogenic effects interference inhibition, and immuno-electron microscopy are required to identify the virus growth precisely. These steps potentially increase the laboratory personnel's risk of infection which is also a critical concern to the whole identification process. Furthermore, even though the skills of extracting specimens have improved, the timing of specimen collection in relation to the onset of disease is still particularly important to assess the presence of viruses.

3.2.3.4 Serological testing methods

There are many different serological tests. The classic approaches include complement fixation tests, flocculation tests, indirect immunofluorescence tests, and immunodiffusion assays. They involve antibody-antigen complexes that can be detected directly, by precipitation techniques, or by labeling the antibody with a radioactive, fluorescent, or enzyme tag. It is especially useful when the agent is difficult to isolate as in viral infections, mycoplasma infection, or when the infection is deep-seated such as streptococcal osteomyelitis or to document prior infections such as rheumatic fever and glomerulonephritis. However, serological tests have two major limitations. The antibody titers may be undetectable during the early evolution phase of the disease which nullifies the possibility of an early diagnosis. Even though the serologic assay has a positive result that is a raised antibody titer, it is often difficult to decide how that titer

correlates with the patient's clinical condition.

3.2.3.5 Factors to be considered in choosing different laboratory testing methods

Many bacteria do not survive well outside the body, for example, obligate anaerobes may be killed by atmospheric oxygen. Some bacteria are very humidity sensitive that they dry up even during specimen transportation. Below are the factors affecting the choice of laboratory testing methods:

i. Time of collection of specimen – it is highly related to the progression of disease in patients. There are more organisms present soon after the onset of symptoms (i.e., during the acute phase of the disease), hence culture techniques should be considered. In the later stages, fewer organisms are present and hence serological testing is preferred for identifying the etiology of the disease.

ii. Source of specimen collection -- for highest likelihood of positive yield- different diseases affect specific parts of our body and hence the sites for specimen collection vary for different diseases, for example, taking throat swabs for pharyngitis and taking sputum, bronchoalveolar lavage fluids, or pleural fluids for pneumonia.

iii. Specimens collection – ‑ it should be collected before antimicrobial agents are used. It is very common for a general practitioner to prescribe antibiotics to patients when they first see them before having any laboratory testing. Hence, should the patient's condition worsen and on admittance to hospital for further treatment, specimen taking at that moment might provide false-negative results.

- Specimen transportation – it is recommended to be transported in appropriate transport medium. Some microorganisms should be inoculated onto media soon after their collection as they are extremely environment

sensitive, so different transport media should be considered if the emergent disease is unknown to microbiologists.

3.3 Difference between diagnosing and screening diseases

Diagnosing a disease always refers to the act or process of identifying or determining the nature and cause of a disease or injury, through evaluation of patient history, and the examination and reviewing of laboratory data (Hornby, 2005). Diagnosis is absolutely useful in confirming the causative agent and the correct use of pharmacological intervention. Diagnostic tests are usually invasive and carry a risk of infection and injury during the procedure.

The screening of a disease, on the other hand, is the medical investigation which does not arise from a patient's request for advice or

for any specific complaints (Wilson, 1971). It is only a systematic attempt to estimate whether certain groups of people have a higher risk of a specific disease among apparently healthy individuals (Cuckle, 2004). According to Wilson (1971), "screening" differs from the usual form of clinical diagnosis in the sense that "the doctor is offering a benefit either to the general population or to some particular group, in contrast to patients with symptoms."(p. 1255). Hence, the emphasis of screening is to specifically justify which screening test can create benefits or bring drawbacks in terms of both resources and ethical issues.

Some guidelines have been developed in the United Kingdom to ensure that physicians are cautious about the consequences and implications when they consider to adopt any screening tests. They must meet the strict criteria after rough estimations of the rating of severity and that of prevalence (Gilbert, Logan, Moyer, & Elliot, 2001). Laboratory results are not perfect even though various methods can be used to identify symptomatic patients who contract infectious diseases. We can thus

conclude that each method has its own limitations along with advantages and thus it is sometimes quite difficult to decide which test to use, particularly during the commencement of a pandemic.

The cost of screening outweighs the consequences and effectiveness of follow-up treatments but the natural history of conditions justify the act of screening. The situation can become catastrophic if many asymptomatic patients are involved during the early commencement stage of infectious diseases as one may not be able to identify the differences unless a large-scale screening campaign is carried out. At this stage, when all information and data seem relevant, the inability to distinguish whether the causative agent is aerobic, anaerobic bacteria or even a virus, hinders one's decision to be made on which test to use.

In real situations, relevant and irrelevant tests are both included to avoid missing any true positive results. These approaches increase the chances of infection for the laboratory personnel as numerous specimens are sent

to the laboratory for testing at the same time. The bulk testing processes eventually impose economic and resource burdens on the health care system. Despite widespread checking, there is no guarantee of complete accuracy, the test results obtained may still mistakenly identify some healthy people as diseased (false-positive results) and some affected people as disease-free (false-negative results) or even in some cases that the laboratory results may lead to wrong clinical decisions. To lower the chances of misdiagnosis, health care professionals, especially the epidemiologists, should incorporate prediction techniques based on statistical calculations to forecast the likelihood of the occurrence of a disease.

As Cuckle (2004) stated, all screening programmes must contain three components: test, disease and preventive action. One simple positive test result with no grey area can define a diseased population from healthy individuals. If the next step is the diagnostic procedure, then there should not have any error in diagnosing. If the next step is

treatment, the ability to prevent the disease or its adverse consequences will have been established. However, all of these outcomes greatly rely on the discriminatory power and predictive value of the screening test. That is the ability to separate diseased patients from healthy individuals and to an extent which high-risk groups can be distinguished from low-risk groups.

3.4 The need to screen for SARS at Emergency Departments

As the pandemic of Avian flu and the new form of Influenza A emerge, more and more articles using SARS as an example to discuss the preparedness and risk assessment on this area (Lim, Ng, & Tsang, 2006; Lim, 2006; Leppin & Aro, 2009) have been published. However, there is very little focus ~~is~~ on the clinical manifestation of the disease, especially at the very early stage when patients are first admitted to Accident and

Emergency Department (AED) during triage when no serological or pathological results are available.

Among 6000 literatures, one hundred articles reviewing the situation of Emergency Department during an epidemic were found from Medline search engine. They covered wide range of topics such as the economical and psychological impact on department and staff of the abrupt increase of attendance rate, changing of disease definition (Centers for Disease Control and, 2003a; Hoey, 2003; World Health Organization, 2003a), infection control issues (Ahmad, Krumkamp & Reintjes, 2009; Zhong & Zeng, 2008), information hotlines, surveillance (Paladini, 2004) and screening service (Kaydos-Daniels et al., 2004). Only three of them dwelt on the clinical presentation of the triage system for patients.

Chong, Tham, Goh & Seow (2005) described different signs and symptoms presented in Emergency Departments from 327 probable and

suspected SARS cases in Singapore, retrospectively. They proved that no diagnostic or screening kit for SARS was available except those defined by the World Health Organization. However, only 90% had fever and around 10% did not have the typical signs and symptoms as suggested by the WHO. No tool exists for easy and rapid screening for SARS upon admission to emergency department and nearly half of the patients discharged were readmitted afterwards with high index of suspicion of contracting SARS. Hence, avoiding discharging a patient suffering from SARS with the intention to prevent close social contact in the community is essential.

For infectious disease, the epidemic consequences of and control measures for emerging disease are governed by their mode of transmission. The initial caregivers are challenged with key responsibility to think of whether this is an introduction of an emergent infectious disease and to obtain appropriate history and information to alert and inform fellow colleagues for source containment, contact

investigation and proper infection control measures.

To review our local experience gained from SARS in 2003, effective communication among health care professionals and the administrators, surveillance, risk assessment, adherence to infection control measures play important roles in containing and controlling the spread of the disease.

3.5 Conclusions

Infectious disease is no longer confined to a regional area, and now it may be able to be transmitted around the world in a day while the development of the required medications does not happen at the same rate. In order to minimize the negative impacts to society and economy, an effective and precise method to identify the potential patients is essential.

There are numerous techniques for diagnosing a patient, but it takes time and financial resources to obtain correct and precise results. Though the skill of obtaining or extracting the specimen for gold standard diagnostic test helps to decrease the contamination problem, time is still an important factor in the combat to keep pandemic under control. Time for the transportation of the specimen, and time for the reagent to act on the cells or the immune system to create certain amount of the markers for reaction are few of the examples of observed delays. Hence, if we can screen the patient accurately to find out the source of disease promptly during the early phase, the propagation of the disease may have the chance to slow down. Although the sensitivity and specificity of a test are virtually constant whatever the prevalence of the condition is, the positive and negative predictive values depend crucially on the prevalence. This is the reason that general practitioners are reluctant to use the test developed exclusively by a specific population, and that is why a good diagnostic test is not necessarily a good screening test.

Therefore, a user-friendly, simple and cheap screening method will be the most valuable solution to solve the disadvantages of the diagnostic tests. Several kinds of prediction methods for screening will be discussed in the next chapter.

CHAPTER 4

PREDICTION METHODS

4.1 Introduction

This chapter presents the pros and cons of the common health care related to statistical prediction methods: dichotomous tests, Baye's Theorem, continuous tests like logistic regression, receiver operating characteristic (ROC) curves. An introduction of using data mining as an alternative method for predicting the accuracy of the diagnosis in this study is discussed and different approaches of/towards prediction in data mining methods is discussed.

4.2 Essential concepts for prediction in health care settings

Diagnostic testing is a critical factor in clinical decision making which may impose undesired or unintended consequences. Most common tests usually provide results along with a continuous quantitative scale (such as cell counts) throughout a range in which clinicians obtain the diagnostic result in the form of disease present or disease absent. With this system, some criteria or cutoff points are established based on comparison. Such cutoff points are usually selected by a gold standard to identify the disease in question or they are based on the statistical and conceptual analysis that balances the rate of false-positive and false-negative results. It is assumed that the disease or diagnosis being considered is mutually exclusive and the result of each diagnostic test is independent from the results of all other tests.

Typically, the distribution of positive and negative laboratory-test results follow some types of distribution curve with different mean points. Patients with or without the disease have a very high or very low value while the majority of patients will have results centered in the middle

around the mean as shown in (McGee, 2010). Patients with diseases are shown in the upper distribution while those shown in the lower distribution are free from infection. The cutoff criterion line 1 distinguishes the infected patients over the right upper pink region while the non-infected patients are shown on the left side of the lower grey region. Some patients above and/or below the selected cutoff point may not be able to be clearly classified. The sensitivity and specificity of the test can be adjusted by shifting the cutoff criterion line to the left and/or the right accordingly.

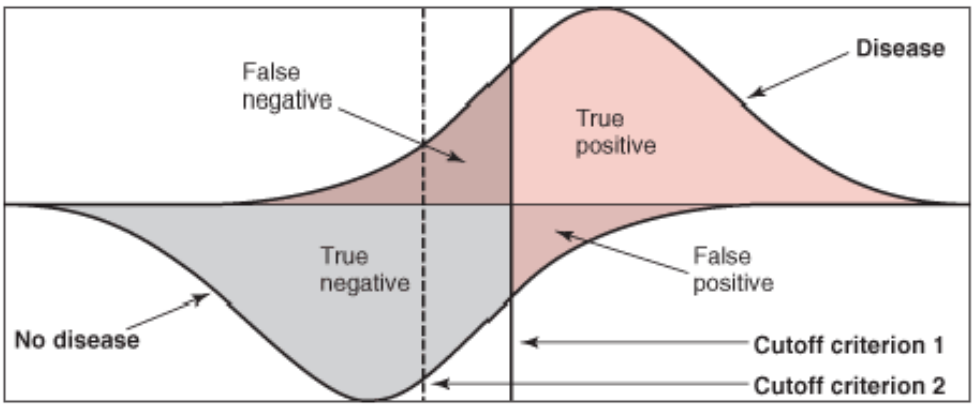

Figure 2 Distribution of test results (McGee, 2010)

For example, if we shift the line to the left reaching to the position of the cutoff criterion line 2, then the test distinguishes more patients with the

disease (an increase in sensitivity) but at the same time an increase can be seen in the number of the false-positive results (a decrease in specificity). Hence, each cutoff criterion is actually adjusting the specific probability of true-positive and/or false-positive results.

The ultimate aim of a diagnostic test is to help the clinician to make a decision on whether a patient contracts a disease and to determine the treatment scheme correspondingly. The validity of a test is based on how strong the decision is being supported to confirm the probability of the disease.

4.3 Common prediction methods used in health care settings

There are many different statistical prediction methods used in health care settings that the most common ones are used for continuous test

scores.

4.3.1 Dichotomous tests

This is a very ideal situation to find no false positive or false negative results. All positive test results denote patients who truly have the disease with a 100% positive predicted value (PPV) while all patients with negative results have a 100% negative predicted value (NPV) and do not have the disease.

However, in reality all tests have false positive or false negative results, and thus the prevalence of a disease needs to be known to further calculate the likelihood of the occurrence of the disease. Flament, Whitaker, Rapoport, Davies, Berg, Kalikow, Sceery and Shaffer (1998) gave an example of the Leyton inventory, a screening instrument for obsessive-compulsive disorder in which the sensitivity was 75%, the specificity was 84% and a predictive value of 18%. From this case, we

can see that the test results provided a definitive diagnosis but the results only estimated the probability of a disease being present or absent. Such kind of results usually varies greatly based on the test's sensitivity and specificity. Clinical scoring systems such as pre-test probability testing are not precise measurements. This scoring system makes use of clinical judgment of symptoms and signs that suggest the disease is present. It is shown that the higher the calculated score the higher the estimation of the occurrence of the patient contracted the disease.

4.3.2 Odds-likelihood calculation based on Baye's Theorem

Unlike sensitivity and specificity (which do not apply to specific patient probabilities), the likelihood ratios allow clinicians to interpret test results in a specific patient, provided there is a known pre-test probability. Baye's Theorem calculates the conditional probability of simultaneous occurrence of a positive diagnostic test and having the

disease. The odds-likelihood involves the revised prior probabilities which express the odds that the test result occurs in patients with the disease versus with those who without the disease. It often occurs in pairs in which one likelihood ratio is for a positive test and the other is for the negative test (Dawson-Saunders & Trapp, 1994).

4.3.3 Presentation of results using Receiver Operating Characteristic (ROC) Curve

The ROC curve is a plot of the sensitivity (or true-positive rate) to the false-positive rate (Dawson-Saunders & Trapp, 1994). By convention, the true-positive fraction is placed on the y-axis while the false-positive fraction is placed on the x-axis. The closer a ROC curve is to the upper left-hand corner of the graph the more accurate it is, because the true-positive rate is one and the false-positive rate is zero. It can also act as a graphical presentation of the tradeoff between the sensitivity and the specificity when the cutoff point is adjusted as shown in Figure 3.

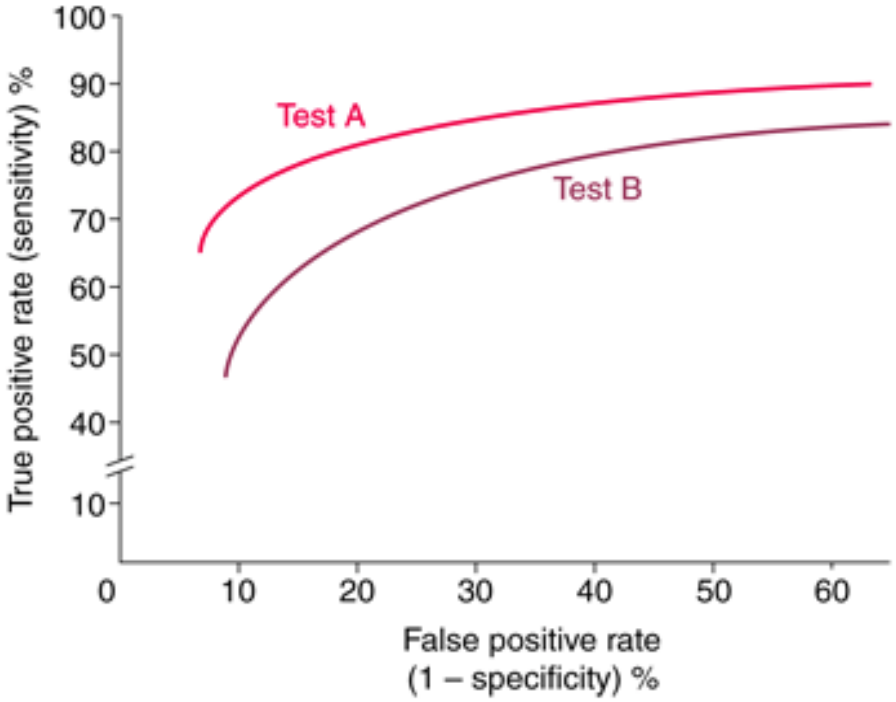

Figure 3 A typical receiver operating characteristic (ROC) curve

(Williams, Hand, & Tarnopolsky, 1982)

It is important to recognize that an empirical ROC curve is a retrospective calculation that using the same data to estimate the score and to assess its performance. Sometimes, it is quite difficult to evaluate how well the score distinguish between future and independent cases during a prospective assessment (Copas & Corbett, 2002).

4.3.4 Logistic regression

Logistic regression is used for predicting the probability of occurrence of an event by fitting data to a log function logistic curve (Dawson-Saunders & Trapp, 1994). It is used extensively in medical and social sciences and also in the marketing in order to predict a customer's propensity to purchase a product or cease a subscription. It is a generalized linear model making use of several predictor variables for binomial regression.

However, most of the regression models have their own assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model leads to erroneous estimations, bias results or even invalid conclusions. Moreover, it is often used to find a linear combination of covariates which best distinguishes between two groups or populations.

4.3.5 Bootstrap calculation

Efron (1982) used re-sampling methods to provide the basis for studying the behavior of estimates. He used sampling with replacement to draw numerous bootstrap sample sets, calculate the estimate from each bootstrap sample sets and then further analyze the sampling distribution (Izrael, Battaglia, Hoaglin, Battaglia & Abt Associate Inc., 2002). This technique is used when the theoretical distribution of a statistic is complicated or unknown or when the sample size is insufficient for straightforward statistical inference.

However, the exact number of bootstrap samples requires for calculating the estimates are not clearly defined and the number of bootstrap sample sets is increased in line with computing power. Moreover, the underlying assumptions such as independence of samples is always being ignored or overlooked as they are not formally stated as in other statistical approaches.

4.4 Use of data mining as an alternative

It is clear that physicians want to rule out the presence of a disease with a reliable negative laboratory result, and they want a high sensitive test to prove the evidence of a disease to have an accurate diagnosis for treatment. Hence, when we want to choose the most appropriate prediction method, we often look for a test with both comparatively high sensitivity and high specificity. The use of data mining can be one of the alternatives, as it can choose the value of sensitivity and specificity arbitrarily.

Data mining is defined as the process for "the extraction of implicit, previously unknown and potentially useful information from data in large databases" (Han & Kamber, 2000). Usually, it will employ one or more computer learning techniques to automatically analyze and extract knowledge (Roiger & Geatz, 2003).

The term ‘data mining’ is often used interchangeably with “knowledge discovery in databases” (KDD). Technically, data mining can be viewed as the core process of knowledge discovery. The knowledge gained from a data mining session is taught through the use of a model or generalization of the data.

All the mined rules, for instance, if attribute X and attribute Y exists then attribute Z will exist (i.e., X&Y⇒Z; both factor X and factor Y causing specified result Z), are governed by the “support” and the “confidence” the researcher specified. Support is the probability that a transaction contains all the attributes (i.e. both factors X, Y and specified result Z coexist in the same sample) whereas confidence refers to the conditional probability that a transaction (sample) having the different attributes (factors X & Y coexist) also contains specified result Z.

Statistics and Data Mining actually have a common aim, that is, they both aim at discovering structure in data. Data mining is sometimes

referred to an exploratory data analysis method involving a large number of the data sets.

Hand (1998) stated that statistics referred to the primary analysis of data collected for a particular question, or with a set of questions in mind, where data mining is entirely concerned with the secondary data analysis or finding unsuspected relationships which are of interest or value to the database owners.

Fundamentally, statistics is a discipline that prefers rigorous data analysis so that the result is not purely based on chance but also on valid inference that can be made from samples. Sometimes it can be detrimental to discover the minute or unexpected relationships as it promotes an overcautious attitude. On the other hand, data mining deals with all cases in the population recruited in the database and thus, any notions of significance testing lose. Though it cannot be used in a prospective way, it makes predictions about the future based on

information from the past and present. It makes use of the monotonicity properties of the goodness-of-fit measures in model search algorithms rather than using probabilistic statements about generalizability based on them (Cabena, Hadjinian, Stadler, Verhcees & Zanasi, 1998). Statistics does not usually recognize important ideas through timely statistical procedures. Data mining can certainly help in seeking patterns as it inherits the adventurous attitude of machine learning properties and makes use of different automatic data analysis techniques in scalable databases.

The term ‘model’, although used in both the statistics and data-mining disciplines, actually refers to different things. In statistical models, the relationships between the variables are analyzed based on some theories, while an atheoretical summary description of the data is employed in data mining. Moreover, the algorithm plays a central role in data mining while the computation procedure and the selection of a proper model are important in statistics (Hand, 1999).

Although data mining is primarily concerned for looking up unsuspected features in data, we often desire to check whether the data support some ideas about the value of a parameter in/during practice. As stated by Roiger & Geatz in 2003, "…we can state a general hypothesis about what we hope to find in the dataset. A hypothesis is an educated guess about what we believe to be true for some of all of the data…" (p.16). It is a form of posing queries and maintaining control of the search for patterns in data.

4.5 Data mining techniques

All data-mining methods are induction-based learning, that is, knowledge is generated by observing specific examples of concepts and eventually a general concept is defined. The most popular techniques are association rule mining, classification with prediction and clustering

methods.

4.5.1 Mining Association Rules in Large Databases

Association rule mining finds frequent patterns, interesting associations, correlations or causal structures among sets of items or objects in transaction databases, relational databases and other information repositories. It is commonly used in business and marketing fields (Kudyba, 2004). The most typical applications include basket data analysis in supermarkets, cross-marketing, catalog designs, and loss-leader analysis. The discovery of interesting association relationships among huge amounts of business transaction records can help in decision-making processes.

According to Han & Kamber (2000), association rules can be classified into several categories based on different criteria:

1. Based on the types of values handled in the rule, associations can be classified into Boolean versus quantitative/ either Boolean or quantitative associations. A Boolean association shows the relationships (presence or absence) among discrete (categorical) objects. In contrast, a quantitative association is a multidimensional association that involves numeric attributes which are discretized dynamically.
2. Based on the dimensions of data involved in the rules, associations can be classified into single-dimensional versus multidimensional. Single-dimensional association involves only one attribute and shows intra-attribute relationships. On the contrary, multi-dimensional association rule involves two or more distinct dimensions or attributes and shows inter-attribute relationships.
3. Based on the levels of abstractions involved in the rule, associations can be classified into single-level versus multilevel. In a single-level association, the items or predicates mined are

not considered at different levels of abstraction. On the other hand, multilevel association does consider multiple levels of abstraction. An example of different levels of abstraction is the item ‘computer’, which is a higher level abstraction of ‘laptop computer’.

4. Based on various extensions to association rule mining, association mining can be extended to correlation analysis, and the mining of maximal frequent patterns and frequent item sets. Correlation analysis and causality analysis, sequential association rule mining or constraints enforced mining are all considering maxpattern and frequent closed item sets.

Different mining methods are used for mining different types of association rules. For example, the Apriori Algorithm is best for finding Iceberg queries such as single-dimensional Boolean association rule from transaction database, the level-cross filtering in mining multilevel association rules from transaction database.

The two key steps in association rule mining are the searching for Frequent Item Set and Rule Generation. The "support" is used to mine the frequent item sets while the "confidence" is used by the rule generation step to qualify the strength of the association rule.

Association rule mining is probably the most significant method to characterize the problem of mining risk patterns as an optimal rule of discovery problem (Li, Fu, et al., 2005). Its unique contribution from the database community in knowledge data discovery has been published in a large number of papers with numerous interesting issues. Brossette, Sprague, Hardin, Jones & Moser (1998), for example, uncovered association rules in hospital infection control and public surveillance data. They presented this novel process and data mining surveillance system which utilize association rules to identify new and interesting patterns in surveillance data in Birmingham in the United States. Although its analysis on Pseudomonas aeruginosa infection did not show statistically significant events are clinically significant, it did identify

some potentially significant shifts in the occurrence of potential antimicrobial resistance of the bacteria. While Li, Fu and Fahey (2008) confirmed that association rule mining was an efficient approach to explore risk patterns associated with the use of ACE inhibitors using the optimal risk pattern set. They pointed out that many studies cannot give much support because the dataset is usually very large and skewed like in a million records. A large majority of them are normal as the disease is usually rare in comparison with the healthy population. Hence, support of any pattern which is actually the ratio of the number of records containing that pattern to the number of all records in the dataset, will be very low/limited. Instead, they explore an anti-monotone property to support efficiently mining optimal risk pattern which seems to be more user friendly to the medical professionals. However, there is room for further research in other types of data such as spatial data, multimedia data and time series data.

4.5.2 Classifications and Prediction

Classification predicts categorical class labels which are to assign a class to find previously unseen records as accurately as possible. Based on the training set, a model can be constructed through the training set. The class labels in a classifying attribute are used in classifying new data. Classification can be done by Bayesian classifiers (predicting class membership probabilities), back propagation (connectionist neural network learning) or concepts based on association rule mining (using either association rule clustering system, or associative classification, or classification by aggregating emerging patterns).

Data classification is a two-step process. The first step is to build a model describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by class label attributes. The model is represented as classification rules,

decision trees or mathematical formulae. This model construction incorporating training samples is often considered as supervised learning as the training data with labels indicate the class of the observations served as the supervision in machine learning. The second step is the model usage for classifying future or unknown objects. The accuracy of the model is estimated/can be tested by comparing the classified result from the model versus a known label of a test sample. Accuracy is shown in terms of the percentage of test set samples that are correctly classified by the model (Roiger & Geatz, 2003).

The decision tree induction is the most frequent type of the classification used. *Iterative Dichotomiser 3 (ID3)* and *Classificatier 4.5 (C4.5)* are greedy algorithms for the induction of decision trees. Each algorithm uses an information-theoretic measure to select the attribute tested for each non-leaf node in the tree. Pruning algorithms attempt to improve accuracy by removing tree branches reflecting noise in the data.

Prediction is a model describing continuous-valued functions that can be used to predict unknown or missing values. This is modeled by statistical techniques of regression using either linear and multiple regression or nonlinear regression (Dunham, 2003). Many nonlinear problems can be converted to linear problems by performing transformations on the predictor variables.

Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes and data transformation such as generalizing the data to higher-level concepts or normalizing the data are common approaches.

The most typical applications for classification and prediction are credit approval, target marketing and medical diagnosis. Han & Kamber (2000) listed out several criteria for comparing and evaluating different classification and prediction methods:

i. Predictive accuracy refers to the ability of the model to correctly predict the class label of new or previously unseen data.

ii. Speed refers to the computation costs involved in generating and using the model.

iii. Robustness is the ability of the model to make correct predictions given noisy data or data with missing values.

iv. Scalability is the ability to construct the model efficiently given large amounts of data.

v. Interpretability refers to the level of understanding and insight that is provided by the model.

Chang & Wang (2006) had used classification tree techniques to analyze the risk factors that can influence the injury severity in traffic accidents. Their non-parametric tree-based (CART) model provides good overall predictions for learning data and testing data by identifying nine critical predictor variables from twenty predictor variables with an overall accuracy of more than 90%.

Trujillano, Badia, Servia and Rodriguez-Pozo (2009) also utilized the classification tree method to predict the severity of critically ill patients retrospectively from the hospital database. Five variables were identified with more than 10 decision rules and when the results were further tested with the conventional logistic regression, the correlation matrix were all >0.76 which was considered to be very high.

When discussing the usage of classification methods for infectious disease situation, data mining is mostly applied in laboratory works, that is, finding some special patterns from the DNA (Zheng, 2005) or any enzymes with specific peptides (Weingart, Lavi & Horn, 2009) or in the use of surveillance tools such as mining the social mixing patterns for infectious disease models in Belgium by Hens, Goeyvaerts, Aerts, Shkedy, Van Damme & Butels (2009) who were building a European distributed clinical data mining network to foster the fight against microbial diseases.

4.5.3 Clustering techniques

In contrast to classification, clustering is considered to be an unsupervised learning as the class labels of the training data are unknown. It is just given a set of measurement or observation with an aim of establishing the existence of classes or clusters in the data. Clustering acts as a stand-alone tool for gaining insights in data distribution or as a preprocessing step for other algorithms.

In order to use clustering effectively, one must have a scalable database which can deal with different types of attributes for discovering clusters with arbitrary shape(s). One should also have minimal requirements for domain knowledge to determine the input parameters. Examples of clustering applications include marketing, land use, insurance and city planning.

In addition, a good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measures used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

There are many clustering methods used to create pattern recognition and spatial data analysis. Partitioning algorithms involves the construction of various partitions and later the evaluation of the various partitions by some criteria. For instance, K-means clustering and K-medoids clustering methods belong to this type. The disadvantage of this method is the specification of the number of clusters needed that can be very difficult to determine. Hierarchy algorithms builds a hierarchical decomposition of a set of data (or objects) using either agglomerative or divisive approaches. Though this method does not require the number of clusters as an input, it demands a termination condition. Density-based

clustering is based on the connectivity and density functions whereas grid-based clustering is based on a multi-level granularity structure. Model-based clustering refers to the hypothesized models for each of the clusters and the idea is to find the best fit of that model to each other. Outlier detection and analysis are very useful for fraud detection and can be performed by statistical, distance-based or deviation-based approaches.

A common technique to evaluate the cluster quality is to calculate the summation/sum of the square-error differences between the instances of each cluster and their corresponding cluster center. The smaller the values calculated the higher the cluster quality (Roiger & Geatz, 2003).

4.5.4 Comparison of the three data mining techniques

There are certain similarities and differences among the three mentioned data mining techniques, below is the summary of the comparison of

these three data mining techniques used/applied.

Table 1 Comparison of the three commonly used data mining techniques

	Association Rule Mining	**Classification & Prediction (Supervised learning)**	**Clustering (Unsupervised learning)**
Technique used	Intelligent search technique	Machine learning; Statistical technique; Models continuous-values functions	Nearest neighbour technique
Aim of the technique	Identifies: 1. Frequent patterns 2. Interesting associations (e.g. spatial, sequential) 3. Interesting correlations 4. Interesting causal structures 5.	1. Predicts categorical class labels 2. Classifies data 3. Predicts unknown or missing values	1. Recognizes patterns 2. Analyzes spatial data
Key Steps	1. Candidate generation 2. Searching for Frequent Item Set 3. Rule Generation	1. Data Preparation 2. Model Construction for training 3. Pruning 4. Model Usage	1. Preprocessing and feature selection 2. Similarity measure 3. Clustering algorithm 4. Result validation 5. Result interpretation and application
Common approaches	1. Types of Database • Transactional • Relational • Information repositories 2. Types of Value • Boolean • Quantitative 3. Dimensions involved • Single dimensional • Multidimensional 4. Levels of abstraction • Single level • Multilevel	1. Bayesian Classifiers 2. Back propagation 3. Neural Network	1. Partitioning • K-means clustering • K-medoids clusters 2. Hierarchical • Agglomerative approach • Divisive approach 3. Density-based 4. Grid based 5. Model based 6. Outlier analysis

Table 1 Comparison of the three commonly used data mining techniques (cont'd)

	Association Rule Mining	**Classification & Prediction (Supervised learning)**	**Clustering (Unsupervised learning)**
Evaluation criteria	1. Objective: Support and Confidence 2. Subjective: Unexpected and Actionable	1. Predictive accuracy 2. Speed 3. Robustness 4. Scalability 5. Interpretability	1. High intraclass similarity 2. Low interclass similarity 3. Ability to discover some or all hidden patterns
Common application	Business and marketing field: 1. Basket data analysis in supermarket 2. Cross-marketing 3. Catalog design 4. Loss-leader analysis	1. Credit approval 2. Target marketing 3. Medical Diagnosis	1. Stand-alone tool for data distribution 2. Preprocessing step for other techniques 3. Customer segmentation 4. Fraud detection 5. Medical analysis
Common problems encountered	1. Not efficient for huge candidate sets and needs/requires multiple scans of database 2. Interest measurements 3. Causation not indicated, hence should not be used for prediction.	1. No one particular approaches is superior than others, only trade-offs among the evaluation criteria 2. Need longer training time for scalable data 3. Not easy to incorporate domain knowledge	1. Difficult to decide appropriate number of clusters 2. Hierarchial methods are biased towards finding spherical clusters 3. Current clustering techniques do not address all the requirements adequately

From the comparison, the supervised learning method (classification and prediction) seems to be the most appropriate method for predicting medical diagnosis as it can successfully predict categorical class label. Association rule mining can identify interesting (or unexpected) associations, correlations and causal structures within the data which is valuable to the identification of new knowledge while unsupervised learning (clustering) can identify special pattern from the data especially the spatial relationship in the data.

4.6 Reviewing existing prediction rules

Leung et al. (2004) tried to determine/set a clinical prediction rule for diagnosing SARS in the Emergency Department using retrospective two-step coefficient-based multivariable logistic regression scoring method with internal validation by bootstrapping. Results were based on the points assigned on the basis of history, physical examination and simple

investigations obtained from 2,649 consecutive patients admitted during the period of 10th March 2003 to 10th May 2003 at 1 hospital and the period of 12th March to 31st May 2003 at the other hospital in Hong Kong. All patients had fever with respiratory and/or gastrointestinal infectious syndrome(s). Only 556 (21%) patients had positive laboratory diagnosis of SARS.

They found that the predictors of SARS on the basis of history (step 1) included previous contact with a patient with SARS and the presence of fever, myalgia and malaise. Adults aged 65 or above and youngsters aged below 18 years, with the presence of sputum, abdominal pain, sore throat and rhinorrhea were inversely related to have contracted SARS. Step 2 showed that haziness or pneumonic consolidation on chest radiographs, and low lymphocyte and platelet counts, in addition to positive contact history and fever, were associated with a higher probability of contracting SARS. A high neutrophil count, the extremes of age, and sputum production were associated with a lower probability

of contracting SARS. The quoted sensitivity and specificity value for Step 1 were 99% and 63% and that of Step 2 were 95% and 57%. Chen et al. (2004), however, concentrated on the patients with fever exposure who attended AED within fourteen days in one Taiwan teaching hospital from 15th March to 2nd April 2003. Seventy patients were enrolled and eight patients (~11%) were diagnosed as probably contracting SARS. Data underwent univariate analysis with two sets of clinical decision rules developed.

The first set called Symptom score relies only on signs and symptoms. Two clinical symptoms, myalgia and diarrhoea, were found to be positively correlated to SARS and other two symptoms including cough and rhinorrhea/sore throat had a negative correlation. The second set, the Clinical score, included two more laboratory results, lymphopenia and thrombocytopenia. The quoted sensitivity and specificity value for the Symptom score were 100% and 75.9% and that of the Clinical score were 100% and 86.3%.

On the other hand, Wang, T. L. et al. (2004) included patients with fever and exposure to SARS who attended the 3 EDs of Taiwan teaching hospitals during 1st March to 20th April 2003. Twenty patients (11.5%) out of 175 patients were identified as probable/suspected SARS cases and their data included radiographic tests, laboratory tests as well as those of clinical signs and symptoms were included for multivariate analysis.

Seven database variables were identified by the authors. Their findings accounted for over 98% of all clinical information at the early stage. These variables included history of exposure or traveling, chest radiographs with multilobar or bilateral infiltration, lymphocytopenia, high C-reactive protein (CRP), monocyte predominant sputum smear, prolonged partial prothrombin time (APTT), and elevated lactate dehydrogenase (LDH). The quoted sensitivity and specificity were 100% and 93% with positive and negative predictive values of 83% and 100%.

These three studies have their own uniqueness. Chen, S.Y. et al (2004) and Wang, T. L. et al (2004) only had limited probable case loads for evaluating the signs and symptoms in the analyzing process. A brief summary of these three studies is presented in Table 2 for comparison.

Table 2 Summary of three SARS prediction studies

Study Group	**Chen et al. (2004)**	**Wang T.L. et al. (2004)**	**Leung et al. (2004)**
Setting	ED of a Taiwan teaching hospital	ED of 3 Taiwan teaching hospitals	ED of 2 Hong Kong hospitals
Inclusion criteria	Fever with exposure	Fever with exposure	Fever with respiratory, gastrointestinal, infectious symptoms
Period	15/3 to 2/4 2003	1/3 to 20/4 2003	10/3 to 10/5 2003 12/3 to 31/5 2003
Samples	8 probable SARS cases 62 non-SARS cases	20 probable SARS cases 155 non-SARS cases	377 + 184 probable SARS cases 897 + 1191 non-SARS cases
Analyze	univariate analysis	multivariate analysis	Multivariate analysis

Clinical predictors	Myalgias Diarrhoea Cough Sore throat/rhinorrhea Lymphopenia Thrombocytopenia	Known exposure/travel history Dyspnoea Myalgia Multilobar infiltrates Lymphocytopenia thrombocytopenia C Reactive Protein Monocyte predominance on sputum's Gram's stain Elevated APPT Elevated LDH	Health care worker Known contact history Fever Rhinorrhea Malaise Chest radiography Leukocyte count Lymphocyte count Neutrophil count Platelet count

4.7 Conclusions

The aim of screening is to predict whether the patient has contracted the disease or not. Many different statistical methods were/have developed for predicting the disease classes of patients. Specificity and sensitivity are phrased in terms of the proportions of the true classes that are correctly classified. With the staggering amount of information available for data analysis, traditional methods of data analysis may not be able to cope with a large amount of information, hence data mining is an effective tool that can give promising results to help health care

providers analyzing the data.

Data mining strategies can be broadly classified as supervised learning or unsupervised learning. Supervised data mining algorithms only allow a single output attribute (that is the dependent variable) whereas the unsupervised data mining is to apply some measure of similarity to divide instances into disjoint partitions. Hence, the learning program builds a knowledge structure by using some measure of cluster quality to group instances into two or more classes. It can be used for both. It requires a large volume of data on hand for processing, and it needs a comprehensive database construct beforehand which may need time and manpower. Since the electronic medical database currently used in HA is usually key in after patient's discharge, though HA is migrating to paperless electronic medical records, it needs time to train up all health care professionals to use computers to record patients' progress.

Three basic techniques of data mining can be used in predicting whether

patients contract the disease. Association rule mining involves finding all of the rules which a particular data attribute is either a consequence or an antecedent. This can be very useful in investigating whether the disease is related to any demographic variables. In contrast, classification involves the need to find the rules that can partition the data into disjoint groups which is extremely important in diagnostic and treatment decision making. Clustering technique, on the other hand, is mainly used to discover structure or similarities within the data.

As data mining deals with all cases in the population recruited in the database, so any notions of significance testing lose out, hence it will be more appropriate to do the mining again if the condition is different from the current one with new specific known criteria.

CHAPTER 5

CONCEPTUAL FRAMEWORK

5.1 Introduction

Once the patient is predicted to have contracted a disease (by using prediction diagnostic tests), what is the next step to prevent (or at least minimize) further transmission of the disease? This chapter mainly aims at exploring how the prediction can help in controlling the spread of the infectious disease in the community.

5.2 Breaking the Chain of Infection

As mentioned in Chapter Two, we know that the interaction among the host (reservoir/susceptible), agent and mode of transmission in the environment are very crucial factors in the chain of infectious diseases.

Breaking the chain of infection is therefore very essential to cease further propagation/spread of an infectious disease.

5.2.1 Reservoir and susceptible host

Reservoirs are the habitats in which an agent normally lives, grows and multiplies. They can be humans, animals or in the environment per se which may or may not be the source from which an agent is transferred.

Human reservoirs may be asymptomatic or act as passive carriers to transmit a disease to others. The transmission of a disease to others commonly takes place when the hosts do not realize they are infected and this sometimes occurs in convalescent carriers who are recovering from their illness but remain capable of transmitting the illness to others. Chronic carriers may, however, harbor the pathogen for months or even years after their initial infection and pass it on to other susceptible hosts when the host's immune status becomes weak. Agents from animal

reservoirs may transmit diseases from animals to animals/ an animal to another animal. And when they attack humans, the disease is transmitted from the animals to humans. Plants, soil and water can also act as environmental reservoirs for some infectious agents.

Disease transmission takes place when humans come into contact with these environmental reservoirs, and different factors such as genetics, immunity, non-specific factors such as coughing reflex, gastric acidity, the presence of enough cilia in the respiratory tract or some disruption to the immune defense system, or factors such as malnutrition and alcoholism, contribute to the increase in vulnerability of a person to be infected. In this study, subjects were vulnerable to different mediated factors including comorbidity, geographic or demographic factors. All of these factors contribute to the manifestation of the clinical signs and symptoms of the disease in the susceptible host.

5.2.2 Portal of exit and entry

The portals of entry and exit are the specific pathways for pathogens to leave its reservoir and enter its susceptible host. For example, influenza viruses exit the respiratory tract of the source host and then enter the respiratory tract of the new host. The process when the pathogens exit and enters the host is known as the mode of transmission. It can be classified as either direct or indirect transmission (Centers for Disease Control and Prevention, 2006c).

Direct transmission means that the susceptible host is directly infected by the infectious agent without using a medium as vectors or vehicles. Direct skin-to-skin contact or droplet spread by sneezing or coughing are examples of such transmission. On the contrary, infectious agents requiring a transporting medium such as air particles or inanimate objects for transmission follow/is called as indirect transmission. A common example of this mode of transmission can be found in airborne

diseases which the droplet nuclei is carried by dust and suspends in the air for infection to take place. Another example of those transmitting medium is mosquitoes which physically carry the infectious agent and pass it on when biting humans.

5.2.3 Appropriate intervention

Most of the time, we break the infection chain by attacking the segment which is most susceptible for intervention. This measure may vary from one disease to another. For some diseases, the most appropriate intervention can be directed to the mode of the transmission so as to reduce or eliminate the number of the source infecting agents. For disease involving respiratory systems, we should interrupt both the direct and indirect transmission pathways (Centers for Disease Control and Prevention, 2006c). First, we must be able to identify who are probably infected from a vast amount of "appear to be healthy" subjects by some screening tests. After the screening procedure, the subject will most

probably undergo further diagnostic tests to confirm the diagnosis before any treatment can be started or to initiate a treatment-scheme based on the course of clinical manifestation. However, in some cases the prophylactic treatment needs to be started immediately before any screening or diagnostic test is carried out owing to the severity of the symptoms. Some will present with the disease outcome directly before any testing or treatment is given, as shown in Figure 4.

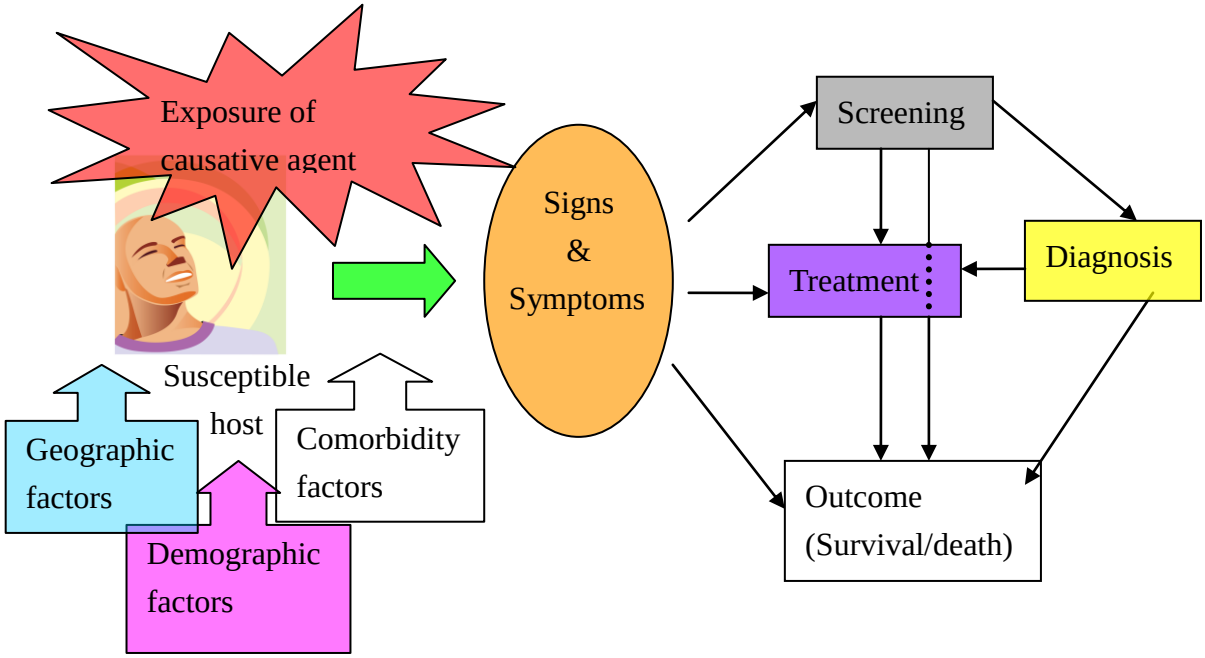

Figure 4 Concepts of screening and diagnosis in relation to infectious diseases

When a brand new type of infectious disease occurs, great challenge is faced by the health care professionals to decide an appropriate

intervention. What kind of patient should we start isolating? Are there any other alternatives that can help us to identify potential patients other than the source of infection? These questions are to be solved by figuring out the corrective measures to keep the infection under control.

5.3 Incorporating the use of data mining techniques into the management of emerging infectious diseases

Advanced information technology and the creation of electronic databases have made the use of concepts such as data warehousing possible that directly capture, store, and analyze data and convert it into useful information and knowledge. Data mining in particular can help control costs, maintain high quality and efficiency of patient care and service, and concurrently it can also serve as a surveillance of infectious diseases.

Considering SARS as an example, with the widespread use of medical

information systems including databases in the Department of Health (DH), Out-patient Department (OPD), Clinical Management System (CMS), and Accident and Emergency Information System (AEIS) from Hospital Authority, one can search information from various systems using the same identity number. A data warehouse can then be created for SARS signs and symptoms as shown in **Figure 5** based on the data from CMS, AEIS, and OPD.

Various procedures such as cleaning, integration, transformation and reliability checking are performed before data analysis.

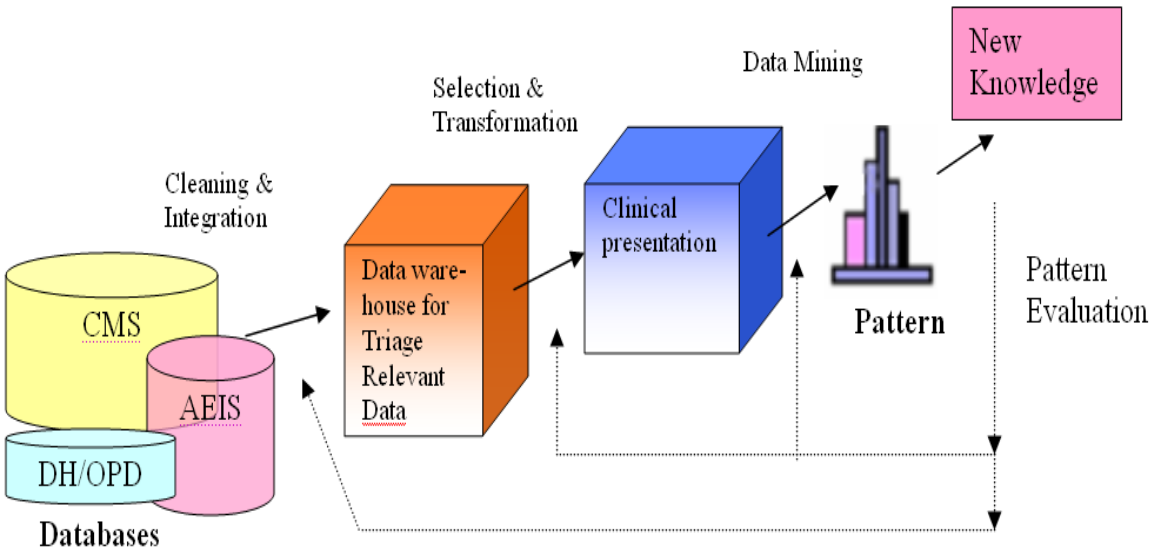

Figure 5 Overview of the steps in data mining process to develop a predictive model relative to patients who are suffering from SARS in Emergency Departments

New knowledge generated from the data mining process can be helpful to break the chain of infection by early identification and isolation of potential source of infection as shown in Figure 7.

The information obtained from databases will then be classified into different categories such as behavioral (smoking habit, drinking habit, etc), environmental (geographic, traveling abroad, etc.), intrinsic (sex, gender, past medical health status, etc.) and extrinsic factors (occupation, etc).

A snowflake scheme of the database design is shown in Figure 6 while the entire flow of the study is presented in Figure 7.

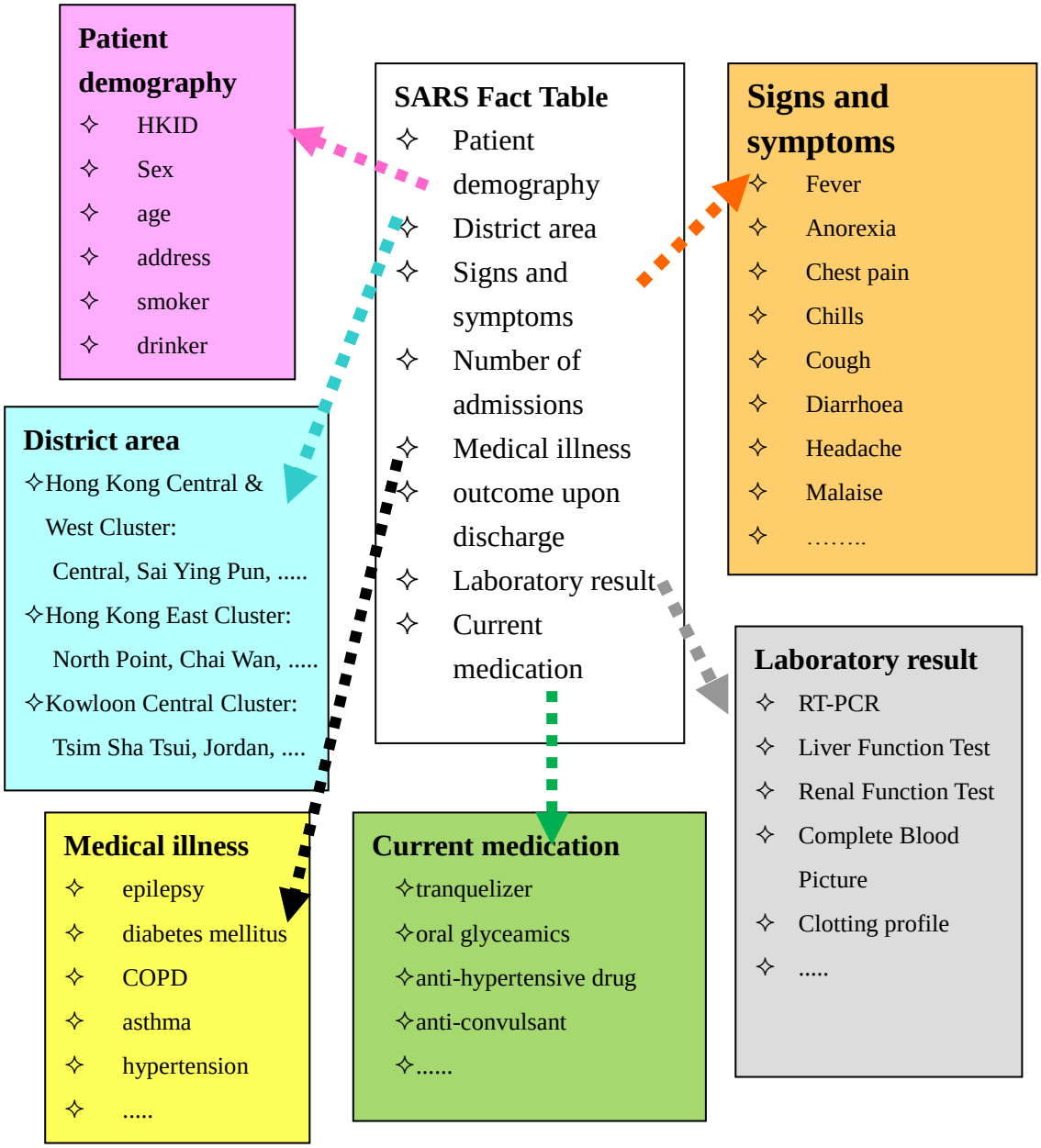

Figure 6 A snowflake schema of the database design

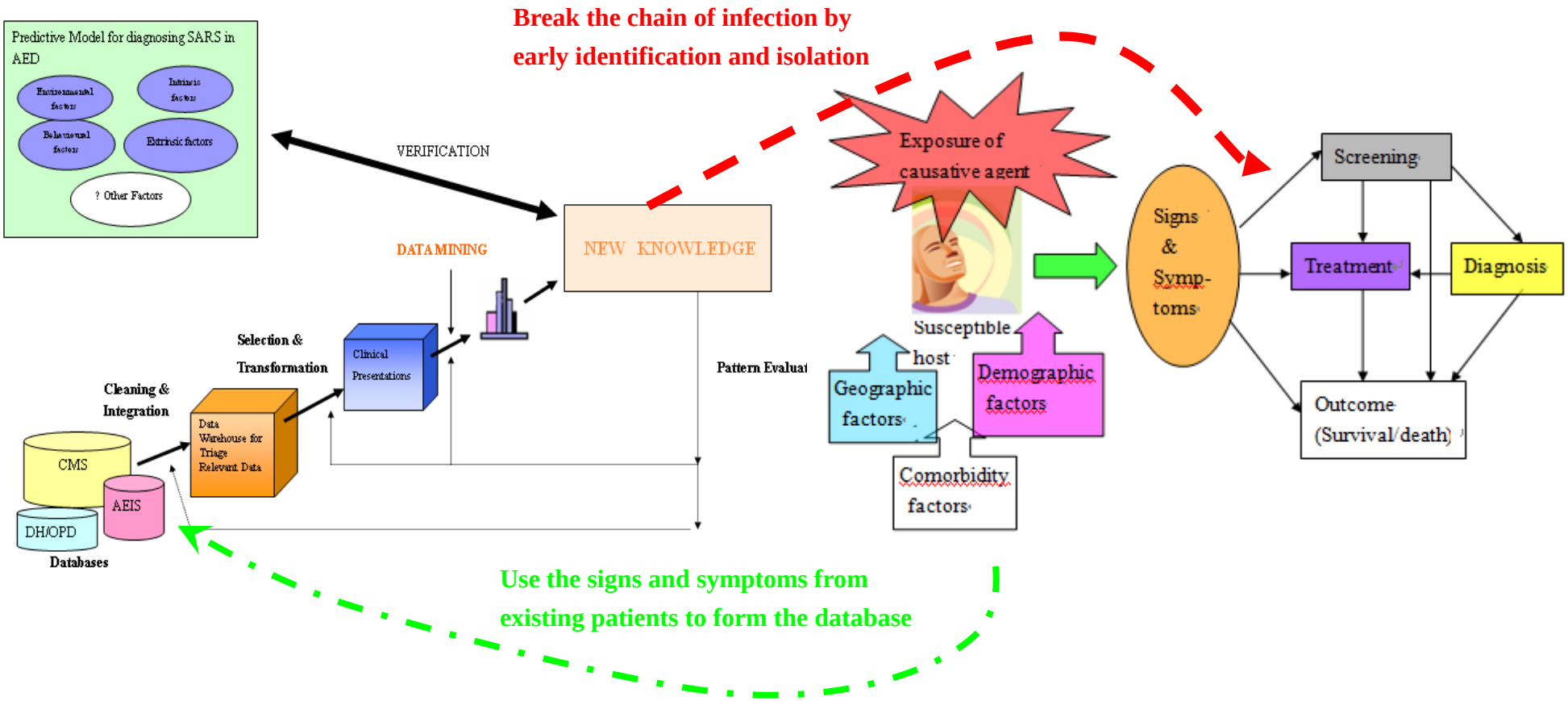

Figure 7 Theoretical framework

5.4 Conclusions

As infectious diseases attack people swiftly and unpredictably, a timely and convenient surveillance system is needed to promptly identify the causing agent and to take appropriate intervention to prevent further spread of the infection.

The WHO (2009) predicts that another influenza pandemic is due (based on the cycle prediction) since the coming of avian influenza pandemic. That the recent H5N1 case has infected eight citizens in Hong Kong (Centre for Health Protection, 2010) reflects that health care workers in general are not sensitive enough to identify potential patients. Hence, we must use smart techniques to identify potential ignition points for the disease. The data-mining technique for the identification of critical predictors for respiratory disease is proposed for this purpose.

CHAPTER 6

STAGE 1 – DATA PREPROCESSING

6.1 Introduction

This chapter describes the design, the inclusion criteria and the procedures of building up a comprehensive database from various sources.

6.2 Research objective and questions

The aim of this stage was to build a preprocessed data warehouse which contains the required data for further analysis at later stages of the study. The demographic and socio-economical profile of the subjects involved was analyzed using this data warehouse.

The research questions are:

i. Is there any gender difference between SARS and non-SARS patients?

ii. Is there any age difference between these two groups of patients?

iii. Is there any different illustration between Amoy and non-Amoy residents?

iv. Is there any difference denote between patients with and patients without morbidities history?

6.3 Methods

6.3.1 Design

It was a retrospective case record review study using observational model in a regional hospital designed for SARS patients in Hong Kong, from 1st February to 30th July 2003.

6.3.2 Inclusion criteria

All of the case notes of adult patients (aged >=18 years old) with preliminary diagnosis of SARS (including those with initial false positive results) in the AED within the period of 1st February to 30th July 2003, were included in the study.

6.3.3 Ethical considerations

The study was undertaken with the assurance of anonymity. All the information gathered was kept strictly confidential and accessible only to the researcher and restricted to the programming personnel. In addition, the database could only be accessed with the security code which would be changed every three months.

The application letter seeking for ethical approval, which explains the

nature and purpose of the study, was sent to the hospital for obtaining authentication. This was a medical record review study that did not concern the care of the patients or ~~with~~ any treatment decisions. In addition, all information precluded personal identifiers and individual consent from patients was not required.

6.3.4 Procedures

A letter requesting access approval was sent to a designated regional hospital for record retrieving.

All case notes including the Clinical Management Sheet, nursing observation charts and laboratory results were reviewed.

6.3.5 Data management

Data were categorized and entered to a database using the Statistical

Package for the Social Sciences (SPSS) (version 17) [now renamed as Predictive Analytics SoftWare (PASW) Statistics]. Data preprocessing such as cleaning, integration, transformation and reliability checking were done in this stage before proceeding to further data analysis. Missing values were identified and some of the blood test results were transformed in order to have the same testing units to be fitted into three other studies (Chen et al., 2004, Wang T.L. et al., 2004 & Leung et al. 2004). For example, in Wang's study, the unit for C-reactive protein is mg/dL while data obtained in current study is mg/L. Some of the nominal data (such as occupation) were recoded into different class labels as well. All recoding and transformation were double checked before performing any mining procedures. About one fourth of the records were randomly chosen and checked against the database for its reliability and accuracy.

6.3.6 Deliverables

A comprehensive data warehouse was made available for further data manipulation and analysis.

6.4 Results

Information obtained from the database was classified into different categories such as behavioral (smoking habit, drinking habits, etc.), environmental (geographic, traveling abroad, etc.), intrinsic (sex, gender, past medical health status, etc.) and extrinsic factors (occupation, etc.).

Five hundred and forty-nine case notes were recorded from the designated hospital and the data were entered into datasets for further analysis. All case notes including the Clinical Management Sheet, nursing observation charts and laboratory results were reviewed. A template, which contained variables as identified from the reviewed case notes that covered all aspects of case and treatment when the patients

were hospitalized for SARS, was constructed. All of the data concerning SARS was input to the data warehouse.

6.4.1 Demographic profile

Among these 549 cases, 232 (42.3%) of them were males and 317 (57.7%) were females. Their ages ranged from 18 to 99 years of age with a mean age of 43.41 (S.D. 15.63). The majority (94.4%) of them were confirmed as SARS cases by the gold standard RT-PCR method. Many of them were also non-smokers and non-drinkers, as shown in Table 3.

Table 3 Frequency distribution of smoking and drinking habits of the subjects (N=549)

	Male (n, %)	Female (n, %)
Smokers	31, 13.4%	6, 1.9%
Non-smokers	201, 86.6%	311, 98.1%
Drinkers	29, 12.5%	3, 0.9%
Non-drinkers	203, 87.5%	314, 99.1%

There is no significant association found between smoking habits and

contracting SARS status [χ^2(4)=4.874, p<0.3] as well as the drinking habits and contracting SARS status [χ^2(4)=7.031, p<0.1].

6.4.2 Socio-economical profiles

Among all the subjects, 239 (39.8%) people had contacted with other SARS patients before the onset of the illness. Out of 239 cases, 19.9% had travelled to other places and the majority of them (91.7%) had travelled to China. There is a significant difference between those who visited Amoy Garden with contracting SARS [χ^2(2)=9.586, p<0.01] and those who travelled abroad [χ^2(8)=17.242, *p*<0.05]. Concerning the occupation of the patients, the disease seemed to attack mostly housewives and health care workers while more than half of the health care workers are nurses as shown in Table 4. There is an association found between financial conditions and confirmed SARS case status [χ^2(6)=17.373, p<0.01] but not for occupation with contracting SARS status [χ^2(13)=21.976, p<0.1].

Table 4 Types of occupation distribution of the subjects (N=549)

Occupation	Number (%)	Job Title	Number (%)
Housewives	97 (17.7)		
Health care workers	87 (15.8)	Nurses	50 (57)
		Health care assistants	12 (14)
		Physicians	11 (13)
		Allied Health care workers	14 (16)
Retired	57 (10.4)		
Elementary workers	56 (10.2)		
Clerk	55 (10.0)		
Professionals	51 (9.3)		
Service	44 (8.0)		
Unemployed	28 (5.1)		
Students	14 (2.6)		
Machinery operators	14 (2.6)		
Others	46 (8.8)		

6.4.3 Geographical profiles

The majority of the subjects (41.8%) lived in Kowloon East district, 11.3% lived in Shatin (in the northern part of the New Territories) and 10.5% lived in Tsuen Wan (in the western part of the New Territories). Amoy Gardens, which is located in the eastern part of Kowloon peninsula, alone accounted for 32.3 % of the subjects. The geographic

distribution of the subjects was presented in the map in Figure 8. There was a significant association between being an Amoy Gardens resident and contracting SARS status [χ^2 (1)=12.905, p<0.001], while no significant association was found for residents of the Lower Ngau Tau Kok Estate and contracting SARS status [χ^2 (1)=1.437, p<0.5].

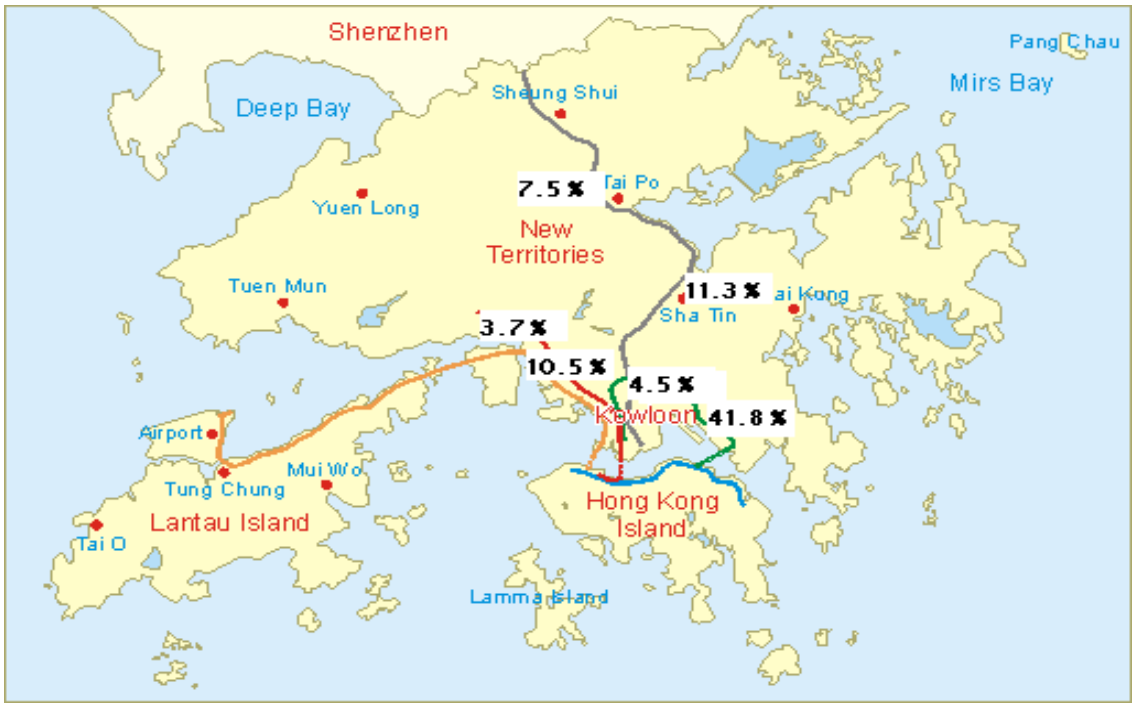

Figure 8 Geographical distribution of the subjects diagnosed with SARS

6.4.4 Co-morbidity profiles

Around 31.2% (171 subjects) had underlying medical illnesses. Hypertension (28.3%) was the most common disease followed by diabetes mellitus (18.2%), chronic obstructive pulmonary disease (12.8%) and cardiovascular disease (12.8%). There is no significant association found between patients having underlying medical illness and contracting SARS status [χ^2 (1)=0.667, p<0.5].

6.4.5 Laboratory result profiles

The first set of blood results of each patient were compared from the data gathered. The majority of the laboratory test results did not show significant differences between SARS and non-SARS patients.

6.4.5.1 Renal function tests

The results from the renal function test did not show significant difference between SARS and non-SARS patients. There were only slightly more cases of hypernatremia (3 cases [0.6%] of serum sodium content > 145mmol/L) and hyperkalemia (12 cases

[2.4%] of serum potassium content >5.1mmol/L) in the SARS group. The range of serum creatinine had a greater distribution in non-SARS patients (47--858 μmol/L) than SARS patients (0.4--1112 μmol/L). Those SARS patients showed no significant difference in various electrolyte levels compared with non-SARS patients. Figure 9, 9a and 9b show the boxplots of selected renal function test results by disease groups for all patients (N=528), for patients without comorbidity (N=356) and those with comorbidity (N=172). From the graphs, there are no significant difference in levels of various renal function tests between SARS and non-SARS patients even stratified for presence and absence of comorbidity.

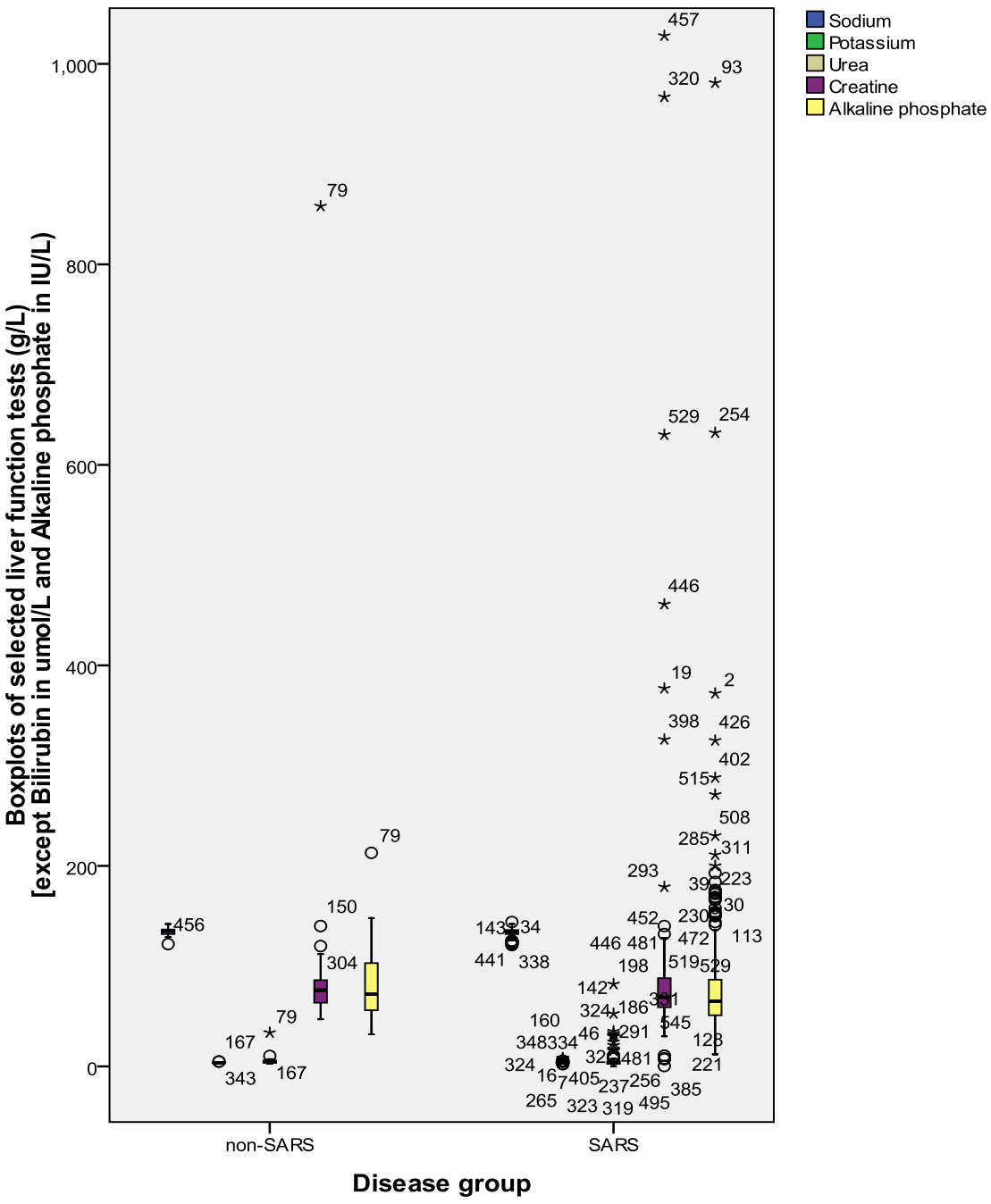

Figure 9 Distribution of selected renal function test results of disease groups (N=528)

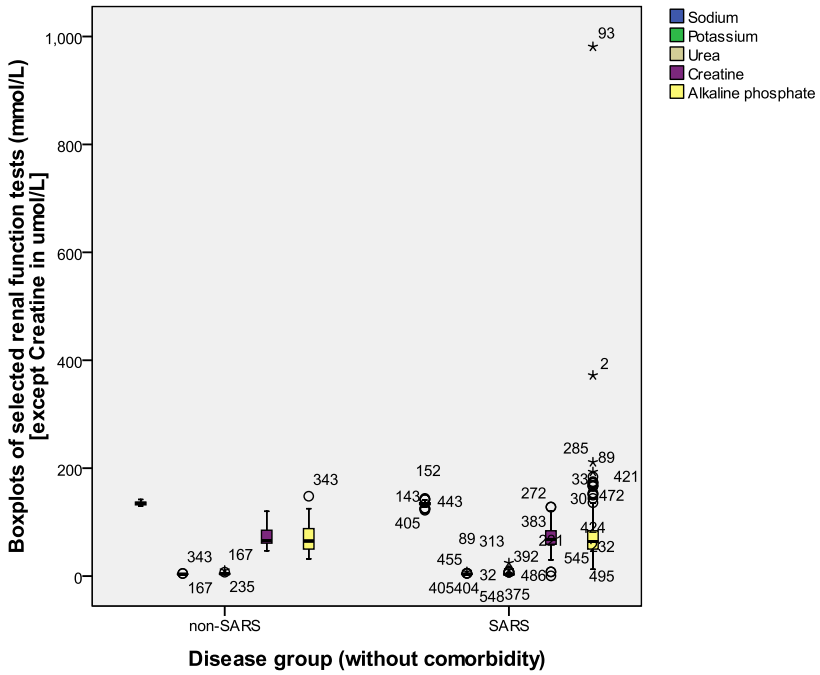

Figure 9a Distribution of selected renal function test results of disease groups for those without comorbidity (N=356)

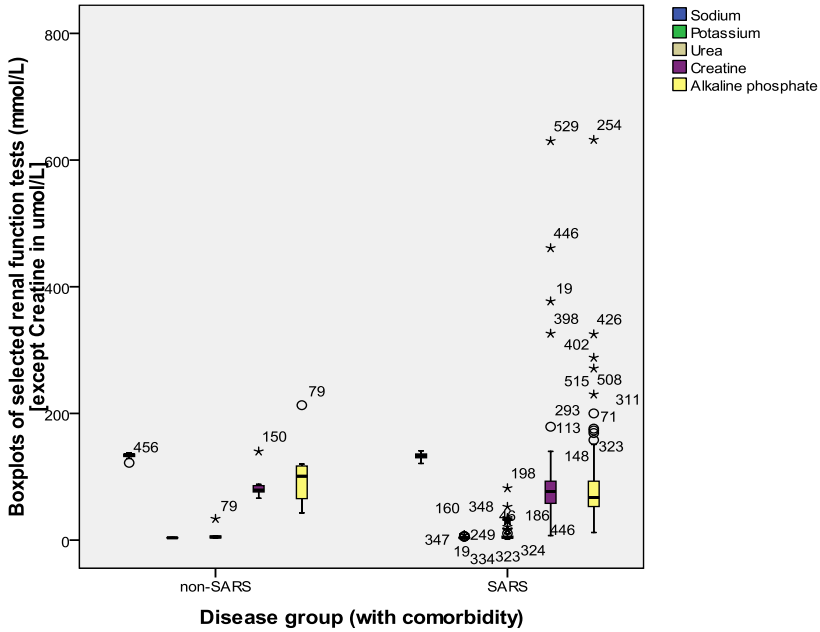

Figure 9b Distribution of selected renal function test results of disease groups for those with comorbidity (N=172)

6.4.5.2 Liver function tests

The results of the liver function test did not show any significant difference between the SARS and non-SARS patients. There were only slightly more cases of high globulin content (3 cases [11.1%] of serum globulin content > 44 g/L or < 9 g/L) in non-SARS patients compared to SARS patients. On the other hand, a higher percentage of SARS patients (20.1%) showed higher serum alanine aminotransferase (ALT) results (ALT >55 IU/L) than the non-SARS group (14.8%). Once again, the SARS patients experience no significant differences in levels of various electrolytes compared to non-SARS patients. Figure 10, 10a and 10b show the boxplots of selected liver function test results by disease groups for all patients (N=525), for patients without comorbidity (N=354) and those with comorbidity (N=171). From the graphs, there are no significant difference in levels of various liver function tests between SARS and non-SARS patients even stratified for presence and absence of comorbidity.

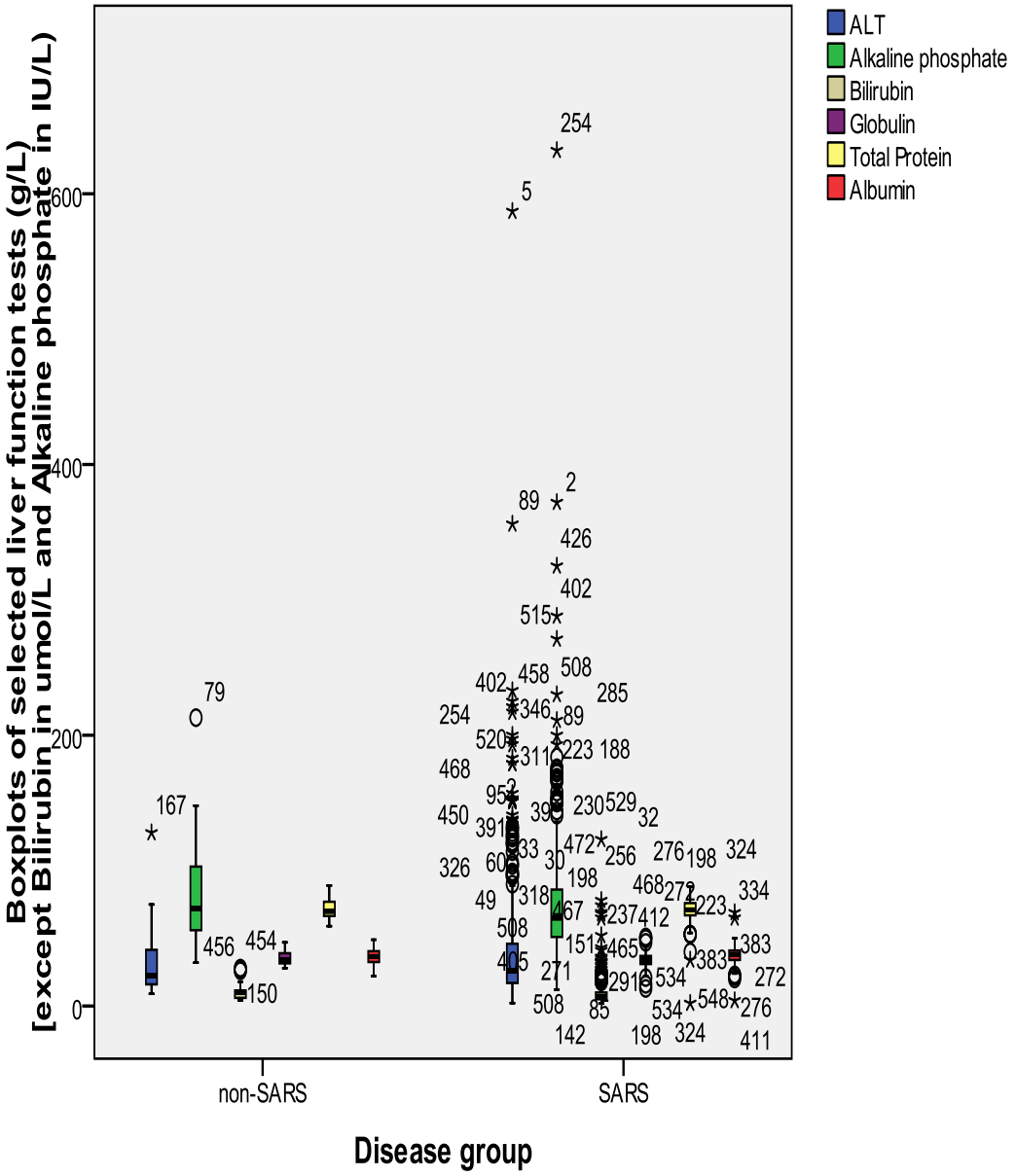

Figure 10 Distribution of selected liver function test results of disease groups for all subjects (N=525)

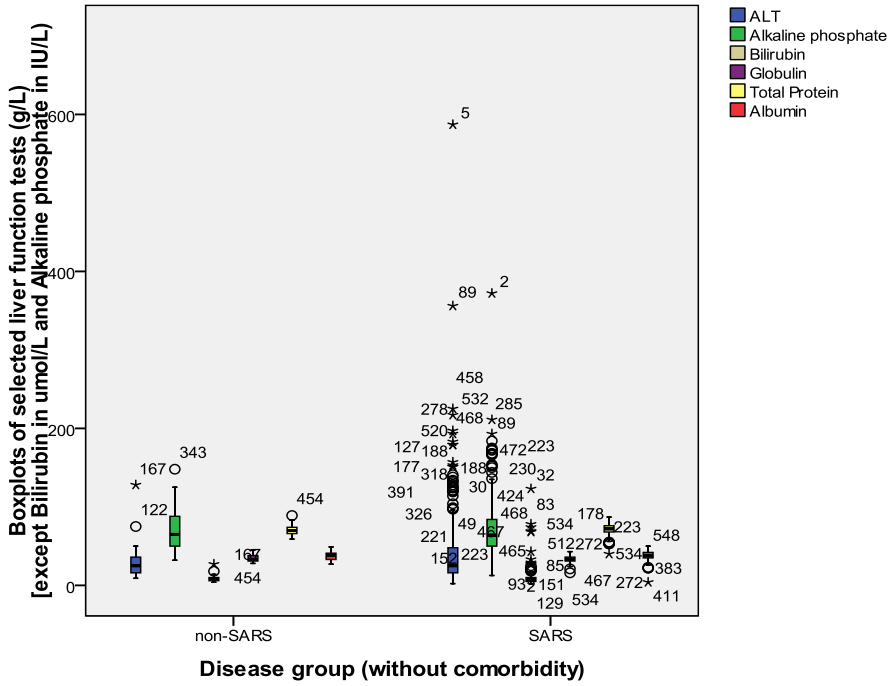

Figure10a Distribution of selected liver function test results of disease groups for those without comorbidity (N=354)

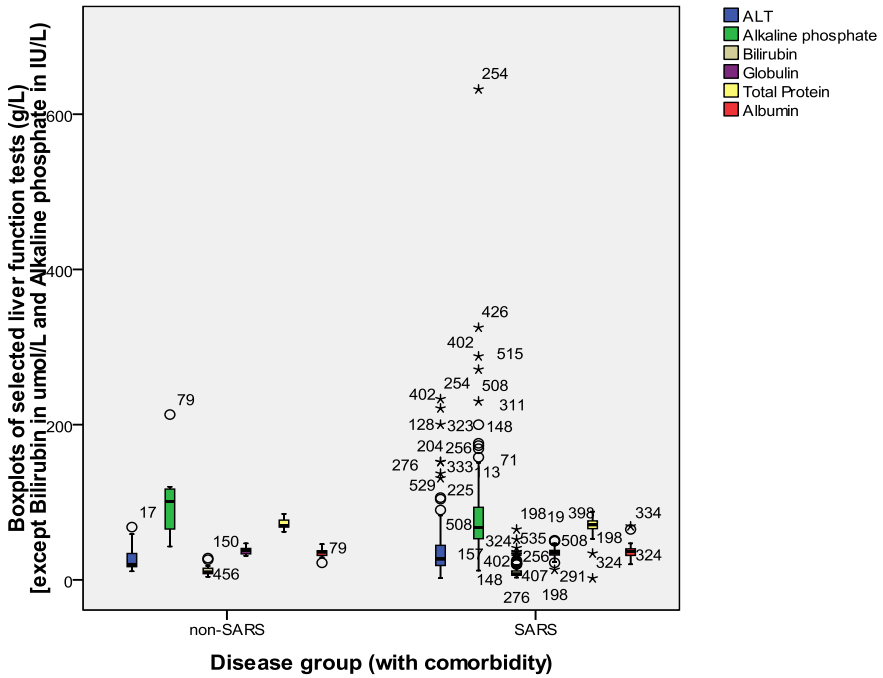

Figure 10b Distribution of selected liver function test results of disease groups for those with comorbidity (N=171)

6.4.5.3 Complete blood profiles

The results in the complete blood profile did not show any significant difference between the SARS and non-SARS patients except serum white blood cells count (WBC) ($t(532)=4.228$; $p<0.0001$) and that of monocyte counts ($p<0.005$).

Figure 11, 11a and 11b show the boxplots of selected complete blood profiles results by disease groups for all patients (N=538), for patients without comorbidity (N=362) and those with comorbidity (N=176). From the graphs, there are no significant difference in levels of selected complete blood profile tests between SARS and non-SARS patients even stratified for presence and absence of comorbidity.

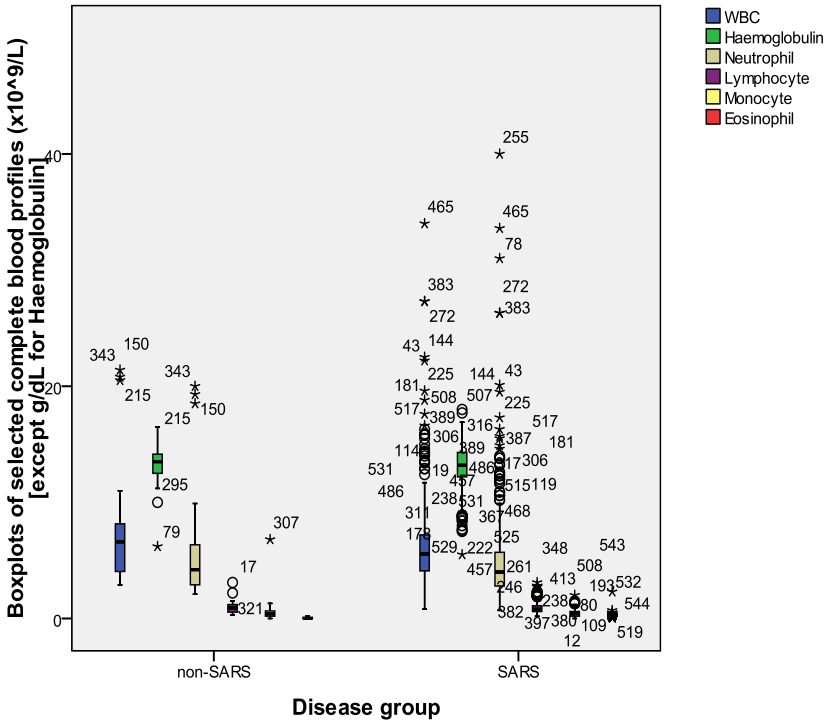

Figure 11 Distribution of selected complete blood profile results of disease groups for all subjects (N=538)

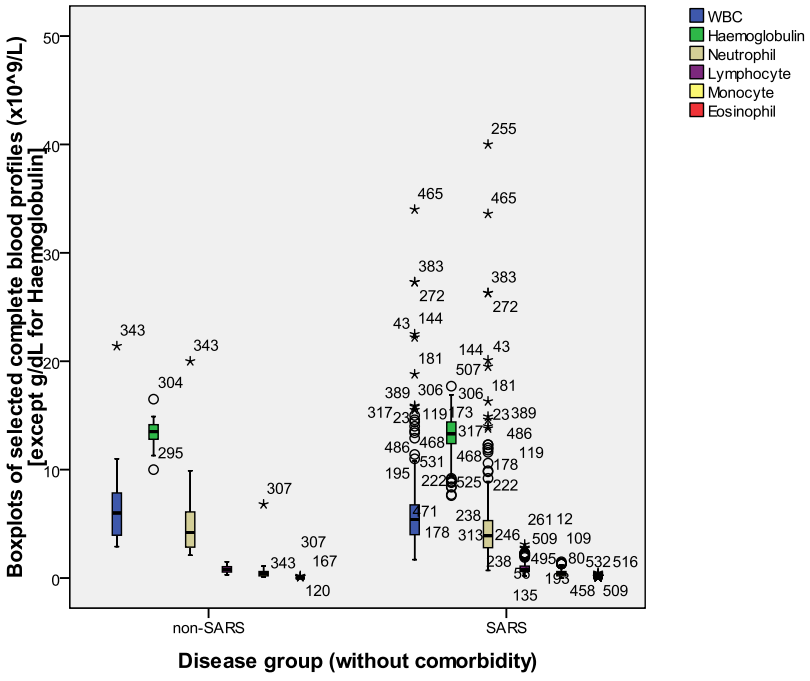

Figure 11a Distribution of selected complete blood profiles results of disease groups without comorbidity (N=362)

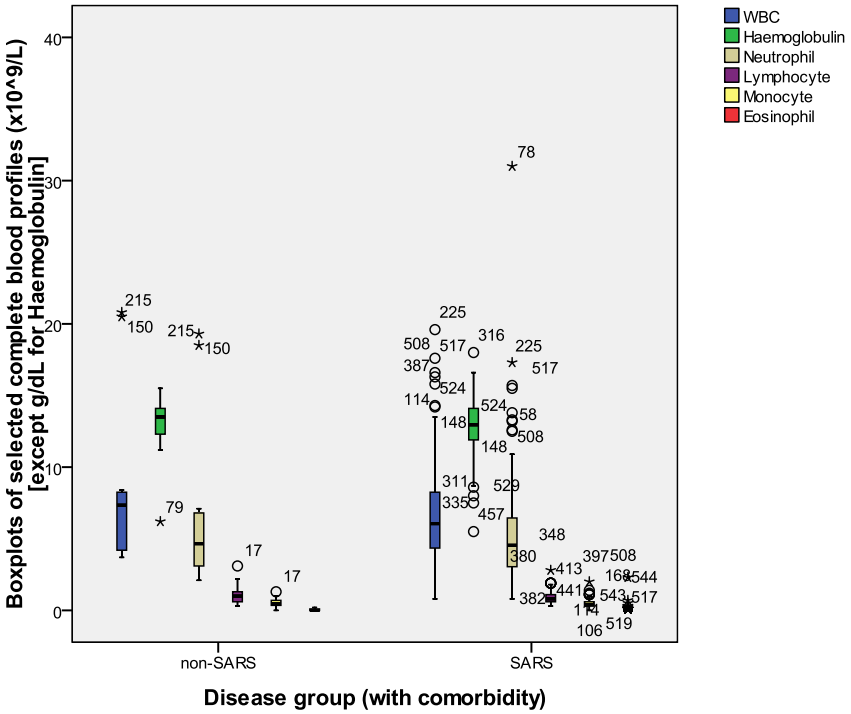

Figure 11b Distribution of selected complete blood profiles results of disease groups with comorbidity (N=176)

6.4.5.4 Clotting profiles

The results in the clotting profile did not show any significant difference between the SARS and non-SARS patients except Activated Partial Thromboplastin Time (APTT) ($t(23)=-2.846$; $p<0.01$). Figure 12, 12a and 12b show the boxplots of selected clotting profiles results by disease groups for all patients (N=477),

for patients without comorbidity (N=301) and those with comorbidity (N=176). From the graphs, there are no significant difference in levels of various clotting profiles between SARS and non-SARS patients even stratified for presence and absence of comorbidity.

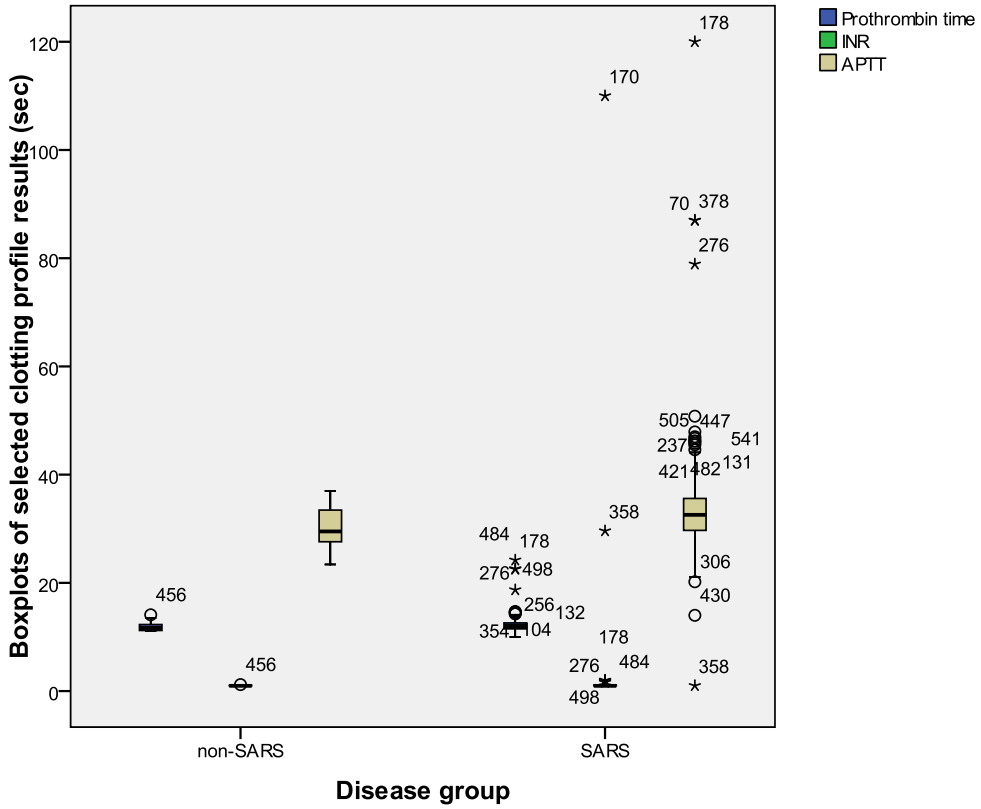

Figure 12 Distribution of selected clotting profile of disease groups for all patients (N=477)

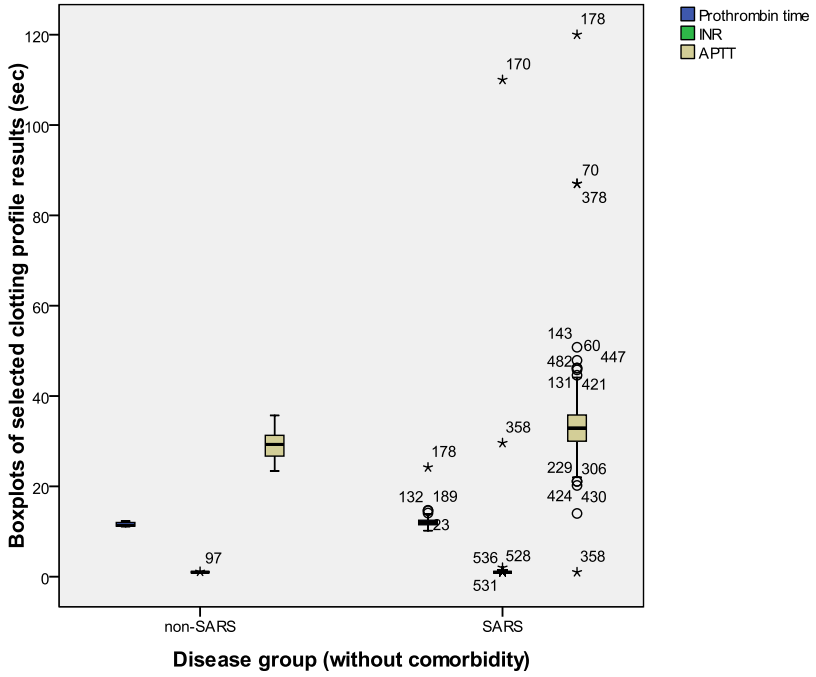

Figure12a Distribution of selected clotting profile results of disease groups for those without comorbidity (N=301)

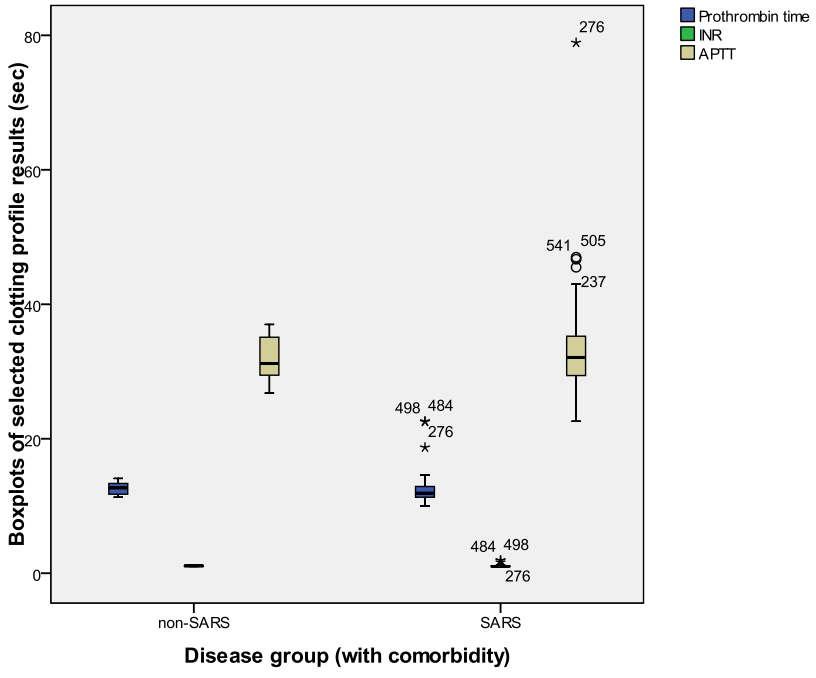

Figure12b Distribution of selected clotting profile results of disease groups for those with comorbidity (N=176)

6.4.5.5 Infection index

The results in the infection index showed significant difference in C reactive protein (CRP) ($t(271)=2.201$; $p<0.05$). Figure 13, 13a and 13b show the boxplots of selected infection index results by disease groups for all patients (N=549), for patients without comorbidity (N=377) and those with comorbidity (N=172). From the graphs, there are no significant difference in levels of various infection index between SARS and non-SARS patients even stratified for presence and absence of comorbidity.

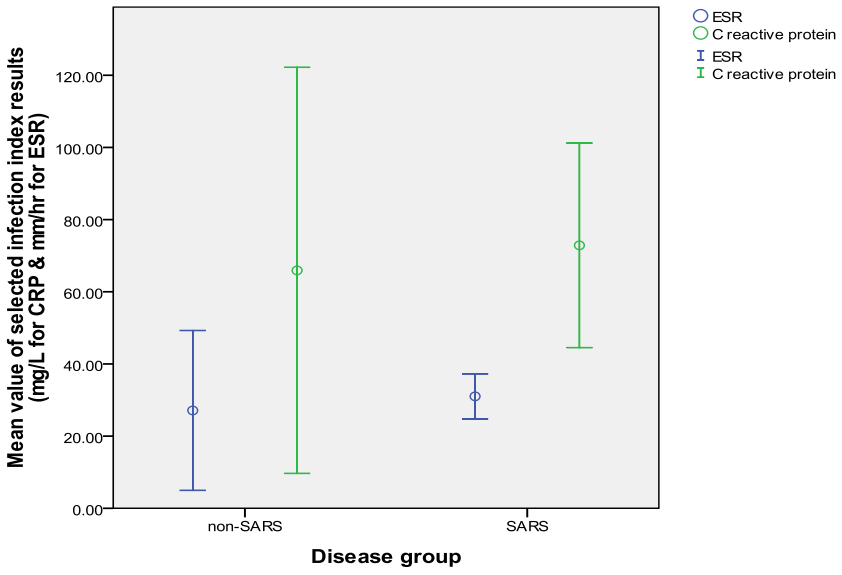

Figure 13 Distribution of selected infection index of disease groups (N=549)

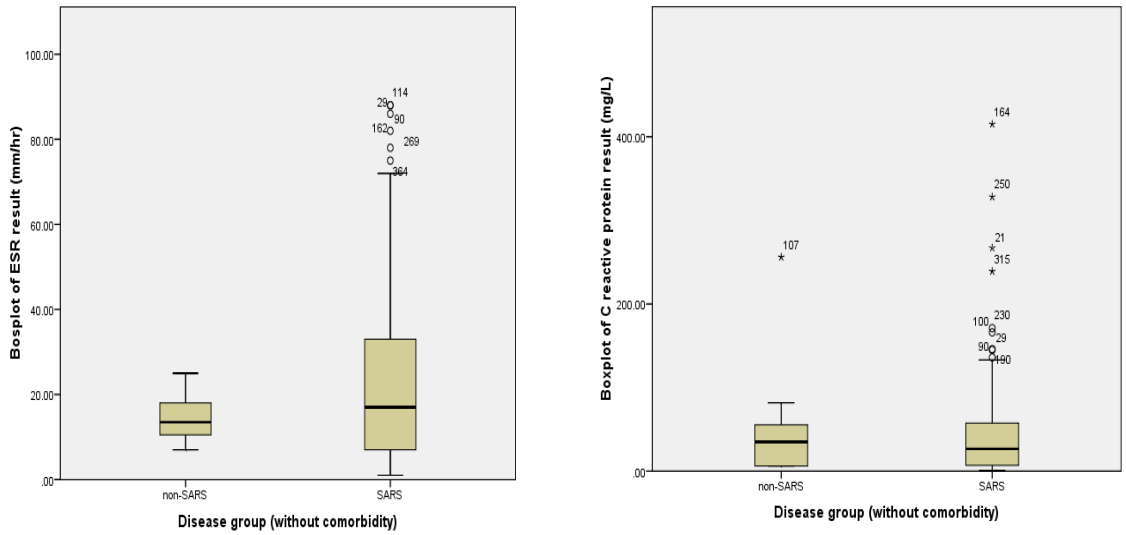

Figure13a Distribution of selected infection index results of disease groups for those without comorbidity (N=377)

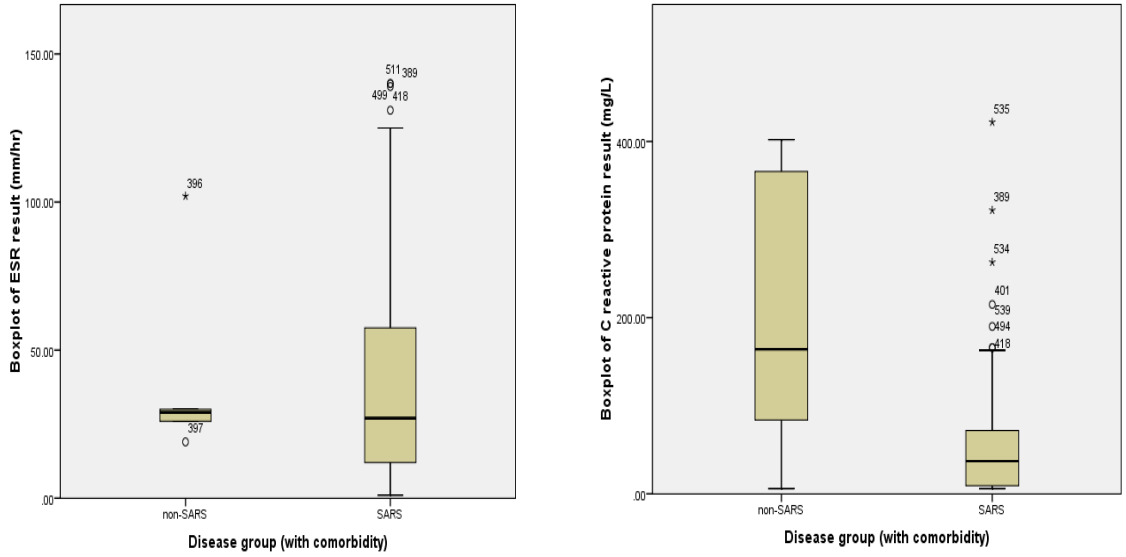

Figure13b Distribution of selected clotting profile results of disease groups for those with comorbidity (N=172)

6.4.5.6 Other blood results

Other blood results showed that the significant differences in those two groups of patients are Troponin I ($t(220)=2.1671$; $p<0.05$) and spot sugar ($t(44.86)=-2..967$; $p<0.05$). Figure 14, 14a and 14b show the boxplots of selected infection index results by disease groups for all patients (N=549), for patients without comorbidity (N=377) and those with comorbidity (N=172). From the graphs, there are no significant difference in levels of various infection index between SARS and non-SARS patients even stratified for presence and absence of comorbidity.

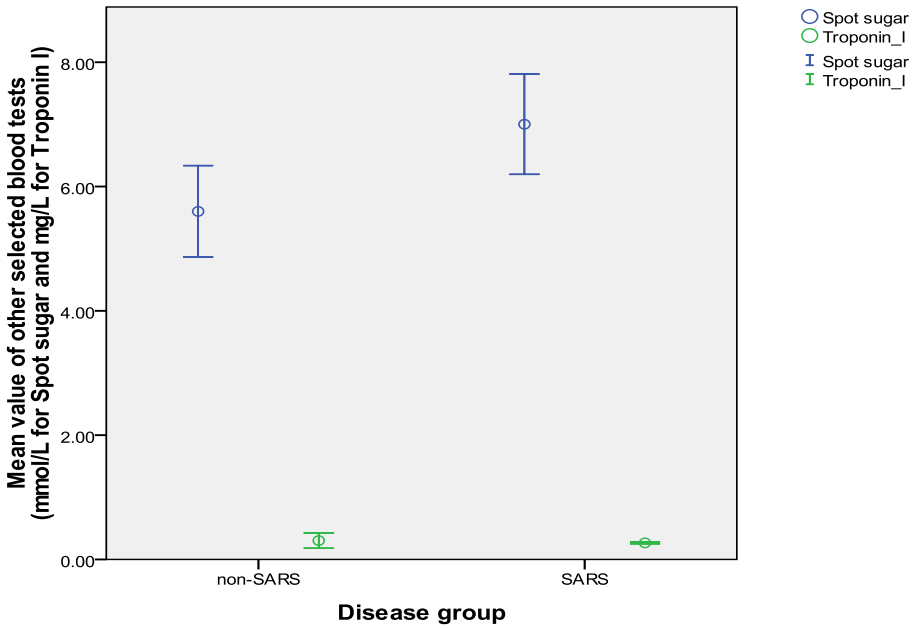

Figure 14 Distribution of selected blood results by disease groups

(N=549)

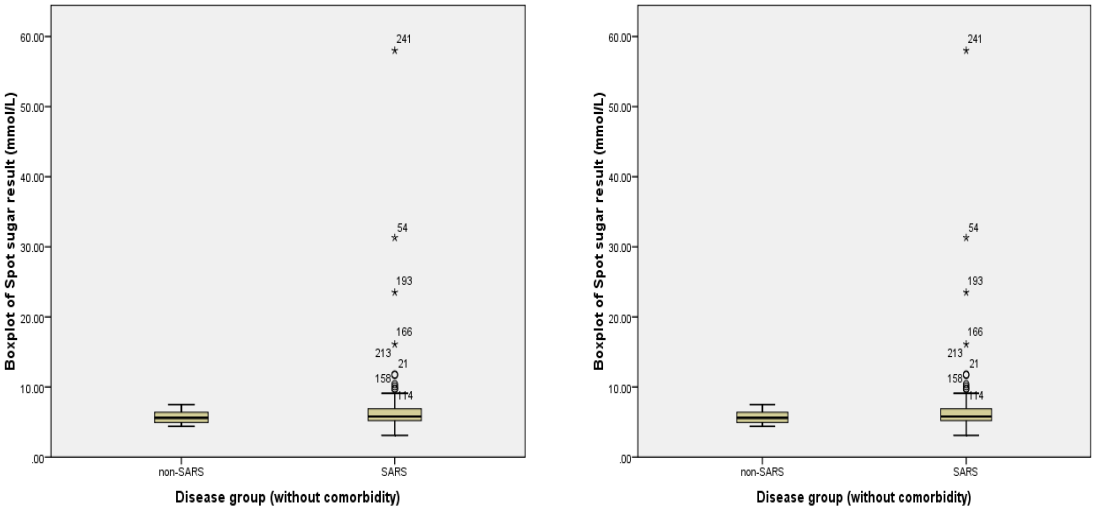

Figure14a Distribution of selected blood test results of disease groups for those without comorbidity (N=377)

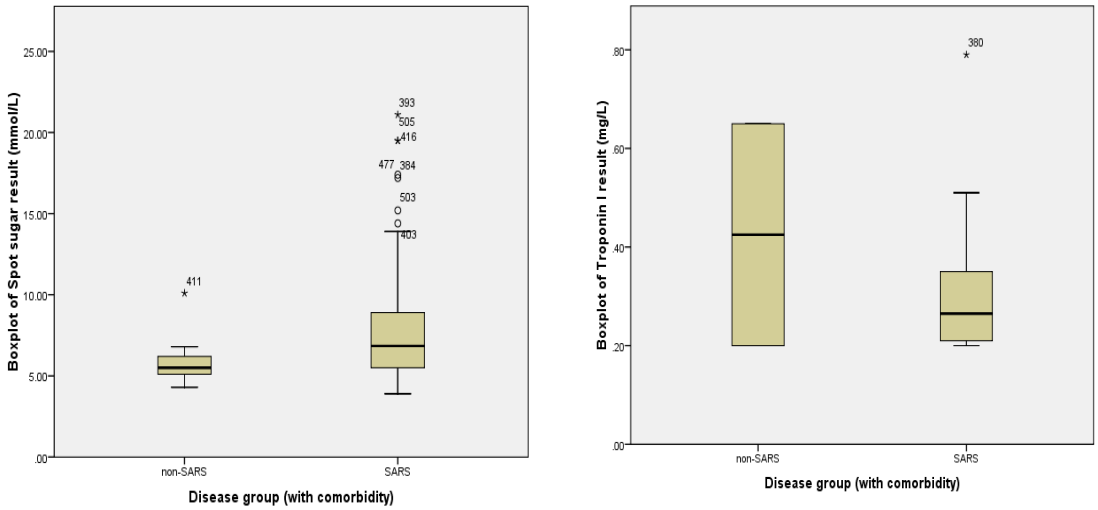

Figure14b Distribution of selected blood test results of disease groups for those with comorbidity (N=172)

6.5 Discussion

A data warehouse which can be used for further extract of relevant SARS data for analysis was successfully constructed in this stage of the study. Different mining techniques were used for analyzing specific environmental, intrinsic and extrinsic factors in order to illustrate that different critical predictors can be used to identify contributing factors under different circumstances.

6.5.1 Demographic profiles

The gender ratio of this study is similar to most of the studies concerning clinical presentations worldwide except for one study which was conducted by Choi et al. (2003) in which only 28% male subjects were included.

Among various countries, the average age of the patients with SARS is around 40 years of age. Hong Kong, however, had the greatest age range from neonates to 99 year-old. With this age span, local health care systems were challenged to provide diverse health care facilities, including adult care, pediatrics and a geriatric service. Chan, Miu, Tsui, Yee and Chan T. Y. (2004) compared the clinical features and outcomes in young and older adults (age $\geqslant$60) with SARS in Hong Kong. They confirmed that older patients were more likely to develop secondary nosocomial infection, be admitted to an intensive care unit and need mechanical ventilation. Various studies (Cao et al., 2003; Cheng &

Kwok, 2004; Dai & Chan, 2003a; Hui, Chan, Wu & Ng, 2004; Tse, Pun & Benzie, 2003; Tsuboi, Fukukawa, Niino, Ando & Shimokata, 2004; Wills & Morse, 2008) showed that older adults were more suitable to be screened with another set of criteria of signs and symptoms as they mostly presented milder and vaguer influenza-like symptoms. Some studies (Cheng & Kwok, 2004; Wills & Morse, 2008) postulated that the young adults experienced more intense signs and symptoms when compared with their older counterparts, due to the strong immune response for younger adults.

The WHO definition criteria for SARS have their limitations when applied to frail older adults, who tend to present with geriatric syndromes such as falls, confusion, incontinence and poor feeding (Dai & Chan, 2003b; Kong et al., 2003). It is important to avoid under- or over-diagnosis of SARS because these geriatric presentations are nonspecific. Older people are more vulnerable to the adverse drug reactions of ribavirin and high-dose corticosteroid which are commonly

used to treat SARS. The high mortality rates (50-75%) reported in older people with SARS may be attributed to late presentation, delayed diagnosis, comorbid conditions and complications from treatment (Chan, Chiu, Lam, Leung & Conwell, 2006).

Older patients diagnosed with SARS had more nonspecific symptoms and their prognoses were poor. RT-PCR was specifically useful in diagnosing SARS in older patients but the role of serological tests in individuals was limited (Chan, Miu, Tsui, Yee & Chan, 2004; Heung, Li, Mak & Chan, 2006; Hui, Chan, Wu & Ng, 2004).

6.5.2 Socio-economical profiles

Contact history is considered to be one of the major factors in the screening of SARS cases, but only a few studies investigated on whether the contact of SARS was caused by occupational health hazards (or not). Large numbers of health care professionals were victims of SARS in

Hong Kong, and it showed great variations in the ratios with most of the other studies. According to WHO, 21 percent out of the 8,447 worldwide cases occurred in health care workers. No definite mortality rates for health care workers but only 9.6 percent of overall mortality obtained (WHO, 2003a). There is only one case study conducted in Taiwan (Chiang, Shih, Su & Perng, 2004) that showed an exceptionally high percentage of patients (71%) involved being health care professionals, otherwise, the result obtained in this study is comparable to most of the other countries.

Traveling history is useful for surveillance processes and reporting systems among countries. It is proven by the case of the swine flu attacks in Mexico and her nearby countries that extensive surveillance systems were working simultaneously in cross country reporting systems (Trifonov, Khiabanian & Rabadan, 2009) and that helped to cease the propagation of the infectious disease to other countries.

6.5.3 Geographical profiles

Residents of Amoy Garden were a special group of patients as they presented very different signs and symptoms. Many studies (Chiu, Chim, & Lo, 2003; Chiu et al., 2004; Li, Duan, Yu & Wong, 2005; McKinney, Gong & Lewis, 2006; Yip, Chang, Yeung & Yu, 2007) specifically analyzed their modes of transmission and questioned whether the quarantine action actually aggregated the disease transmission.

Chiu et al. (2004) compared the incidence of watery diarrhoea in Taiwan with those in Amoy Gardens. They showed that the hospital-acquired SARS cases in Taiwan were infected mainly via respiratory routes and were less commonly associated with diarrhoea. Lower intestinal viral load, when the virus spreads, may contributive to lower rates of diarrhoea and lower positive rates obtained from rectal swab RT-PCR testing. Hence, geographical factors have to be considered as one of the risk factors for SARS as well as other airborne disease in the future.

Therefore, if we are facing another unknown diseases in the future (regardless it is airborne or droplet), we should also consider their geographic factors as those factors play crucial roles in identifying the potential cases timely.

6.5.4 Co-morbidity profiles

For many different kinds of illness, co-morbidity always increases the chances of having adverse outcomes or increases the readiness of contracting certain kinds of diseases. For example, a patient with diabetes mellitus would be more prone to have burn and scald injuries and if he suffered from a severe scald or was involved in a traffic accident, his recovery time or his wound-infection rate would then increase (McCampbell, Wasif, Rabbitts, Staiano-Coico, Yurt & Schwartz, 2002; Memmel, Kowal-Vern & Latenser, 2004). In this study, those with known underlying medical illnesses did not show significant association with the occurrence of being more prone to contract the disease.

There were relatively few studies in this area and most of them look at whether the presence of underlying diseases causes a higher case-fatality rate. Such a case was conducted by the Hospital Authority SARS Collaborative Group which examined 1312 laboratory-confirmed patients' case notes. After adjusting the odds ratios, they found that older in age, males, with elevated pulse rates and elevated neutrophil counts were predicted to have higher hospital mortality during the first 10-days of illness (Chan, Tsui & Wong, 2007). Similar findings have been found by Wang, J. T. et al. (2004) and Liu et al. (2004).

6.5.5 Laboratory result profiles

Lee et al. (2003) described the clinical and laboratory features of 138 cases of suspected SARS cases in Hong Kong. It is striking to discover that 44.8% of the patients had thrombocytopenia, 45% had elevated

levels of D-dimers and 42.8% had a prolonged activated partial-thromboplastin time. Such discovery suggests the presence of a form of disseminated intravascular coagulation or pulmonary-induced coagulation and fibrin polymerization with consumption of platelets and clotting factors (Jr. Taylor, Toh, Hoots, Wada & Levi, 2003). Elevated D-dimer levels have also been reported to have found in patients with acute lung injury and patients with acute respiratory distress syndrome (Wenzel et al., 2002).

However, some patients did not exhibit these typical clinical features as those stated by Li, Zhao, Chen & Zhou (2003). These patients were initially excluded, despite their close contact with confirmed SARS patients, because their symptoms could be considered as those of contracting a common cold. No specific diagnostic approaches were carried out when they were sick because the causative agent of SARS was not identified until March 2003. Their serum specimens tested positive for IgG against SARS-CoV by ELISA indicating that they were

infected with SARS-CoV. Mild SARS-CoV infection may not be easily clinically defined, and those patients may then potentially spread the disease if they were not isolated.

Wiwanitkit (2007) conducted a meta-analysis study to document the frequency of the lymphopenia in SARS. Overall, 637 SARS patients, selected from five available reports concerning the prevalence of lymphopenia in SARS among different populations, were analyzed retrospectively and 492 cases were presented with lymphopenia. The overall prevalence rate of lymphopenia in SARS is 77.2 %. Hence, lymphopenia as an important characteristic in patients with SARS, it was not an indicator that "SARS is a viral – induced lymphopenia disease" where no significant correlation is found between the population ethnicity and the prevalence rate ($p > 0.05$).

Hui, Chan, Wu & Ng (2004) reported a meta-analysis study that the common laboratory positive test results for SARS patients included

lymphopenia, thrombocytopenia, raised alanine transaminases, lactate dehydrogenase, and creatinine kinase. The constellation of compatible clinical and laboratory findings, together with certain characteristic radiological features and lack of clinical responses to a broad spectrum of antibiotics, should arouse suspicion of SARS. Measurement of serum RNA by the real-time reverse transcriptase-polymerase chain reaction technique had a detection rate of 75%–80% in the first week of the illness.

Majority of studies (Fan, Yieh, Peng, Lin, Wang & Chang, 2006; Leong et al., 2006; Liu et al., 2004; Tang et al., 2004; Wang, J.T. et al., 2004; Wilder-Smith, Earnest & Paton, 2004) described that people suffering from SARS would experience leucopenia, lymphopenia, thrombocytopenia and elevated CRP. However, the condition was not entirely true compared with the findings of this study. Many viral or bacterial infections would cause the similar responses, so how are we going to differentiate SARS from other viral or bacterial infections?

In addition, some patients did not exhibit these typical clinical features as stated in Li, Zhao, Chen & Zhou (2003). These patients were initially excluded despite their close contacts with confirmed SARS patients because their symptoms could be explained as a common cold and no specific diagnostic approaches were carried out when they were sick since the causative agent of SARS was not identified until March 2003. Their serum specimens were positive for IgG against SARS-CoV by ELISA. Those results strongly indicate that patients were infected with SARS-CoV, although their signs and symptoms did not meet the criteria for the SARS case definition. Mild SARS-CoV infection may not easily be defined clinically, and such patients may potentially spread the disease if they are not isolated.

6.6 Conclusion

With respect to the demographic and socio-economical profile of those who contracted SARS, there was a significant association between financial status and visit to Amoy Gardens, and also between financial status and Amoy Gardens residence. The expected association between occupation and underlying medical illness was not supported regarding the status of contracted SARS. The financial status did have some contributions to the status of contracted SARS. For laboratory results, serum white cells counts, infection index, Troponin I and spot sugar showed significant differences between SARS and non-SARS patients.

With the current available data from the data warehouse, we can evaluate the accuracy of existing prediction rules reported by other researchers in classifying patients into SARS-contracting group.

CHAPTER 7

STAGE 2 - EVALUATION OF EXISTING PREDICTION RULES

7.1 Introduction

Identifying patients who are tainted with infectious disease and starting the necessary precaution promptly before it starts propagating depends greatly on the accuracy of the diagnosis and the time taken for providing a diagnosis (including provisional diagnosis). There were some prediction rules developed by other researchers expecting to differentiate SARS patients from others by examining their presenting signs and symptoms as well as the laboratory results.

This chapter will first present the existing prediction rules reported by others and evaluate the precision of how these prediction rules can

identify patients who contracted SARS from others based on the database constructed in Stage One of this study.

7.2 Research objective

This stage mainly aims at testing whether the available data fit into the existing prediction rules developed by other researchers.

7.3 Research questions

The research questions of this stage are:

i. What is the predictive value quoted in different studies based on their own database?

ii. Do the existing prediction rules developed by other researchers have the same high quoted predictive value using our data?

iii. How will the predictive power changed if using the data from the data warehouse constructed in Stage One?

7.4 Methods

This stage incorporates the model testing method to evaluate how well the existing prediction rules can predict the status of whether a patient has contracted SARS using the data from database constructed in Stage One.

7.4.1 Procedures

Using the data warehouse prepared in Stage 1, the data were randomly split into two groups, around 80 percent (439 cases) were used to form a model (training data) and the rest (110 cases) were used to test the model (testing data). This ratio is the most common arbitral ratio used in data

mining for supervised learning (Dunham, 2003; Roiger & Geatz, 2003).

As discussed in Chapter 4, there were three prediction models identified which were carried out in the most similar situation or inclusion criteria as the current study. They were the Clinical Score and Symptom Score from Chen et al (2004); the Step I & Step 2 prediction from Leung et al (2004) and the Predictor model from Wang et al (2004). Each of these published studies identified under which circumstances to diagnose the patient as SARS positive or SARS negative by calculating a score for screening. The testing data were then fitted into each of these different prediction models and the subsequent prediction scores were derived according to their formulas and differentiated whether the patients were diagnosed as SARS positive and SARS negative accordingly. As described before, the researcher had used the same definition of each study as far as possible. These calculated results were then compared with the laboratory result (RT-PCR) to check whether the case was identified correctly as a SARS-positive one. Different predictive values

of various prediction models were computed accordingly.

7.4.2 Data management

Data were tested by fitting into various prediction rules with the help of SPSS (version 17), a software, which is now renamed as Predictive Analytics Soft Ware (PASW) to compare their predictive value, specificity and sensitivity.

7.5 Results

7.5.1 Profile difference between training and testing data

As mentioned, 80 percent of the data were randomly split into two groups: training data and testing data. The profile of the two groups was relatively comparable to each other. The details of the profile difference are shown in Table 5.

Table 5 Profile difference between training and testing data (N=549)

Factors		Training data (439)	Testing data (110)
Gender	Male : Female	40:60	50:50
Age	18-65 years old	88.8%	92.7%
	>65 years old	11.2%	7.3%
Smoking habit	Non-smoker	90%	92.7%
Drinking habit	Non-drinker	93.2%	91.8%
Financial status	Self financed	83.4%	77.3%
Occupation	Health care worker	17.1%	10.9%
	Housewife	19.1%	11.8%
	Retired	10.7%	9.1%
Geographical factors	Amoy residents	30.3%	41%
	Lower Ngau Tau Kwok Estate residents	1.1%	1.8%
Travel history	Yes	20%	19.1%
	Travel to China	18.2%	18.2%
Comorbidity status	With comorbidity	31.7%	33.6%
Disease status	Confirmed SARS	93.6%	97.3%

7.5.2 Data fitting

The testing data in this study were fitted into three existing clinical prediction rules. The raw scores of their testing result are presented as error bar as shown in Figure 15 below. Please note that the *Chan's prediction rule* is shown as *"Clinical score" and "Symptom score"* in

the Figure 15.

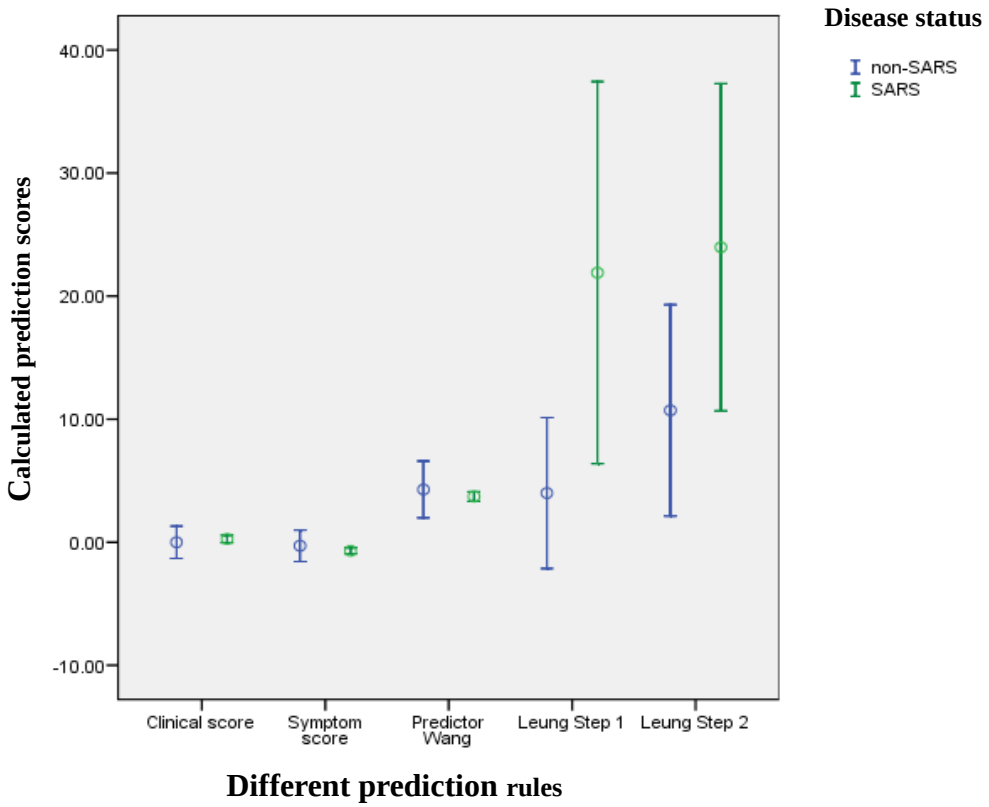

Figure 15 Raw predictive scores of existing prediction rules (N=110)

The distribution of the raw scores from various studies did not show any significant differentiations of the SARS cases and non-SARS cases. The sensitivity, specificity, positive predictive value and negative predictive value of these prediction rules for predicting SARS occurrence are shown in Table 6.

Table 6 Quoted and calculated values of sensitivity, specificity and predictive values for different prediction models for predicting the occurrence of SARS

	Chen et al (2004)		Wang et al. (2004)	Leung et al. (2004)	
	Clinical score	Symptom score	Predictor	Step 1	Step 2
Quoted sensitivity	100	100	100	99	95
Calculated Sensitivity	43.14	48.04	16.50	95.10	92.16
Quoted specificity	86.3	75.9	93	63	57
Calculated Specificity	57.14	42.86	66.67	0	42.86
Positive predictive value	93.62	92.45	89.47	93.27	95.92
Negative predictive value	6.45	5.36	4.44	0	27.27

According to the table above, we can summarize that the values for sensitivity and the specificity are relatively low and incoherent even among the studies, which means that the test results may not be able to identify the cases confidently. We are obliged to explore other prediction rules that can improve the sensitivity and the specificity.

7.6 Discussion

7.6.1 Profile difference between training and testing data

The training data had a relatively significant difference in the distribution of the occupation. The training data had a larger percentage for the health care worker and housewife than that of the testing data. The testing data also had more samples on Amoy residents. However, as the data was randomly split into the two groups, and the testing procedures for Stage 2 and that of Stage 4 were all the same. The effect would be even out. For the confounding factors such as age and gender, the use of randomization as for the case-control study did not apply here as this was a medical record review study.

7.6.2 Data fitting

The use of clinical prediction rules helps physicians to interpret clinical

information in order to estimate the probability of a diagnostic outcome and to classify patients according to the risk of a disease and the potential benefit from the therapy. The rules are always derived from studies involving thousands of patients and sophisticated mathematical analysis employed. It is expected that subjects under similar situations of exposure show similar results with a certain percentage of accuracy. However, when the testing data of this study are entered into various existing prediction rules, the sensitivity and the specificity of diagnosing SARS dropped rapidly, showing a wide margin from the author's quotation as shown in Table 6.

One may suspect that the variation in case definition during the initial stage and the post stage might cause the difference of the results. It must be admitted that with the gradual accumulation of knowledge during the SARS outbreak, the changes in case definitions during the epidemic inevitably caused confusion in the management of different individual cases. The author personally experienced a total different protocol

during the course of the outbreak in collecting specimen for investigation. All of these changes and errors must be avoided in subsequent infectious disease outbreaks, otherwise, further confusion and chaos will be experienced by more people/ it will again cause confusion and chaos to a society.

During the waiting period for the laboratory results, a specific area is needed to accommodate and isolate the patients. According to the differences observed between the quoted specificity and the calculated specificity from different studies in Table 6, around 30% of the patients experienced inappropriate isolation. If we apply any of the existing prediction rules for classifying future cases for SARS, this extra 30% wrong diagnose will require an extra 30% isolation facilities requirement. This will eventually pose a financial and resource burden to our health care system for future pandemics. For future (potential) pandemics, the health care system needs to provide more isolation facilities, which are sufficiently equipped to deal with the patients.

One may say that the increase in the quoted specificity and yet the increase in isolation facilities will not do any harm to a society anymore as SARS may not be returned. However, some Japanese scientists (Kawana, 2003; Taguchi, 2003; Teruya, 2003) predicted the return of the disease would be possible and thus, they were monitoring it closely during the winter in 2003. From their data, it could be seen that a second wave of infection occurred in Amoy Gardens. It was suspected that this outbreak was caused by contaminated air travelling from the sewage systems via chimneys to the atmospheric air. However, the index patient would be required to excrete a large amount of virus into the environment. A single viral discharge from the index patient has a finite window of infectiousness. Although some researchers showed that the SARS virus could live up to 4 days in diarrhoeal fluid and the survival time was estimated to be 24-48 hours on dry surfaces (WHO, 2003c), the path of the virus to release to an environment was difficult to determine.

SARS is still a threat to the general public because the route of

transmission varies in different situations and there is still no promising treatment identified so far, nor any vaccines for prevention (Hattori, 2005; Imai, Takahashi, Hasegawa, Lim & Koh, 2005; Nishiura et al., 2005). Hence, finding a reliable and efficient way to screen and identify potential patients who could contract this highly contagious disease along with the appropriate isolation strategies, if required, is one of the most essential steps to prevent further propagation of the disease.

7.7 Conclusions

It is expected that the prediction rules attain its purpose of proper identification of individuals infected with SARS instead of other diseases with high sensitivity and specificity. Using different data around the world might cause some slight differences in sensitivity and specificity. Nonetheless, the change should be small enough that it still retains its function in differentiating affected patients.

The poor sensitivity and specificity of quoted and calculated values obtained from the data of this study illustrated that another new prediction model must be developed in order to improve the prediction power.

CHAPTER 8

STAGE 3 - ATTRIBUTE IDENTIFICATION

8.1 Introduction

This chapter describes the procedures and results of identifying attributes of a new predictive model from the database constructed at Stage One using a data mining method. Further analysis was performed to prepare for different specific situations like geographical factors, gender differences, age groups and involvement of underlying medical illnesses.

8.2 Research objectives

This stage mainly aims at identifying the critical clinical variables by using the data mining technique.

8.3 Research questions

The research questions of this stage are:

i. Which are the critical clinical variables of SARS by data mining based on the comprehensive database?

ii. What are the associations between the clinical variables and the outcomes of SARS?

iii. Are gender differences found in the critical clinical variables?

iv. Are those critical clinical variables mined age specific?

v. Are the critical clinical variables mined illustrated any differences between Amoy and non-Amoy residents?

vi. Are the critical clinical variables mined denoted any differences between patients with or without morbidities history?

8.4 Methods

Some clinical predictors from various sources were identified and they

were then further analyzed by using the data mining method.

8.4.1 Procedures

In Stage One, the identification of the three studies is based on the situation/inclusion criteria carried out in the most similar situation/inclusion criteria as the current study, so that we can try to fit our data into their prediction rules. However, we found that the prediction rules from these three studies cannot serve its main function to differentiate affected patients from healthy subjects with poor sensitivity and specificity calculated (P.159). In order to develop another new model, we should start from the very beginning to decide which attributes (factors) should be included, hence a new and broader search of literature was conducted to look for any missing attributes (factors) in the previous studies.

A review of the literature in all available aspects of care and treatment

was carried out to explore the relevant clinical variables to be included in data-mining processes.

8.4.2 Data management

Data were then mined with the help of the software Clementine (version 12) and the upgraded version of PASW Modeler 13 to identify possible critical clinical variables. All associations and classification rules mining in this study started the testing with the support and confidence level of 0.8 and 0.9 respectively, which means it trials from very strong or high level towards very low level of support and confidence. The number of cluster started with 10 clusters in the clustering method.

8.4.3 Deliverables

Different associations among various clinical variables were found and critical clinical variables were identified.

8.5 Results

When one searches for “Severe acute respiratory syndrome” on the Internet with the Medline search engine, more than six thousand works can be found. This section focuses on the available literature and past research studies concerning clinical data, triage upon admission, and prediction rules for SARS patients.

8.5.1 General clinical predictors for adults

SARS is clinically characterized by fever, a dry cough, myalgia, dyspnoea, lymphopenia and abnormal chest radiograph results (Centers for Disease Control and Prevention, 2003; Donnelly et al., 2003; Lee et al., 2003; Peiris, Chu et al., 2003; T. Tsang & Lam, 2003). It is common that different clinical manifestations are noted for specific diseases. In different parts of the world, different signs and symptoms were experienced by patients who suffered from SARS during the outbreak.

Vu et al. (2004) documented the most common signs and symptoms in Vietnam. These are malaise (82.3%), lymphopenia (79.3%), fever>38^{o}C (79.0%), thrombocytopenia (40.3%), dry cough (22.6%), myalgia, headache, chills, chest pain (24.2%), dyspnoea (19.4%), diarrhoea, productive cough, vomiting, sore throat, rhinorrhea and leucopenia. While Poutanen et al. (2003) collected epidemiologic, clinical and diagnostic data from the first ten reported worldwide cases and stated that the most common presenting symptoms were fever, malaise, nonproductive cough and dysponea.

Booth et al. (2003) accounted for self-reported fever (99%) and documented elevated temperature (85%), nonproductive cough (69%), myalgia (49%), and dyspnoea (42%) as common signs and symptoms. Common laboratory features included elevated lactate dehydrogenase (87%), hypocalcemia (60%) and lymphopenia (54%) as the most common symptoms found in Canadian patients.

Choi et al. (2003) confirmed from 267 hospitalized patients that fever (99%), chills (74%), malaise (63%), and myalgia (50%) were the most common presenting symptoms for Hong Kong SARS patients while Donnelly et al. (2003) reported that fever (94%), influenza-like symptoms (72.3%), chills (65.3%), malaise (64.3%), loss of appetite (54.6%), and myalgia (50.8%) were the five most commonly reported signs and symptoms in Hong Kong.

On the contrary, Peiris et al. (2003) identified fever (100%), chills (74%) and non-productive cough (62%) as the most common clinical signs of the community outbreak of coronavirus-associated pneumonia whereas (Lew et al., 2003; Wang, J.T. et al., 2004) believed that only fever (89%) was the most common signs and symptoms for those critically ill patients with SARS. Interestingly, Liu et al., (2004) pointed out fever (98%), chills (68%), malaise (62%) and myalgia (57%) were the most common clinical features of severe acute respiratory syndrome in Taiwan whereas

Chiang et al (2004) denied Liu et al's findings and believed fever (100%), dyspnoea (86%), malaise (79%), diarrhoea (79%), and dizziness (79%) were the most common features instead.

Among all of the previous studies with similar conditions, 18 variables were identified as the most common clinical predictors for data mining which included fever, influenza-like symptoms, chills, rigor, cough (productive or non-productive), malaise, myalgia, sore throat, dyspnoea/shortness of breath (SOB), diarrhoea, headache, nausea/anorexia, vomiting, running nose, night sweat, abdominal pain, arthralgia, chest pain and dizziness. Table 7 summarizes the reported signs and symptoms of SARS from different researchers.

Table 7 Summary of clinical presentation among different studies

	Lee et al (2003)	Peiris et al 2003a)	Poutanen et al (2003)	Booth et al (2003)	Lew et al (2003)	Donnelly et al (2003)	Wang J.T.et al (2004)	Chiang, Shih, Su & Perng (2004)	Liu et al (2004)	Chang et al (2005)	Choi et al (2003)	Current study
Geographic area	HK	HK	Canada	HK	Singapore	HK	Taiwan	Taiwan	Taiwan	Taiwan	HK	HK
Subject number	138	50	10	144	199	1425	76	14	53	84	267	549
Signs and symptoms (%)												
Fever	100	100	100	99.3	89	94	100	100	98	94	99	73
Influenza-like						72.3						
Chills	73	74		27.8		65.4		57	68		74	61
Rigor	73					43.7	30.3				41	32
Cough (non productive)	57	62	100	69.4	39	50.4	47	71	42	4	43	43
Cough (Productive)	30			4.9		27.8		29	26		20	23
Myalgia	61	54	20	49.3	39	50.8	48.7	64	57	57	52	37
Malaise		50	70	31.2		64.3		79	62			55
Running nose	23	24		2.1							11	
Sore Throat	23	20	30	12.4		23.1	9.2	21	25	13	14	14
Dyspnoea			80	14.7			40.8	86	40			
SOB		20		41.7	14.5	30.6					20	21
Diarrhoea	20	10	50	23.6		27	31.6	79	36	35	15	16
Headache	56	20	30	35.4		50.1	18.4		45	45	33	34
Nausea	20			19.4		22.2	11.8	14	11		7	7
Vomitting						14	3.9	14	9			
Anorexia						54.6					23	25
Chest pain				10.4				14			7	8.5
Arthralgia				10.4								
Dizziness	43			4.2		30.7		79	32		17	16
Abdominal pain				3.5		12.6			9	5		
Rhinorrhea				2.1			2.6					
Coryza						24.6						
Night sweat						27.8						

8.5.1.1 Use of association rule mining

The data were inputted into the system for association rule mining in the red circle shown in Figure 16, however, no rules were successfully mined even the confidence level was put down to 40 %, which is considered to be low.

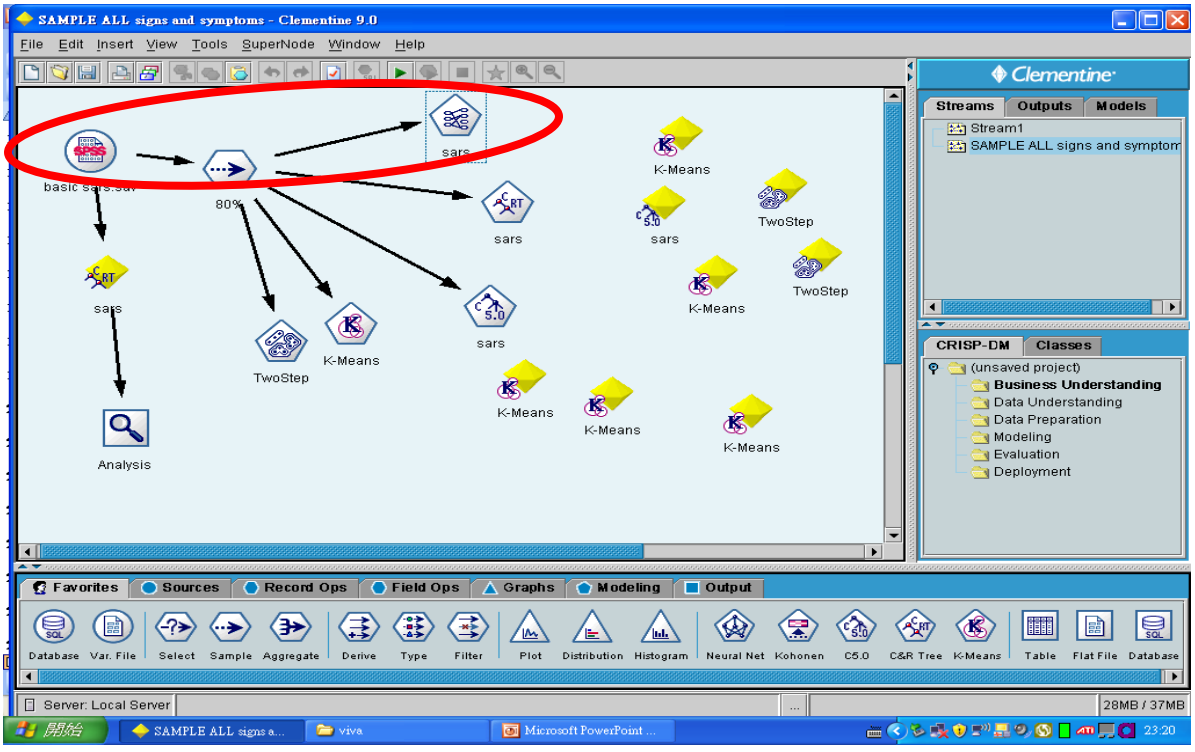

Figure 16 Example of association rule mining

8.5.1.2 Use of classification techniques

The data were then put for classification data mining in the red circle shown in Figure 17 and the model test procedure shown in green circle.

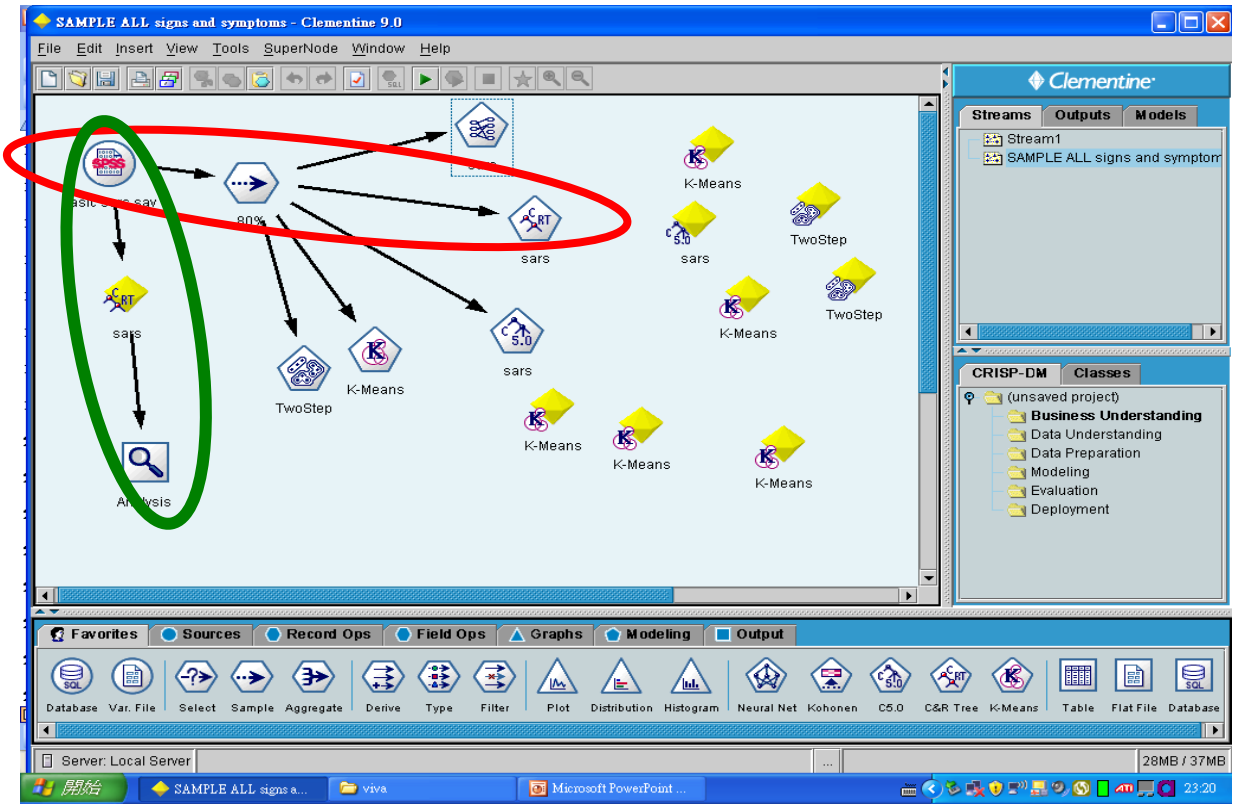

Figure 17 Example of classification rule mining

The software generated the decision tree induction with flow-chart like structure, one of the examples is shown here with

internal node denoting the test on the attribute; branch representing the outcome of the test and the leaf nodes for different class labels as shown in Figure 18.

The software identified sore throat, headache, vomit, sputum, chest pain, malaise and dizziness as the critical clinical predictors for the occurrence of SARS.

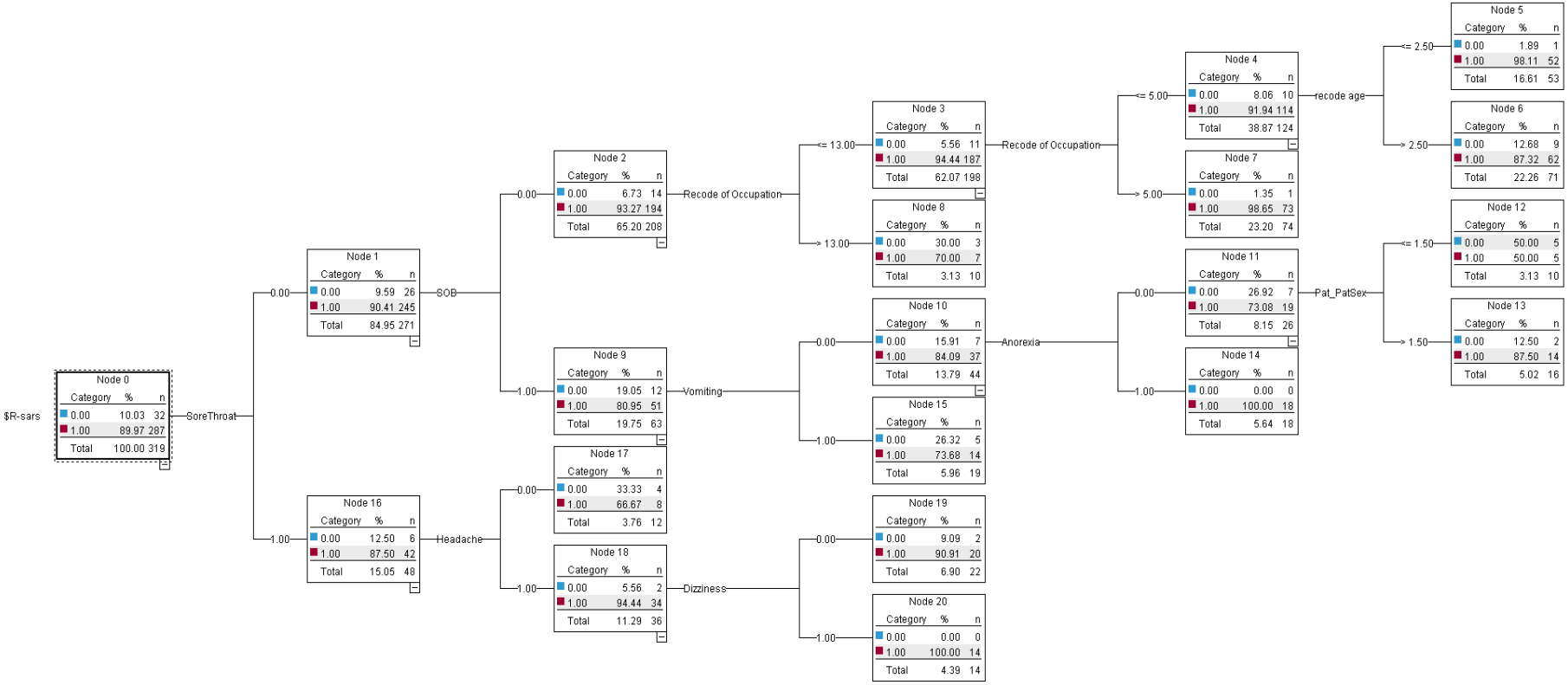

Figure 18 Example of the results mined by decision tree induction

8.5.1.3 Use of clustering techniques

The data was further tested with the clustering technique to check whether they can form different clusters sharing similar characteristics. The programme was started with the assigned minimum of 3 clusters and further moved up to more finite ones, i.e., moved up to more clusters.

Figure 19 and Figure 20 are the graphical result displays generated by the software after the clustering process. Clustering is an unsupervised machine learning method, it only requires the operator to set how many clusters he/she wants, then the software will fits the data in. For example if the operator fixed the software to separate the data into 3 clusters only (as shown in Figure 19). We can see that the system had used the K Means method to separate the data into 3 clusters. Then it will display the frequency distribution of subjects with each specific attributes

(clinical symptoms) among each clusters as bar charts. For example, in cluster 1 of attribute “Anorexia”, you can consider the blue bar as “anorexia positive” which is significantly higher than that of the orange bar, which considered as “anorexia negative”. But when you take a look in cluster 3 of the same attribute, you can see that the two coloured bars are more or less with the same height, that is you cannot easily identified. So for “Anorexia” this attribute, Cluster 1 can be considered as “Anorexia positive patients”, but no other clusters can show “Anorexia negative patients”. But if we considered “Headache” this attribute, we can see that the length of the two bars can identifies cluster 1 as “Headache positive patients” while cluster 3 as “Headache negative patients”. When we view this graph as a whole, we can see that the software cannot differentiate clearly the data in only 3 clusters.

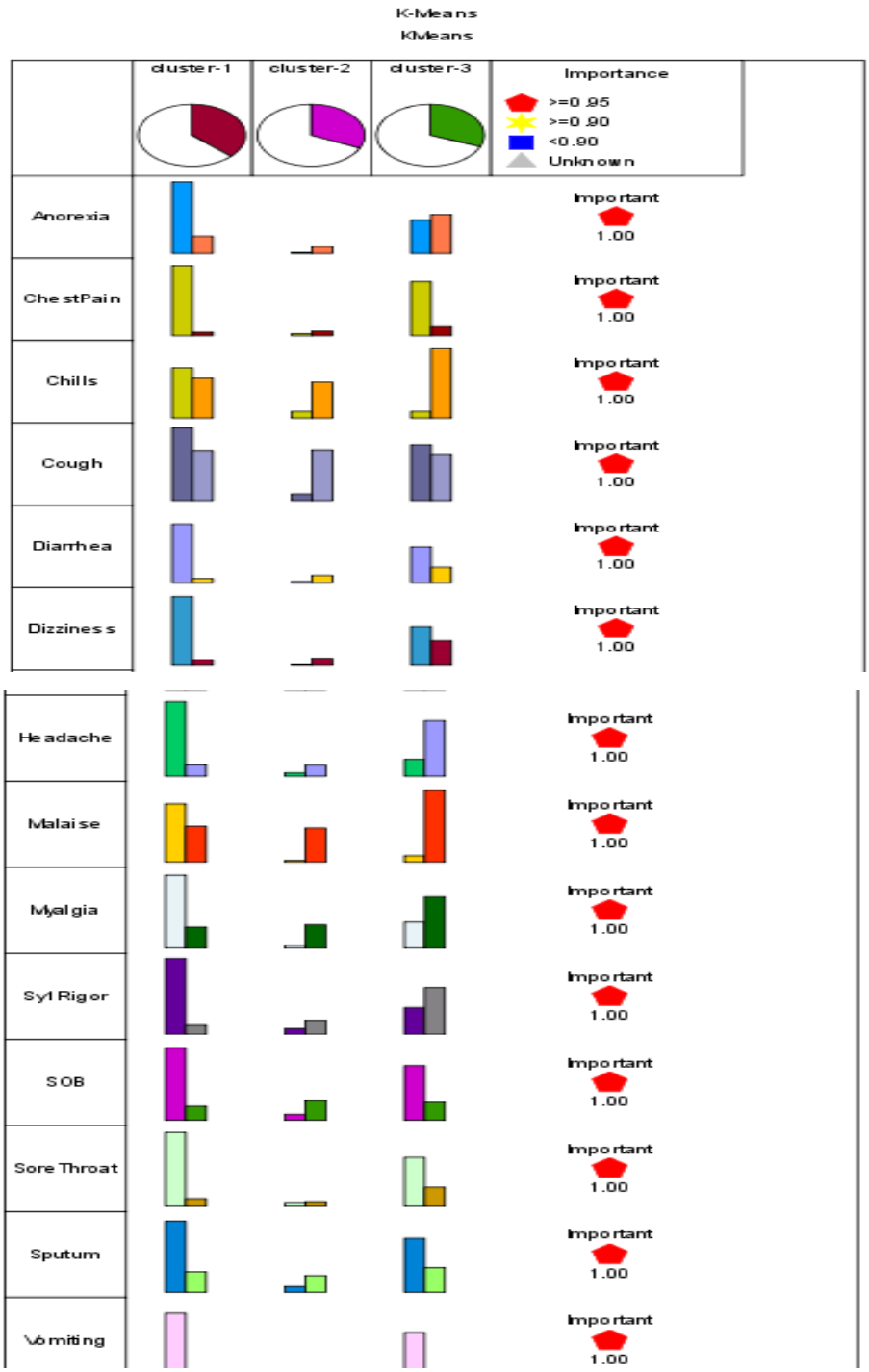

Figure 19 Results showing the clinical predictors with different significances (number of clusters =3)

From the graphical result as shown in Figure 20, we can see on the left side that the software divides the data into three clusters and the distribution of subjects with specific clinical symptoms within each cluster displayed as bar charts.

Taking “Chills” as an example, the subjects with chills can be predominantly picked up in cluster 2 but not in other clusters, and the importance is 1.0 that is very significant result. Hence, “Chills” can be considered as one of the critical clinical predictors. Similarly, headache, malaise, myalgia can also be considered as critical predictors.

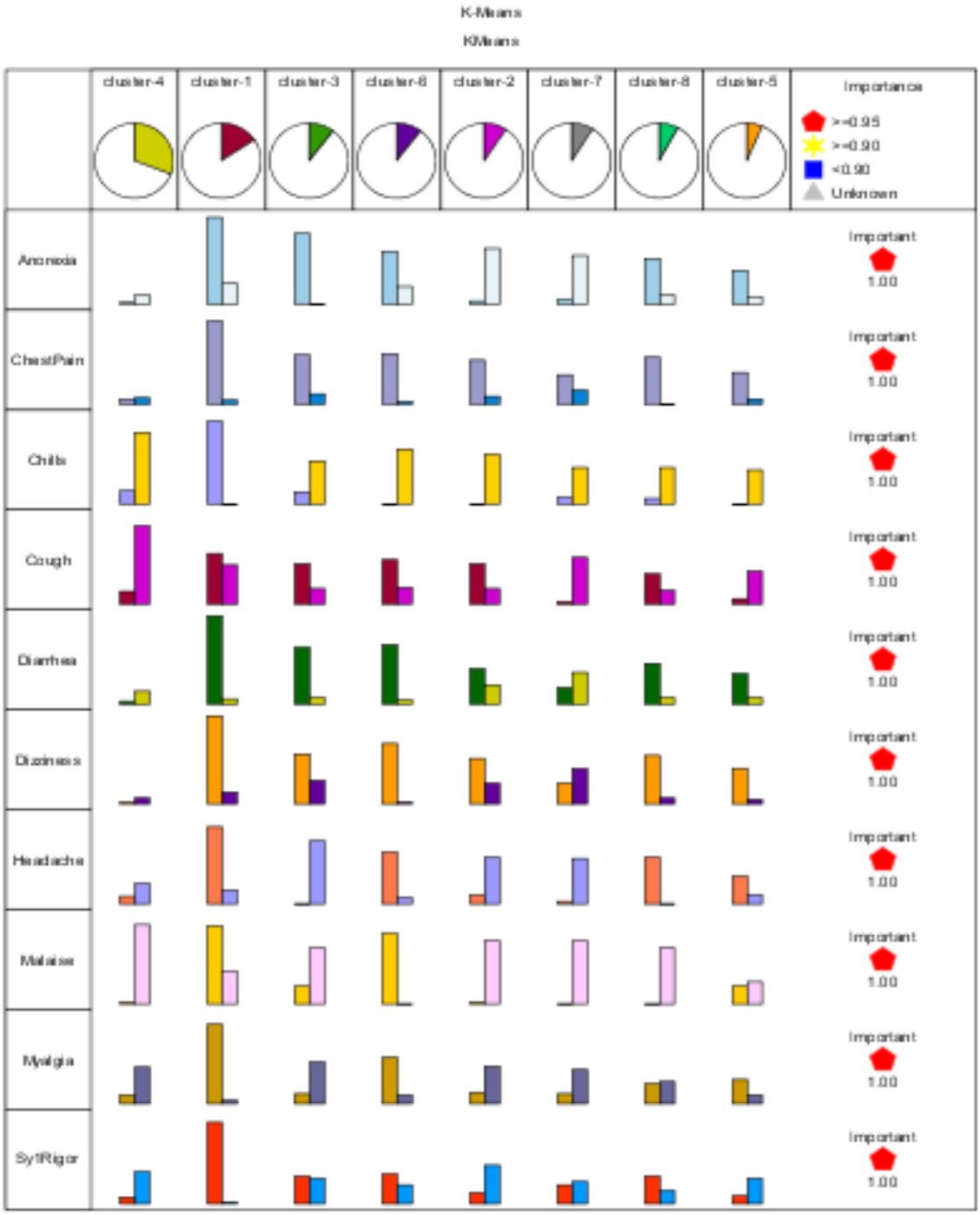

Figure 20 Results showing clinical predictors with different significance shown (number of clusters =8)

Two other steps of the clustering method were also incorporated in which the programme would automatically estimate the optimal number of clusters for the training data as shown in Figure 21. Table shows the summary of the critical clinical predictors mined in various conditions.

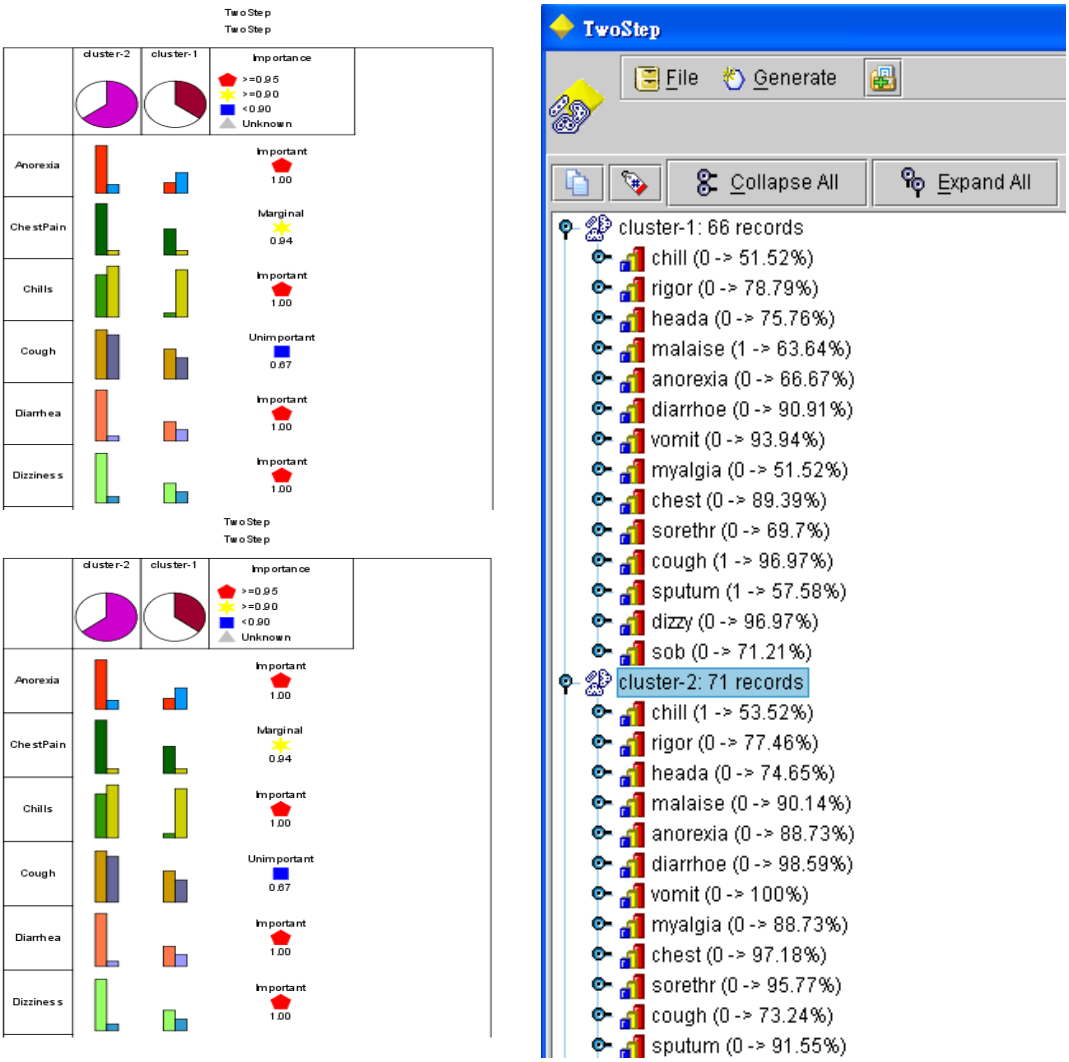

Figure 21 Results showing two-step clustering technique

Table 8 Summary of significant clinical predictors by different data mining techniques															
Dataset\ variables		*anore*	*chest*	*chill*	*cough*	*diarr*	*dizzy*	*heada*	*malaise*	*myalgia*	*rigor*	*sob*	*soreth*	*sputum*	*vomit*
C&RT*			√				√	√	√				√	√	√
Clustering (no.)															
K means(8)	1 (7)														
	2 (45)	√		√				√	√	√	√				
	3 (70)			√				√	√	√					
	4 (148)	√		√	√	√	√	√	√	√	√	√	√	√	√
	5 (31)			√	√				√		√			√	
	6 (50)			√											
	7 (43)	√		√	√	√	√	√	√	√	√	√	√	√	
	8 (74)			√					√	√					
(6)	1 (134)	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	2 (39)	√			√				√					√	
	3 (69)			√	√				√	√					
	4 (100)														
	5 (47)	√		√	√	√	√	√	√	√	√	√		√	
	6 (88)			√				√	√	√	√				

* C&RT denotes Classification and Regression Tree

Table 8 Summary of significant clinical predictors by different data mining techniques (cont'd)

Dataset\ variables		*anore*	*chest*	*chill*	*cough*	*diarr*	*dizzy*	*heada*	*malaise*	*myalgia*	*rigor*	*sob*	*soreth*	*sputum*	*vomit*
(5)	1 (82)				√				√					√	
	2 (63)	√		√	√	√	√	√	√	√	√	√			
	3 (139)	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	4 (103)			√											
	5 (96)			√				√	√	√	√				
(4)	1 (135)				√										
	2 (92)	√		√				√	√	√	√				
	3 (105)			√				√	√	√					
	4 (145)	√	√	√	√	√	√	√	√	√	√	√	√	√	√
(3)	1 (176)														
	2 (152)	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	3 (148)	√		√				√	√	√	√				
2 steps (10…>4)	1 (66)				√				√				√	√	
	2 (71)			√											
	3 (73)			√				√	√	√	√				
	4 (51)	√		√	√	√	√	√	√	√	√				
3	1 (100)														
	2 (84)			√	√				√					√	
		√		√				√	√	√	√				

8.5.2 Different clinical predictors for different situations

Further identification of the contributing factors for prediction under different circumstances was performed by creating different data cubes based on different attributes. A few examples were shown below for illustration.

8.5.2.1 Clinical predictors for geographical factors

A data cube for Outcome upon discharge (whether it is a SARS case) versus District_key (Residential area) versus Signs and symptoms was constructed and presented for mining any special prediction rules for geographical differences.

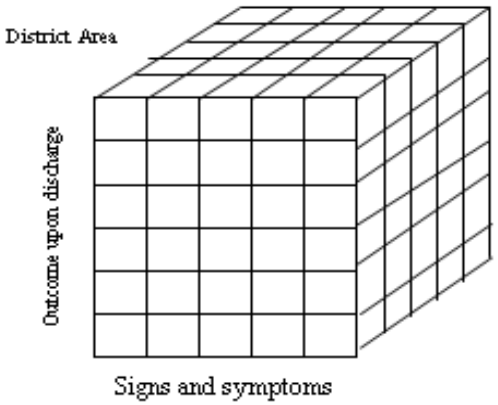

Figure 22 Data cube constructed for mining geographical difference predictions

8.5.2.1.1 Use of association rule mining

No rules successfully mined for identifying the critical clinical predictors for patients residing in Amoy Garden and non-Amoy Garden even the confidence level was put down to 40 %.

8.5.2.1.2 Use of classification

The software identified malaise as the critical predictive factor for Amoy Garden residents and non-Amoy Garden residents.

8.5.2.1.3 Use of clustering

The software identified chills and malaise as the critical predictive factors for Amoy Garden residents and chills, cough and malaise for non-Amoy Garden residents.

8.5.2.2 Clinical predictors for different genders

1. Outcome upon discharge versus Gender versus Signs and symptoms presented

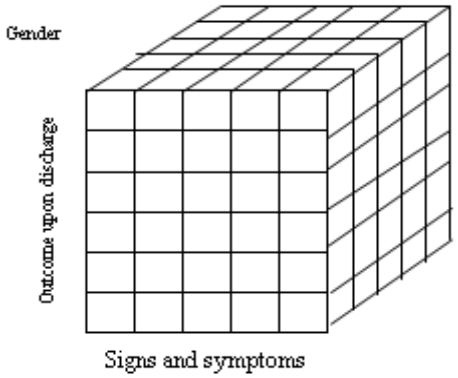

Figure 23 Data cube constructed for mining gender difference predictions

8.5.2.2.1 Use of association rule mining

No rules successfully mined even the confidence level was put down to 40 %.

8.5.2.2.2 Use of classification

Sore throat, fever and cough were identified by the software as the critical predictive factors for male patients. The same findings were not indicated among female patients.

8.5.2.2.3 Use of clustering

The software identified chills and malaise as the critical predictive factors for both males and females.

8.5.2.3 *Clinical predictors for young adults versus elderly patients*

Hence, another data cube for Outcome upon discharge versus Age versus Signs and symptoms was constructed and presented

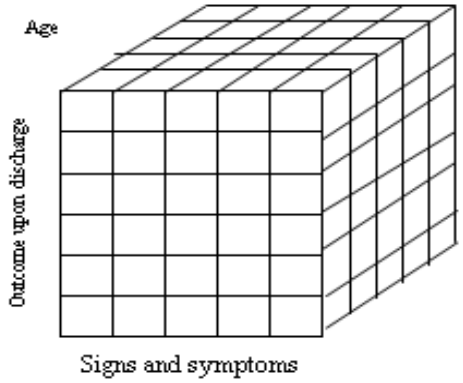

Figure 24 Data cube constructed for mining age difference predictions

8.5.2.3.1 *Use of association rule mining*

No rules successfully mined even with the confidence level being put down to 40 %.

8.5.2.3.2 *Use of classification*

This time sore throat, dizziness and shortness of breath were identified as the critical predictive factors for older adults whereas no specific clinical predictors were identified for

young adults.

8.5.2.3.3 Use of clustering

The software identified chest pain, sore throat and dizziness as the critical predictive factors for older adults but no rules mined for younger adults.

8.5.2.4 Clinical predictors for different co-morbidities

Outcome upon discharge versus Medical history versus Signs and symptoms presented

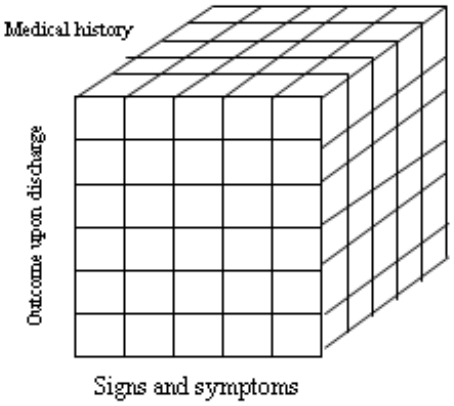

Figure 25 Data cube constructed for mining comorbidity difference predictions

8.5.2.4.1 Use of association rule mining

No rules successfully mined even the confidence level was put down to 40 %.

8.5.2.4.2 Use of classification

Fever, shortness of breath and chills were identified for patients without underlying medical illness.

8.5.2.4.3 Use of clustering

Shortness of breath, sore throat and vomiting were critical predictive factors identified for patients without underlying medical illness while chest pain and shortness of breath were predictors for those who had underlying medical illness.

8.6 Discussion

8.6.1 General clinical predictors for adults

The common clinical presentation of SARS patients in the database constructed at Stage One were fever (73%), chills (61%) and malaise (55%) which did not wholly congruent with those found in other research findings. The percentage of fever (73%) was not as high as it was reported in other studies, which were about 90%. However, fever is usually characterized as body's immune responses towards bacterial or viral infection. It is not

specific enough to work as critical clinical predictors. Senanayake (2006) stated fever varies according to individuals, time of day and anatomical site. And its occurrence is prevailing. The governing authority uses "fever" as the critical screening variables, the definition must be clearly stated as in pandemic situations because case definitions will be used by laymen to monitor themselves for signs of illnesses. In addition, many authorities nowadays use remote-sensing infrared thermographic camera (Chan, Cheung, Lauder, Kumana & Lauder, 2004; Chiu et al., 2005; Ng, 2005) as screening device in country borders for surveillance. In that case, the definition of fever would be different as skin temperature of fever patients may not be as high as that of oral temperatures. While the thermo-scanner is valuable in screening great number of people, the tool serves to identify those individuals with warm skin temperature. The results from this approach create further temperature deviation from the primary screening tool which measures the oral temperature of a feverish person, which considered to be a reliable dimension of infection.

The other common symptoms observed from the database were chills and malaise. Chills, which is a shaking body movement often occurs during high fever, usually considered as the consequences of fever. Chills will

subside gradually as the illness progressed (Lee et al., 2003). Malaise is a subjective feeling of general discomfort and uneasiness (Hornby, 2005) and is notably experienced when an individual is having infection or other diseases. All these factors were common signs and symptoms that presented in other studies as well (Leong et al., 2006; Leung et al., 2008; Sampathkumar, Temesgen, Smith & Thompson, 2003).

When all the potential critical clinical predictors were inputted into the software for data mining, headache and malaise were substantially identified by both classification and clustering technique while other predictors including sore throat, vomit, sputum, chest pain, dizziness, chills and myalgia were considered to be critical by either method.

8.6.2 Different clinical predictors for different situations

The technique for identifying the critical predictors for adults in various specific situations had been used repeatedly by other researchers to compare clinical presentations between SARS and non-SARS patients.

8.6.2.1 Clinical predictors for geographical factors

Malaise is identified as the critical clinical predictors for differentiating whether the patient had contracted SARS by both classification and clustering methods. If clustering method is used alone, chill is also considered to be one of the critical clinical predictors. This result is different from what other researchers postulated where diarrhoea and nausea should rank the highest probability in stratifying SARS and non-SARS patients (McKinney, Gong & Lewis, 2006; Yip, Chang, Yeung & Yu, 2007).

Instead, Li, Duan, Yu, & Wong (2005) and McKinney, Gong & Lewis (2006) made suggestion in relation to the reason for the special clinical presentation of SARS patients who was Amoy Garden residents. The researchers suggested that the dryness of the pipelines and sewage system failed to retain the germs but helped promoting a speedy spread of disease in the same building block. Some other investigators Ng (2003) believed that there might have animal vectors (rats and rodents) that cause the wide transmission of disease within Amoy Garden. To date, no confirmative/conclusive findings have been announced (Chiu, Chim,

& Lo, 2003; Li, Duan, Yu & Wong, 2005; McKinney, Gong, & Lewis, 2006).

8.6.2.2 Clinical predictors for different gender

Many studies investigated the relationship between diseases with the occurrence of different gender. However, the critical predictors mined cannot well delineate the gender difference using clustering although some researchers (Liang, et al, 2003; Lam, et al, 2004; Tsuboi, et al, 2004) believed that gender role influence where men and women spend their time, and the infectious agents they come into contact can be possible factors influencing the outcome. WHO (2007) had concluded that only pregnant women had greater concern in transmission and control the spread of disease in the future.

8.6.2.3 Clinical predictors for young adults versus elderly

The critical predictors mined for age specific condition do have some differences which need further investigation. Dizziness and

sore throat are the identified predictors by both classification and clustering methods. Shortness of breath is found to be one of the critical predictors for the older adults while chest pain is considered to be significantly helpful in identifying SARS case among young adults.

Shortness of breath is relatively common among older people especially for those who had underlying respiratory or cardiac problems. However, it is sometimes difficult to delineate which case is caused by SARS and which one is related to the underlying medical illness if a large number of patients are to be screened inside the Emergency Department (Goel, Gupta, Singh & Lenka, 2007).

Studies had suggested the discrepancy of presenting signs and symptoms from young versus older adults was due to their fast and healthy immune system of the young adults which cause the swift disease progression and the immanent response (Chan, Miu, Tsui, Yee & Chan, 2004).

8.6.2.4 Clinical predictors for different co-morbidities

Many studies investigated the relationship between mortality rates with the occurrence of an underlying disease. For examples, Chan et al (2006) stated that both age and existence of other diseases before SARS were significantly correlated with prognosis while Wang J. T. et al. (2004) believed that underlying disease and initial CRP level were predictive of death in SARS patients. Chen et al. (2005) suggested that diabetes mellitus, ischaemic heart disease and congestive heart problem would cause the rapidly fatal outcome and severe complications.

8.6.3 Different data mining techniques

In association rules mining, it is mainly to deal with whose consequent may contain multiple conditions and attribute relationships. An output attribute is that one association rule can be an input attribute in another rule. It is somehow an affinity analysis. It will be the best choice when attributes are allowed to play multiple roles in the data mining process. Hence, if we

changed to whether the patient has certain kinds of condition, such as hyponatremia or hypernatremia, there will be a better chance to get a positive result.

As mentioned in Section 4.5, classification mining is mainly to assign a class to find previously unseen records as accurately as possible. Hence, it is most commonly recruited as a decision making technique. The rules mined are actually originated from a supervised training of the available training data and tested ~~it~~ against the testing data.

8.7 Conclusion

Chills, sore throat and malaise were reckoned as the most critical clinical predictors for the identification of SARS in general. Shortness of breath and headache are to be taken into account for age specific group (condition) such as the elderly.

CHAPTER 9

STAGE 4 - MODEL FITTING AND TESTING

9.1 Introduction

This chapter evaluates the effectiveness of the newly identified critical clinical predictors in determining whether a patient contracted SARS (or not) by using the same testing data of Stage Two. A conclusion will be presented at the end of this chapter.

9.2 Research objective

This stage mainly aims at fitting the testing data into the mined models to check and compare the predictive values with those conducted in other studies.

9.3 Research questions

The research question of this stage is:

- Can any of these prediction rules with high effective and high precise rate be found based on the clinical variables?

9.4 Method

This stage incorporated the model testing method to evaluate how well the newly identified prediction factors are by using the same testing data of Stage Two.

9.4.1 Procedure

The testing data was then entered into the newly identified prediction models and the subsequent prediction scores were calculated. The status of the subjects contracting SARS was checked against the laboratory result (using RT-PCR). Predictive values of the prediction model were calculated.

9.4.2 Data management

Data was tested by being fit into various prediction rules with the help of the software, PASW, to compare their predictive value, specificity and sensitivity.

9.5 Results

9.5.1 General clinical predictors for adults

The researcher made use of the testing model formed at Stage 3, that is, chills, sore throat, and malaise which were the most critical clinical predictors. The accuracy of the testing dataset to classify patients with the identified critical clinical predictors was measured. Over 88% in correct classification was attained and this result was considered to be fairly high in accuracy.

9.5.2 Differences in clinical prediction for different situations

The percentage of correct classification attained under different situations was presented in Table 9.

Table 9 Percentage of correct classification of testing dataset under different situations

	Percentage of correct classification of cases (%)	
	Classification & Reduction Tree	Clustering
Geographical factors (Amoy residents vs. Non-Amoy residents)	92	74.36
Gender (Male vs. Female)	97.53	80
Age (Young vs. Old)	64	58.8
Comorbidity (with vs. without underlying medical illnesses)	94.57	66.67

The testing data of the database in this study were fitted into the newly identified clinical prediction rules. The raw scores of their testing result was presented as error bars together with other prediction rules mentioned in this study as shown in

Figure 26 below.

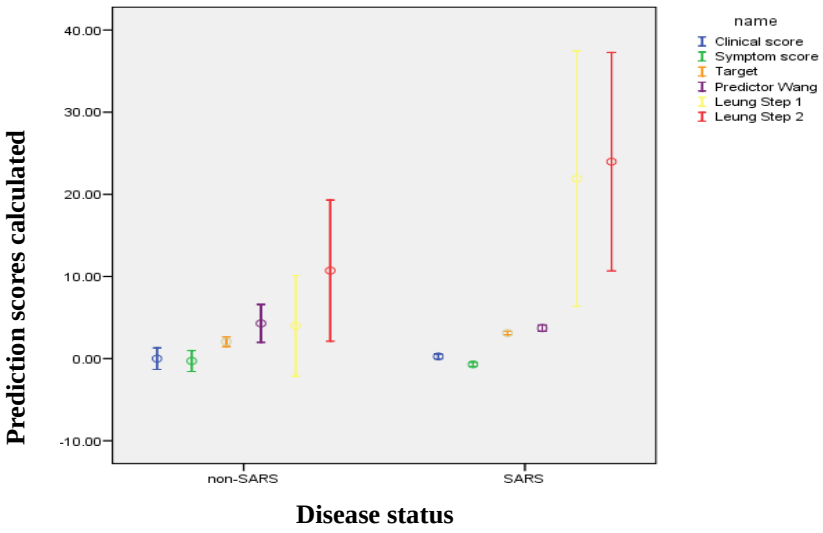

Figure 26 Distribution of raw predictive scores of various prediction rules (N=110)

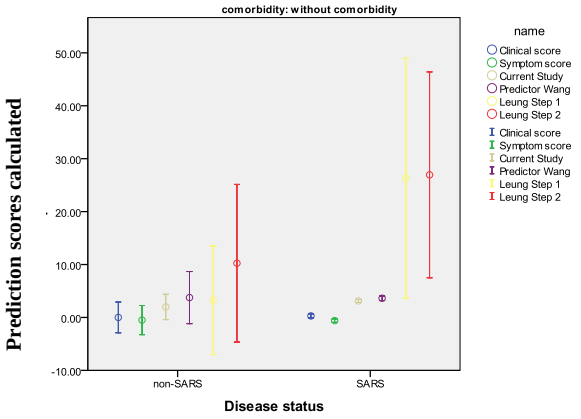

Figure 26a Distribution of raw predictive scores of various prediction rules by disease group without comorbidity (N=73)

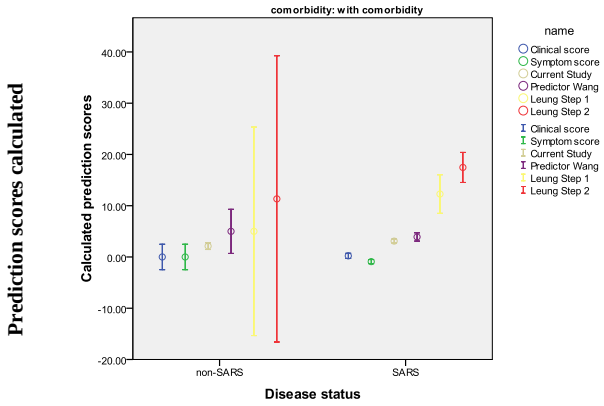

Figure 26b Distribution of raw predictive scores of various prediction rules by disease group with comorbidity (N=37)

From the Figure 26a & 26b, there are no significant difference in identifying SARS and non-SARS patients with various existing prediction rules even stratified for presence and absence of comorbidity.

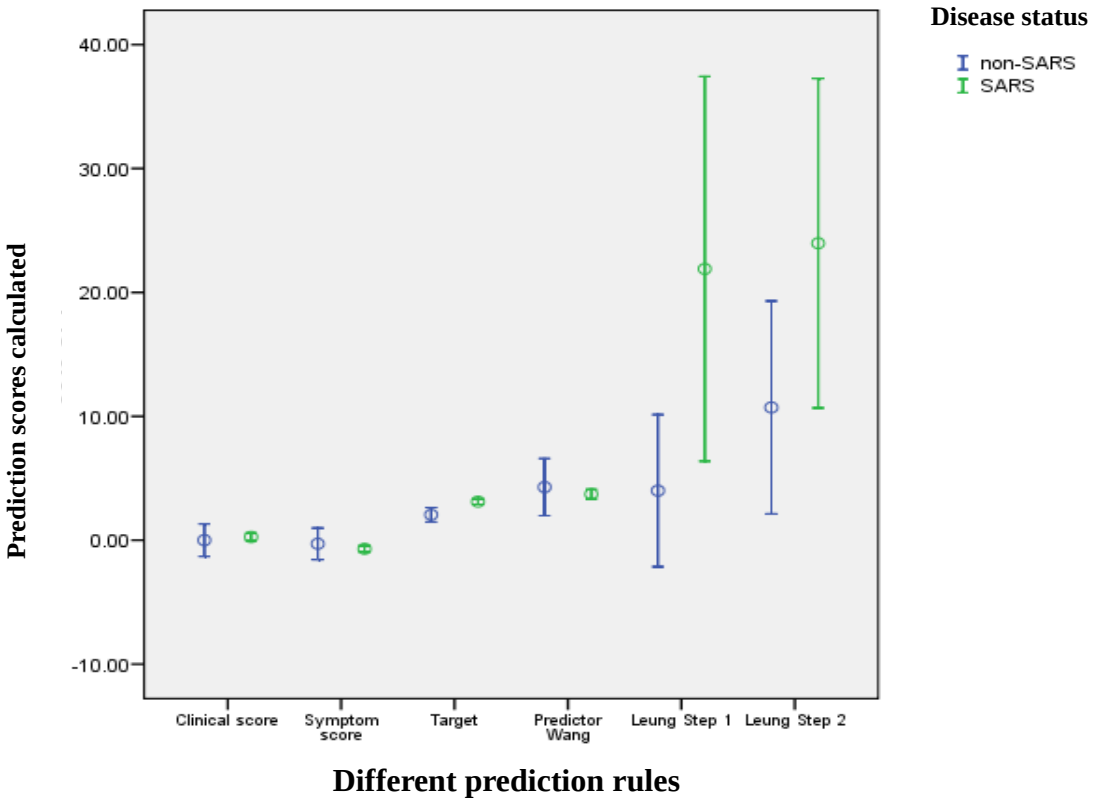

Figure 27 Comparison of raw predictive scores of existing prediction rules (N=110)

The distribution of the raw scores from various studies does not show a clear-cut differentiation between SARS cases and non-SARS cases. And the sensitivity, specificity, positive predictive value and negative predictive value of these prediction rules in predicting SARS occurrence were calculated as shown in Table 10.

Table 10 Calculated values of sensitivity, specificity and predictive values for different prediction models in predicting the occurrence of SARS

	Chen et al (2004)		Wang et al. (2004)	Leung et al. (2004)		Current study
	Clinical score	Symptom score	Predictor	Step 1	Step 2	Target
Quoted sensitivity	100	100	100	99	95	--
Calculated Sensitivity	43.14	48.04	16.50	95.10	92.16	86.27
Quoted specificity	86.3	75.9	93	63	57	--
Calculated Specificity	57.14	42.86	66.67	0	42.86	71.43
Positive predictive value	93.62	92.45	89.47	93.27	95.92	96.70
Negative predictive value	6.45	5.36	4.44	0	27.27	22.22

From here, we can summarize that the sensitivity and the specificity of the newly identified predictors are having better predictive values when compared with other existing prediction models developed by other researchers.

9.6 Discussion

9.6.1 General clinical predictors for adults

It is crucial for the health care management team to have the latest information concerning any new disease so as to provide appropriate and efficient information to their front line workers. Knowledge on general clinical predictors for patients supports the education needs of health professionals on prevention at earlier stage before actual outbreak of infectious incident happen. Individuals with experience in working on SARS case definitions are able to identify potential risks by means of the clinical predictors. The training is best delivered within a centralized surveillance and assessment system by an expert team with clinical and public health expertise such as knowledge of the epidemiological features of the disease and the availability of necessary additional resources (Timen, et al, 2006).

Employing an efficient and accurate prediction for classifying the patients correctly with the disease is extremely essential. The result from this study can shed some light in how to prepare better future practice.

As seen from the research questions (p. 21-22) and results (p.179, 192), apart from factors like gender, age, comorbidity, we need to consider geographical factors as one of the risk factors for SARS as well as other airborne disease. Hence, in the future, if we are facing another unknown diseases (no matter it is airborne or droplet), we should also consider their geographic factors as it may have contribution in identifying the potential cases timely.

9.6.2 Using of data mining technique

Traditionally, it is common for the health care professional to do the manual analysis and interpretation periodically so as to turn the medical data into knowledge. However, the time required for the whole process will be timely especially for large data volume. It usually requires a lot of manpower, especially if it involves the change of different thresholds which require performing all the processes again. Suggestion of using another prediction method is instrumental.

In this study, the patterns mined were further tested and compared with other prediction models in the same area against accuracy. All the information helps us to analyze clinical variables which are critical in predicting SARS

occurrence and whether the variables are different under different headings (for example, residential area, gender difference, with or without co-morbidity). The nontrivial method used is not a straightforward computation of predefined quantities like computing the average value of a set of numbers, instead, it is a user-oriented, domain specific automatic pattern recognition with thresholds chosen by the user. Eventually the mined rules were compared with the existing prediction rules done by other researchers to check whether data mining can be a more beneficial method in arousing the citizens' as well as the health care professionals' awareness and to minimize the transmission of respiratory infectious disease in the future.

Data mining technique can be broadly classified as supervised learning or unsupervised learning. Supervised data mining algorithm only allows a single output attribute (that is the dependent variable) whereas the unsupervised data mining is to apply some measure of similarity to divide instances into disjoint partitions. Hence, the learning program builds a knowledge structure by using some measure of the cluster quality to group instances into two or more classes. It was adopted in this study to identify the hidden pattern and characteristics of the attributes from large volume of clinical data. In association rules mining, it is mainly to deal with which

consequent may contain multiple conditions and attribute relationships. An output attribute is an association rule that can be an input attribute in another rule. It is somehow an affinity analysis. It will be the best choice when attributes are allowed to play multiple roles in the data mining process. Hence, if we changed it to whether a patient has certain kinds of condition, such as hyponatremia or hypernatremia, there will be a better chance to get a positive result.

Subsequently, the classification technique was put to action based on the association rule mining results found at the first stage of the study. Possible models for classifying the occurrence of SARS were identified. Decision Tree Induction which caters for classifying data sets with millions of examples and hundreds of attributes with reasonable speed was applied. Its comparatively faster learning speed than other classification methods and easy conversion to simple understandable classification rules are noted advantages.

Decision tree induction for continuous-valued attributes, which can dynamically define new discrete-valued attributes that partition the continuous attribute value into a discrete set of interval, was used to handle the missing attribute values by simply assigning the most common value of

the attribute by their probability to each of the possible values. This step serve to construct new attributes based on existing ones that are sparsely represented, thus avoiding fragmentation, repetition and replication. As mentioned before, classification mining is mainly to assign a class to find previously unseen records as accurately as possible. Hence, it is most commonly recruited as a decision making technique. The mined rules are actually originated from a supervised training of the available training data and are tested against the testing data.

Finally, clustering technique is helpful to locate certain high-risk group of clients that were more susceptible to contract SARS, for example, spatial clustering located estates of high SARS occurrence. Eventually, a statistical model was constructed for clinical management of SARS in the future based on data from three-quarters of all cases.

Hui, Chan, Wu, & Ng (2004) reported that the major clinical features of the SARS patients include persistent fever, chills/rigor, myalgia, malaise, dry cough, headache, and dyspnoea. Older subjects may present without the typical febrile response. Measurement of serum RNA by real time reverse transcriptase-polymerase chain reaction technique has a detection rate of 75% – 80% in the first week of the illness. The testing model yielded by the

classification technique can attain 82.35% correct classification of the testing data.

9.7 Conclusion

From this study, one can see that the use of data mining can be an alternative way for predicting whether the client presenting with specific critical predictors is actually contracting the disease or not before laboratory test available. But in this study, it is just a retrospective case notes review. The contribution of the study may be confirmed if it can be tested in real time situation in the future.

At the very beginning, it is obviously an application for decision making to determine which patients are likely to have contracted the disease and require further isolation facilities. But if the database is ready after initial setup of the computer system, this logic can be used as surveillance since the system can yield the result in a short period of time.

Among these three types of data mining techniques (association rule mining, classification and clustering), classification rule mining can help identifying

critical factors for rule mining where clustering technique can also serve the same purpose with good specificity and sensitivity.

Hence, from the result of this study, we could recommend Hospital Authority to employ data mining as the surveillance method of which the classification rule mining and clustering technique can help to identify critical factors with good specificity and sensitivity. However, as this method requires a large volume of data on hand for processing, it needs a comprehensive database construct beforehand which may need more time and manpower. Since the electronic medical database currently used in HA is usually input after patient's discharge, HA is migrating to paperless electronic medical records. But it needs time to train all health care professionals on how to use computers to record patients' progress. But once, the database infrastructure has consolidated, the screening processing can be very fast and efficient compared with the existing method.

The limitation of using data mining is noteworthy because a large volume of data is required on hand for processing. In the initial phase of an infectious disease, the computer software is possible to support the health administrators with known minimum confidence level input for analysis. On the other hand, the software promotes differentiation between relevant

variables from their irrelevant counterparts so that those unwanted variables will be deleted (to avoid confusion). As it stated on P.87, data mining deals with all cases in the population recruited in the database, so any notions of significant testing lose. Hence, it will be more appropriate to do the mining again if the condition is different from the current one with new specific known criteria.

Although SARS may not come back again in the near future, the Avian influenza together with any other potential health hazards will affect many lives (Tsang et al., 2006), and so we should pay attention. Such pandemic disease not only affects an individual's life but also poses great burden in our healthcare system in terms of manpower and resources (Tolomiczenko et al., 2005). Emergency Department as the first spot flooded with patients suffering from suspected infection striking the healthcare system provides a clear picture of the actual demand on healthcare services. (Chen, Cheng, Chung & Lin, 2005).

Early detection to avoid the spread during incubation period is always the best solution to prevent the disease from getting out of control (Rea et al., 2007). Adequate public health education focusing on measures taken by individuals to reduce the spread of infection (Jefferson et al., 2007) and

ways to increase body resistance (Tai, 2006) can supplement the surveillance to combat pandemic infection.

References

Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., et al. (2004). Microevolution and history of the plague bacillus, Yersinia pestis. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 17837-17842.

Ahmad, A., Krumkamp, R., & Reintjes, R. (2009). Controlling SARS: a review on China's response compared with other SARS-affected countries. *Trop Med Int Health*.;14 Suppl1:36-45.

Arita, I., Nakane, M., Kojima, K., Yoshihara, N., Nakano, T., & El-Gohary, A. (2004). Role of a sentinel surveillance system in the context of global surveillance of infectious diseases.[erratum appears in Lancet Infect Dis. 2004 Aug;4(8):533]. *The Lancet Infectious Diseases, 4*(3), 171-177.

Bachmann, L. M., Kolk, E., Koller, M. T., Steurer, J., & ter Riet, G. (2003). Accuracy of the Ottawa ankle rules to exclude fractures of the ankle and mod-foot: A systematic review. *BMJ, 326*, 417-423.

Backmann, L. M., Haberzeth, S., Steurer, J., & ter Riet, G. (2004). The accuracy of the Ottawa knee rule to rule out knee fractures: A systematic review. *Annals of Internal Medicine, 140*, 121-124.

Berkelman, R. L., Byran, R. T., Osterholm, M. T., LeDuc, J. W., & Hughes, J. M. (1994). Infectious disease surveillance: a

crumbling foundation. *Science, 264*, 368-370.

Booth, C. M., Matukas, L. M., Tomlinson, G. A., Rachlis, A. R., Rose, D. B., Dwosh, H. A., et al. (2003). Clinical features and short-term outcomes of 144 patients with SARS in the greater Toronto area. *JAMA, 289*(21), 2801-2809.

Brossette, S. E., Sprague, A. P., Hardin, J. M., Jones, K. W. T., & Moser, S. A. (1998). Association rules and data mining in hospital infection control and public health surveillance. *Journal of American Medical Informatics Association, 5*, 373-381.

Buchsbaum, D. G., Buchanan, R. G., Centor, R. M., Schnoll, S. H., & Lawton, M. J. (1991). Screening for alcohol abuse using CAGE scores and likelihood ratios. *Annals of Internal Medicine, 115*, 774-777.

Burnet, F.M., & White, D.O. (1972). Natural history of infectious disease (4th ed.) Preface.p. ix Cambridge University Press. Melbourne. Retrieved 18 Nov, 2010, from http://www.google.com/books?hl=zh-TW&lr=&id=ifQ3AAAAIAAJ&oi=fnd&pg=PR9&dq=when+infectious+disease+first+appeared&ots=Fh-MJZMJ6Y&sig=NQFndFIHipknOKsDzFSaPDIted0#v=onepage&q=when%20infectious%20disease%20first%20appeared&f=false

Cabena, P., Hadjinian, P., Stadler, R., Verhcees, J., & Zanasi, A. (1998). *Discovering Data Mining: From Concept to Implementation.* .

New Jersey: Prentice Hall.

Cao, B., Liu, Z. Y., Wang, M. Z., Cai, B. Q., Xu, Z. J., Bai, Y., et al. (2003). Clinical diagnosis, treatment and prognosis of elderly SARS patients. *Chung-Kuo i Hsueh Ko Hsueh Yuan Hsueh Pao Acta Academiae Medicinae Sinicae, 25*(5), 547-549.

Centers for Disease Control and, P. (2003a). Revised U.S. surveillance case definition for severe acute respiratory syndrome (SARS) and update on SARS cases--United States and worldwide, December 2003. *MMWR - Morbidity & Mortality Weekly Report, 52*(49), 1202-1206.

Centers for Disease Control and, P. (2003b). Severe acute respiratory syndrome (SARS) and coronavirus testing--United States, 2003.[erratum appears in MMWR Morb Mortal Wkly Rep. 2003 Apr 18;52(15):345]. *MMWR - Morbidity & Mortality Weekly Report, 52*(14), 297-302.

Centers for Disease Control and Prevention. (2003). Update: severe acute respiratory syndrome--United States, 2003. *Morbidity & Mortality Weekly Report, 52*(18), 411-413.

Centers for Disease Control and Prevention. (2006a). Pandemic Flu: Key Facts. Retrieved 18 Oct, 2009, from http://www.cdc.gov/flu/pandemic/pdf/pandemicflufacts.pdf

Centers for Disease Control and Prevention. (2006b). *Principles of Epidemiology in Public Health Practice* (3rd ed.). Atlanta: U.S. Department of Health and Human Services.

Centers for Disease Control and Prevention. (2006c). *Principles of Epidemiology in Public Health Practice: An introduction to applied epidemiology and biostatistics Self Study- course SS1000* (3rd ed.). Atlanta: Centres for Disease Control and Prevention (CDC).

Centers for Disease Control and Prevention. (2006d). Update on Multi-State Outbreak of E. coli O157:H7 Infections From Fresh Spinach, October 6, 2006. Retrieved 2 Jan, 2009, from http://www.cdc.gov/foodborne/ecolispinach/100606.htm

Centre for Health Protection. (2010, 18 November 2010). Update on H5N1 case. *Press releases* Retrieved 20 Nov, 2010, from http://www.chp.gov.hk/en/content/116/22276.html

Chan, J. C. K., Tsui, E. L. H., & Wong, V. C. W. (2007). Prognostication in severe acute respiratory syndrome: a retrospective time-course analysis of 1312 laboratory-confirmed patients in Hong Kong. *Respirology, 12*(4), 531-542.

Chan, L. S., Cheung, G. T. Y., Lauder, I. J., Kumana, C. R., & Lauder, I. J. (2004). Screening for fever by remote-sensing infrared thermographic camera. *Journal of Travel Medicine, 11*(5), 273-279.

Chan, S. M. S., Chiu, F. K. H., Lam, C. W. L., Leung, P. Y. V., & Conwell, Y. (2006). Elderly suicide and the 2003 SARS epidemic in Hong Kong. *International Journal of Geriatric Psychiatry, 21*(2), 113-118.

Chan, T. Y., Miu, K. Y., Tsui, C. K., Yee, K. S., & Chan, M. H. (2004). A comparative study of clinical features and outcomes in young and older adults with severe acute respiratory syndrome. *Journal of the American Geriatrics Society, 52*(8), 1321-1325.

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffice injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention, 38*, 1019-1027.

Chen, C. Y., Lee, C. H., Liu, C. Y., Wang, J. H., Wang, L. M., & Perng, R. P. (2005). Clinical features and outcomes of severe acute respiratory syndrome and predictive factors for acute respiratory distress syndrome.[see comment]. *Journal of the Chinese Medical Association: JCMA, 68*(1), 4-10.

Chen, S. Y., Su, C. P., Ma, M. H. M., Chiang, W. C., Hsu, C. Y., Ko, P. C., et al. (2004). Predictive model of diagnosing probable cases of severe acute respiratory syndrome in febrile patients with exposure risk. *Ann Emerg Med, 43*(1), 1-5.

Chen, W. K., Cheng, Y. C., Chung, Y. T., & Lin, C. C. (2005). The impact of the SARS outbreak on an urban emergency department in Taiwan. *Medical Care, 43*(2), 168-172.

Cheng, H. M., & Kwok, T. (2004). Mild SARS in elderly patients.[comment]. *CMAJ Canadian Medical Association Journal, 170*(6), 927.

Chiang, C. H., Shih, J. F., Su, W. J., & Perng, R. P. (2004). Eight-

month prospective study of 14 patients with hospital-acquired severe acute respiratory syndrome. *Mayo Clinic Proceedings, 79*(11), 1372-1379.

Childs, J., Fritz, J., Flynn, T., Irrgang, J., Johnson, K., Majkowski, G., et al. (2004). A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: A validation study. *Annals of Internal Medicine, 141*, 920-928.

Chiu, R. W. K., Chim, S. S. C., & Lo, Y. M. D. (2003). Molecular epidemiology of SARS--from Amoy Gardens to Taiwan.[comment]. *New England Journal of Medicine, 349*(19), 1875-1876.

Chiu, W. T., Lin, P. W., Chiou, H. Y., Lee, W. S., Lee, C. N., Yang, Y. Y., et al. (2005). Infrared thermography to mass-screen suspected SARS patients with fever. *Asia-Pacific Journal of Public Health, 17*(1), 26-28.

Chiu, Y. C., Wu, K. L., Chou, Y. P., Fong, T. V., Tsai, T. L., Kuo, C. M., et al. (2004). Diarrhoea in medical care workers with severe acute respiratory syndrome. *Journal of Clinical Gastroenterology, 38*(10), 880-882.

Choi, K. W., Chau, T. N., Tsang, O., Tso, E., Chiu, M. C., Tong, W. L., et al. (2003). Outcomes and prognostic factors in 267 patients with severe acute respiratory syndrome in Hong Kong. *Annals of Internal Medicine, 139*(9), 715-723.

Chong, W. C., Tham, K. Y., Goh, H.K., & Seow, E. (2005).

Presentation of severe acute respiratory syndrome (SARS) patients in a screening centre. *Singapore Medical Journal, 46*(4):161-164.

Colwell, R. R. (1996). Global climate and infectious disease: the cholera paradigm. *Science, 274*(5295), 2025-2031.

Copas, J. B., & Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika 89*, 315-331.

Cuckle, H. (2004). Principles of screening. *The Obstetrician & Gynaecologist, 6*(1), 21-25.

Dai, L. K., & Chan, J. (2003a). Challenges in diagnosis and management of SARS in elderlys. Retrieved 23 Dec, 2009, from http://medicine.org.hk/hkgs

Dai, L. K., & Chan, J. (2003b). Challenges in diagnosis and management of SARS in elders. *Hong Kong Geriatrics Society Interhospital Geriatric Meeting, May 30, 2003* Retrieved 17 Dec, 2008, from http://medicine.org.hk/hkgs/

Dawson-Saunders, B., & Trapp, R. G. (1994). *Basic & Clinical Biostatistics* (2nd ed.). Connecticut: Prentice-Hall International Inc.

Donnelly, C. A., Ghani, A. C., Leung, G. M., Hedley, A. J., Fraser, C., Riley, S., et al. (2003). Epidemiological determinants of spread

of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet, 361*(9371), 1761-1766.

Dunham, M. H. (2003). *Data Mining: Introductory and Advanced Topics.* Upper Sadle River: Prentice Hall.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadeiphia: Capital City Press.

Elliott, T., Worthungton, T., Osman, H., & Gill, M. (2007). *Lecture Notes. Medical microbiology & inflection.* (4th ed.). U.S.A.: Blackwell.

Fan, C. K., Yieh, K. M., Peng, M. Y., Lin, J. C., Wang, N. C., & Chang, F. Y. (2006). Clinical and laboratory features in the early stage of severe acute respiratory syndrome. *Journal of Microbiology, Immunology & Infection, 39*(1), 45-53.

Flament, M. F., Whitaker, A., Rapoport, J.L., Davies, M., Berg, C. Z., Kalikow, K., Sceery, W., & Shaffer, D. (1988) Obsessive Compulsive Disorder in Adolescence: An Epidemiological Study. Journal of the American Academy of Child & Adolescent Psychiatry: 27(6):764-771.

Flynn, T., Fritz, J., Whitman, J., Wainner, R., Magel, J., Rendeiro, D., et al. (2002). A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine, 27*, 2835-2843.

Gilbert, R., Logan, S., Moyer, V. A., & Elliot, E. J. (2001). Assessing

diagnostic and screening tests: Part1. Concepts. *Western Journal of Medicine, 174*(June), 405-409.

Goel, S., Gupta, A. K., Singh, A., Lenka, S. R. (2007). Environmental epidemiology practitioners: looking to the future. *Journal of Hospital Infection, 66*(2):142-147.

Halsall, P. (1996). Boccaccio: The Decameron, "Introduction". Retrieved 27 Dec, 2008, from http://www.fordham.edu/halsall/source/decameronintro.html

Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Technique*. San Fransisco: Morgan Kaufmann.

Hand, D. J. (1998). Data Mining: Stastics and More? *The American Statistician, 52*(2), 112-118.

Hand, D. J. (1999). Statistics and Data Mining: Intersecting Disciplines. *SIGKDD Explorations, 1*(1), 16 -19.

Harti, G. (10 August 2010). Transcript of virtual press conference with Dr Margaret Chan, Director -General, World Health Organization and Dr Keiji Fukuda, Special Adviser to the Director-General on Pandemic Influenza. *Pandemic (H1N1) 2009 press briefings* Retrieved 11 August, 2010

Hattori, T. (2005). [Newly emerging infections (including SARS)]. *Nippon Naika Gakkai Zasshi - Journal of Japanese Society of Internal Medicine, 94*(9), 1915-1920.

Hawker, G. A., Jamal, S. A., Ridout, R., & Chase, C. (2002). A clinical preduction rule to identify pre-menopausal woman with low bone mass. *Osteoporosis International 13*, 400-406.

Henderson, D. A., Courtney, B., Inglesby, T. V., Toner, E., & Nuzzo, J. B. (2009). Public Health and Medical Responses to the 1957-58 Influenza Pandemic [Electronic Version]. *biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 7, 1-9. Retrieved 5 Jan 2010 from http://www.upmc-biosecurity.org/website/resources/publications/2009/pdf/2009-08-05-public_health_medical_responses_1957.pdf.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Dame, P., & Buetels, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium [Electronic version]. *BMC Infectious Diseases*, 9(5), 1-18. Retrieved 29 June, 2010, from http://www.biomedcentral.com/1471-2334/9/5

Heung, L. C. L., Li, T., Mak, S. K., & Chan, W. M. (2006). Prevalence of subclinical infection and transmission of severe acute respiratory syndrome (SARS) in a residential care home for the elderly. *Hong Kong Medical Journal*, *12*(3), 201-207.

Heymann, D. L. (2005). Social, behavioural and environmental factors and their impact on infectious disease outbreaks.[comment]. *Journal of Public Health Policy, 26*(1), 133-139.

Hilleman, M. (2002). Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control. *Vaccine, 20*(25-26),

3068-3087.

Hoey, J. (2003). Updated SARS case definition using laboratory criteria. *CMAJ Canadian Medical Association Journal*, *168*(12), 1566-1567.

Hornby, A. S. (2005). *Oxford Advanced Learner's English-Chinese Dictionary* (7th ed.). Hong Kong: Oxford University Press.

Hui, D. S. C., Chan, M. C. H., Wu, A. K., & Ng, P. C. (2004). Severe acute respiratory syndrome (SARS): epidemiology and clinical features. *Postgraduate Medical Journal*, *80*, 373-381.

Hui, R. K. H., Zeng, F., Chan, C. M. N., Yuen, K. Y., Peiris, J. S. M., & Leung, F. C. C. (2004). Reverse transcriptase PCR diagnostic assay for the coronavirus associated with severe acute respiratory syndrome. *Journal of Clinical Microbiology, 42*(5), 1994-1999.

Imai, T., Takahashi, K., Hasegawa, N., Lim, M.-K., & Koh, D. (2005). SARS risk perceptions in healthcare workers, Japan. *Emerging Infectious Diseases*, *11*(3), 404-410.

International Food Safety Network. (2006). Outbreaks of foodborne illness linked to fresh lettuce and spinach since 1993 Retrieved 4 Oct, 2010, from http://foodsafety.ksu.edu/en/article-details.php?a=3&c=14&sc=98&id=903

Izrael, D., Battaglia, A. A., Hoaglin, D. C., Battaglia, M. P., & Abt Associate Inc. (2002). Use of the ROC Corve and the Bootstrap in Comparing Weighted Logistic Regression Models.

Retrieved 1 Nov, 2010, from http://www2.sas.com/proceedings/sugi28/275-28.pdf

Jefferson, T., Foxlee, R., Del Mar, C., Dooley, L., Ferroni, E., Hewak, B., et al. (2007). Interventions for the interruption or reduction of the spread of respiratory viruses. *Cochrane Database of Systematic Reviews*(4), CD006207.

Jr Taylor, F. B., Toh, C. M., Hoots, W. K., Wada, H., & Levi, M. (2003). *Towards a definition, clinical and laboratory criteria, and a scoring system for DIC. Offical communicaations of the Scientific and Standization Committees and the International Society on Thrombosis and Haemostasis*. Chapal Hill: University of North Carolina at Chapel Hill School of Medicine.

Kawana, A. (2003). [Infection control measures for SARS during epidemics of influenza]. *Nippon Rinsho - Japanese Journal of Clinical Medicine, 61*(11), 2019-2024.

Kaydos-Daniels, S. C., Olowokure, B., Chang, H.-J., Barwick, R. S., Deng, J.-F., Lee, M.-L., et al. (2004). Body temperature monitoring and SARS fever hotline, Taiwan. *Emerging Infectious Diseases, 10*(2), 373-376.

Khan, N. A. (2008). *Microbial pathogens and human diseases*. Enfield NH: Science Publshers.

Kong, T. K., Dai, L. K., Leung, F., Au, Y., Yung, & Chan, H. (2003). Severe acute respiratory syndrome (SARS) in elders. *Journal*

of the American Geriatrics Society, 51(8), 1182-1183.

Kudyba, S. (2004). *Managing Data Mining: Advice from Experts.* Hersey: Cyber Tech Publishing.

Kutyrev, V. V. (2008). [Quarantine infectious diseases and sanitary control of territories in modern conditions]. *Zhurnal Mikrobiologii, Epidemiologii i Immunobiologii*(1), 17-23.

Lam, C. M., Wong, S. F. Leung, T. N., Chow, K. M., Yu, W. C., Wong, T. Y., Lai, S. T. & Ho, L.C. (2004). A case-controlled study comparing clinical course and outcomes of pregnant and non-pregnant women with severe acute respiratory syndrome. BJOG: an International Journal of Obstetrics and Gynaecology, 111:771-774.

Lee, N., Hui, D., Wu, A., Chan, P., Cameron, P., Joynt, G. M., et al. (2003). A major outbreak of severe acute respiratory syndrome in Hong Kong. *New England Journal of Medicine, 348*(20), 1986-1994.

Leong, H.-N., Chan, K.-P., Oon, L. L. E., Koay, E. S. C., Ng, L.-C., Lee, M.-A., et al. (2006). Clinical and laboratory findings of SARS in Singapore. *Annals of the Academy of Medicine, Singapore, 35*(5), 332-339.

Leppin, A. & Aro, A. R. (2009). Risk perceptions related to SARS and avian influenza: theoretical foundations of current empirical research. *International Journal of Behaviour Medicine, 16*(1):7-29.

Leung, G. M., Hedley, A., J., Kong, J., Lam, T. H., Lau, F. L., Rainer, T., et al. (2008). A clinical prediction rule for diagnosing severe acute respiratory syndrome in the emergency department. *Hong Kong Med J, 14*(Suppl 5), S8-10.

Leung, G. M., Rainer, T. H., Lau, F. L., Wong, I. O. L., Tong, A., Wong, T. W., et al. (2004). A clinical prediction rule for diagnosing severe acute respiratory syndrome in the emergency department. *Annals of Internal Medicine, 141*(5), 333-342.

Lew, T. W. K., Kwek, T.-K., Tai, D., Earnest, A., Loo, S., Singh, K., et al. (2003). Acute respiratory distress syndrome in critically ill patients with severe acute respiratory syndrome. *JAMA, 290*(3), 374-380.

Li, G., Zhao, Z., Chen, L & Zhou, Y. (2003). Mild Severe Acute Respiratory Syndrome. *Emerging Infectious Diseases, 9*(9):1182-1183.

Li, J., Fu, A. W. C., & Fahey, P. (2008). Efficient discovery of risk patterns in medical data [Electronic Version]. *Artificial Intelligence in Medicine*, 1-13. Retrieved 6 Mar 2009 from http://www.cse.cuhk.edu.hk/~adafu/Pub/riskpattern08.pdf.

Li, J., Fu, A. W. C., He, H., Chen, J., Jin, H., McAullay, D., et al. (2005). Mining Risk patterns in Medical Data. Retrieved 27 Jan, 2010, from http://www.cse.cuhk.edu.hk/~adafu/Pub/sigkdd05.pdf

Li, Y., Duan, S., Yu, I. T. S., & Wong, T. W. (2005). Multi-zone modeling of probable SARS virus transmission by airflow between flats in Block E, Amoy Gardens. *Indoor Air, 15*(2), 96-111.

Liang, J., Bennett, J. M., Sugisawa, H., Kobayashi, E. & Fukaya, T. (2003). Gender differences in old age mortality: roles of jealth behaviour and baseline health status. *Journal of Clinical Epidemiology 56*(6):572-82.

Liao, J. W., Lu, J. H., Guo, Z. M., Wang, G. L., Zhang, D. M., Chen, L.-J., et al. (2007). A retrospective serological study of severe acute respiratory syndrome cases in Guangdong province, China. *Chinese Medical Journal, 120*(8), 718-720.

Lim, M. K. (2006). Bird flu: pandemic flu preparation: an unheeded lesson from SARS.[comment]. *BMJ, 332*(7546), 913.

Lim, W., Ng, K.-C., & Tsang, D. N. C. (2006). Laboratory containment of SARS virus. *Annals of the Academy of Medicine, Singapore, 35*(5), 354-360.

Liu, C. L., Lu, Y. T., Peng, M. J., Chen, P. J., Lin, R. L., Wu, C. L., et al. (2004). Clinical and laboratory features of severe acute respiratory syndrome vis-a-vis onset of fever. *Chest, 126*(2), 509-517.

Lovis C, Douglas T, Pasche E, Ruch, P, Colaert D & Stroetmann K. DebugIT: Building a European distributed clinical data mining network to foster the fight against microbial diseases.

Retrieved 28 July, 2010, from http://www.empirica.com/publikationen/documents/2009/2009%20PSIP%20Paper.pdf

McCampbell, B., Wasif, N., Rabbitts, A., Staiano-Coico, L., Yurt, R. W., & Schwartz, S. (2002). Diabetes and Burns: Retrospective Cohort Study. *Journal of Burn Care & Rehabilitation, 23*(3), 157-166.

McGee, D. L. (2010). Testing. Retrieved 19 Aug, 2010, from http://www.merckmanuals.com/professional/sec22/ch328/ch328e.html

McKinney, K. R., Gong, Y. Y., & Lewis, T. G. (2006). Environmental transmission of SARS at Amoy Gardens. *Journal of Environmental Health, 68*(9), 26-30; quiz 51-22.

Memmel, H., Kowal-Vern, A., & Latenser, B. A. (2004). Infections in Diabetic Burn Patients. *Diabetes Care, 27*(1), 229-233.

Morelli, G., Song, Y., Mazzoni, C. J., Eppinger, M., Roumagnac, P., Wagner, D. M., et al. (2010). Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity [Electronic Version]. *Nature Genetics*. Retrieved 15 Nov 2010 from

Moreno RP, Rhodes A & Chiche J-D (2009). The Ongoing H1N1 Flu Pandemic and the Intensive care community: challenges, opportunities, and the duties of scientific societies and intensivists. Intensive Care Medicine 35:2005-2008.

Murray, P. R., Rosenthal, K. S., & Pfaller, M. A. (2005). *Medical Microbiology* (5th ed.). philadelphia: Elsevier Mosby.

Ng, E. Y. K. (2005). Is thermal scanner losing its bite in mass screening of fever due to SARS? *Medical Physics, 32*(1), 93-97.

Ng, S. K. C. (2003). Possible role of an animal vector in the SARS outbreak at Amoy Gardens. *Lancet, 362*(9383), 570-572.

Nishiura, H., Kuratsuji, T., Quy, T., Phi, N. C., Van Ban, V., Ha, L. E. D., et al. (2005). Rapid awareness and transmission of severe acute respiratory syndrome in Hanoi French Hospital, Vietnam. *American Journal of Tropical Medicine & Hygiene, 73*(1), 17-25.

Oshitani H, Kamigaki T & Suzuki A (2008). Major Issues and Challenges of Influenza Pandemic Preparedness in Developing Countries. Emerging Infectious Diseases 14 (6):875-880

Paladini, M. (2004). Daily Emergency Department Surveillance System --- Bergen County, New Jersey. *MMWR - Morbidity & Mortality Weekly Report, 53 Suppl*, 47-49.

Peiris, J. S. M., Chu, C. M., Cheng, V. C. C., Chan, K. S., Hung, I. F. N., Poon, L. L. M., et al. (2003a). Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study. *Lancet, 361*(9371), 1767-1772.

Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Guan, Y., Yam, L. Y. C., Lim, W., et al. (2003b). Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet, 361*(9366), 1319-1325.

Portney, L. G. & Watkins M. P. (2009). *Foundations of Clinical Research: Applications to Practice* (3rd ed.). New York: Prentice Hall.

Poutanen, S.M., Low, D.E., Henry, B., Rose, D., Green, K., Tellier, R., Draker, R., Adachi, D., Ayers, M., Chan, A.K., Skowronski, D.M., Salit, I., Simor, A.E., Slutsky, A.S., Doyle, P.W., Krajden, M., Petric, M., Brunham R.C., McGeer, A.J. (2003). Identification of severe acute respiratory syndrome in Canada. *New England Journal of Medicine. 348*(20):1995-2005.

Rainer, T. H. (2004). Severe Acute Respiratory Syndrome: Clinical Features, Diagnosis, and Management. Current Opinion in Pulmonary Medicine; *10*(3)

Rea, E., Lafleche, J., Stalker, S., Guarda, B. K., Shapiro, H., Johnson, I., et al. (2007). Duration and distance of exposure are important predictors of transmission among community contacts of Ontario SARS cases. *Epidemiology & Infection, 135*(6), 914-921.

Roiger, R. J., & Geatz, M. W. (2003). *Data Mining: A-tutorial-based primer.* . Boston: Addison Wesley.

Saijo, M., Ogino, T., Taguchi, F., Fukushi, S., Mizutani, T., Notomi, T., et al. (2005). Recombinant nucleocapsid protein-based IgG

enzyme-linked immunosorbent assay for the serological diagnosis of SARS. *Journal of Virological Methods, 125*(2), 181-186.

Sampathkumar, P., Temesgen, Z., Smith, T. F., & Thompson, R. L. (2003). SARS: Epidemiology, Clinical Presentation, Management, and Infection Control Measures. *Mayo Clin Proc, 78*, 882-890.

Senanayake, S. N. (2006). The limitation of fever in case definitions for avian influenza and SARS. *Communicable Diseases Intelligence, 30*(2), 250.

Surak, J. G. (2007). A Recipe for Safe Food: ISO 22000 and HACCP. *Quality Progress, 40*(10), 21-27.

Tablan, O. C., Anderson, L. J., Besser, R., Bridges, C., Hajjeh, R., Cdc, et al. (2004). Guidelines for preventing health-care--associated pneumonia, 2003: recommendations of CDC and the Healthcare Infection Control Practices Advisory Committee. *Morbidity & Mortality Weekly Report Recommendations & Reports, 53*(RR-3), 1-36.

Taguchi, F. (2003). [SARS coronavirus]. *Uirusu, 53*(2), 201-209.

Tai, D. Y. H. (2006). SARS: how to manage future outbreaks? *Annals of the Academy of Medicine, Singapore, 35*(5), 368-373.

Tang, P., Louie, M., Richardson, S. E., Smieja, M., Simor, A. E., Jamieson, F., et al. (2004). Interpretation of diagnostic

laboratory tests for severe acute respiratory syndrome: the Toronto experience. *Canadian Medical Association Journal, 170*(1), 47-54.

Taubenberger, J. K. & Morens, D. M. (2006). 1918 Influenza: the Mother of All Pandemics. Emerging Infectious Diseases 12 (1):15-22

Taylor, F. B. Jr, Toh, C. H., Hoots, W. K., Wada, H., Levi, M. (2001). Towards definition, clinical and laboratory criteria, and a scoring system for disseminated intravascular coagulation. *Thromb Haemost, 86*:1327-30.

Teruya, K. (2003). [Current state of understanding of SARS and infection control measures]. *Nippon Ronen Igakkai Zasshi - Japanese Journal of Geriatrics, 40*(6), 553-558.

The University of Hong Kong. (2005). Influenza Pandemic - Are You Prepared? Retrieved 15 March, 2010, from http://hku.hk/uhs/avianflu/pandemic_uhs.htm

Tolomiczenko, G. S., Kahan, M., Ricci, M., Strathern, L., Jeney, C., Patterson, K., et al. (2005). SARS: coping with the impact at a community hospital. *Journal of Advanced Nursing, 50*(1), 101-110.

Trifonov, V., Khiabanian, H., & Rabadan, R. (2009). Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus [Electronic Version]. *The New England Journal of Medicine,* 361, 115-119 from

http://www.nejm.org/doi/full/10.1056/NEJMp0904572.

Trujillano, J., Badia, M., Servia, L. M., J., & Rodriguez-Pozo, A. (2009). Stratification of the severity of critically ill patients with classification trees [Electronic Version]. *BMC Medical Research Methodology*, 9, 1-12. Retrieved 14 Dec 2009 from http://www.biomedcentral.com/1471-2288/9/83.

Tsang, K. W., Shim, Y.-S., Wong, T. K. S., Liam, C. K., Eng, P., Lam, W. K., et al. (2006). Possible case scenarios and logistic issues in H5N1 pandemic. *Respirology, 11*(5), 520-522.

Tsang, T., & Lam, T. H. (2003). SARS: public health measures in Hong Kong. *Respirology, 8 Suppl*, S46-48.

Tse, M. M. Y., Pun, S. P. Y., & Benzie, I. F. F. (2003). Experiencing SARS: perspectives of the elderly residents and health care professionals in a Hong Kong nursing home. *Geriatric Nursing, 24*(5), 266-269.

Tsuboi, S., Fukukawa, Y., Niino, N., Ando, F., & Shimokata, H. (2004). [Age and gender differences as factors related to depressive symptoms among community-dwelling middle-aged and elderly people]. *Shinrigaku Kenkyu - Japanese Journal of Psychology, 75*(2), 101-108.

U. S. Department of Health & Human Service. (2010). History of Flu Pandemics. Retrieved 29 May, 2010, from http://pandemicflu.gov/individualfamily/about/pandemic/history.html

Virella, G. (1997). Microbiology and infectious diseases In (3rd ed., pp. 91-97; 205-211). Baltimore: Williams & Wilkins.

Vu, H. T., Leitmeyer, K. C., Le, D. H., Miller, M. J., Nguyen, Q. H., Uyeki, T. M., et al. (2004). Clinical description of a completed outbreak of SARS in Vietnam. *Emerging Infectious Diseases, 10*(2), 334-338.

Wang, J. T., Sheng, W. H., Fang, C. T., Chen, Y. C., Wang, J. L., Yu, C. J., et al. (2004). Clinical manifestations, laboratory findings, and treatment outcomes of SARS patients. *Emerging Infectious Diseases, 10*(5), 818-824.

Wang, T. L., Jang, T. N., Huang, C. H., Kao, S. J., Lin, C. M., Lee, F. N., et al. (2004). Establishing a clinical decision rule of severe acute respiratory syndrome at the emergency department. *Annals of Emergency Medicine, 43*(1), 17-22.

Weingart, U., Lavi, Y. & Horn, D (2009). Data mining of enzymes using specific peptides [Electronic version]. *BMC Bioinformatics*, 10(446), 1-18. Retrieved 29 June, 2010, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811123/pdf/1471-2105-10-446.pdf

Wells, P. S., Hirsch, J., Andersson, D. R., Lensing, A., Foster, G., kearon, C., et al. (1997). Value of assessment of pre-test probability of deep-vien thrombosis in clinical management. *Lancet, 350*, 1795-1798.

Wenzel, C., Kofler, J., Locker, G. J., Laczika, K., Quehenberger, P., Frass, M., et al. (2002). Endothelial cell activation and blood coagulation in critically ill patients with lung injury [Abstract]. [Electronic Version]. *Wien Klin Wochenschr*, 114, 853-858.

Wilder-Smith, A., Earnest, A., & Paton, N. I. (2004). Use of simple laboratory features to distinguish the early stage of severe acute respiratory syndrome from dengue fever. *Clinical Infectious Diseases, 39*(12), 1818-1823.

Williams, P., Hand, D. J., & Tarnopolsky, A. (1982). The problem if screening for uncommon disorders - a comment on the Eating Attitudes Test. *Psychological Medicine, 12*, 431-434.

Wills, B. S. H., & Morse, J. M. (2008). Responses of Chinese elderly to the threat of severe acute respiratory syndrome (SARS) in a Canadian community. *Public Health Nursing, 25*(1), 57-68.

Wilson, J. M. G. (1971). Principles of Screening for Disease. *Current status and value of laboratory screening tests*

Wiwanitkit, V. (2007). Lymphopenia in severe acute respiratory syndrome: a summary on its frequency. *Nepal Medical College Journal, 9*(2), 132-133.

World Health Organization. (2003a). Case Definitions for Surveillance of Severe Acute Respiratory Syndrome (SARS). Retrieved 28 Apr, 2008, from http://www.who.int/csr/sars/casedefinition/en/

World Health Organization. (2003b). Use of Laboratory Methods for

SARS Diagnosis. Retrieved 3 Dec, 2008, from www.who.int/csr/sars/labmethods/en/

World Health Organization. (2004). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003.

World Health Organization. (2007). Addressing sex and gender in epidemic-prone infectious diseases. Retrieved 13 Jan, 2010, from http://influenzatraining.org/documents/s17632en/s17632en.pdf

World Health Organization. (2009). Pandemic (H1N1) 2009 - update 71. Retrieved 13 Jan, 2010, from http://www.who.int/csr/don/2009_10_23/en/index.html

World Health Organization. (2010a). Part II. Global Health Indicators. *World Health Statistics*

World Health Organization. (2010b). World Health Statistics 2010. Retrieved 18 Jun, 2010, from http://www.who.int/whosis/whostat/EN_WHS10_Full.pdf

Yip, C., Chang, W. L., Yeung, K. H., & Yu, I. T. S. (2007). Possible meteorological influence on the severe acute respiratory syndrome (SARS) community outbreak at Amoy Gardens, Hong Kong. *Journal of Environmental Health, 70*(3), 39-46.

Yu, S., Qiu, M., Chen, Z., Ye, X., Gao, Y., Wei, A., et al. (2005). Retrospective serological investigation of severe acute respiratory syndrome coronavirus antibodies in recruits from

mainland China. *Clinical & Diagnostic Laboratory Immunology, 12*(4), 552-554.

Zheng, R.Y. (2005). Mining SARS-CoV protease cleavage data using non-orthogonal decision trees: a novel method for decisive template selection Bioimformatics 21(11) [Electronic version]. Retrieved 28 July, 2009, from http://portal.acm.org/citation.cfm?id=1094214

Zhong, N. & Zeng G. (2006). What we have learnt from SARS epidemics in China. *British Medical Journal*, 333(7564): 389–391