

國立台灣大學電機資訊學院資訊工程學系

博士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Doctoral Dissertation

應用機器學習演算法識別罹患帶狀疱疹之高危險群

Applying Machine Learning Algorithms to Identify

Patients with High Risk of Developing Herpes Zoster

李定達

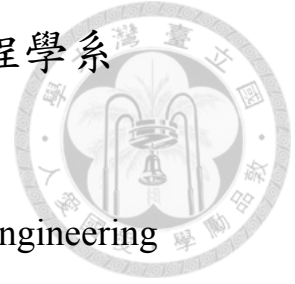
Ding-Dar Lee

指導教授:歐陽彥正博士

Advisor: Yen-Jen Oyang, Ph.D.

中華民國 102 年 7 月

July 2013



致謝



能夠順利在台大資工取得博士學位，首先要感謝的是指導教授歐陽彥正老師，歐陽老師對於我這個不具資工背景的學生始終帶著耐心開導啟發，不時關心我的進度與現況，即使我的研究能力沒能突破，老師也都沒有對我放棄。另外還要謝謝趙坤茂老師當初鼓勵我報考本所與在資工領域的啟蒙，孫維仁老師對研究方向的指導與建議，以及楊孟翰老師在我研究過程及論文寫作上的大力協助。在這段求學過程中，特別要謝謝實驗室先後期的學長姐和學弟妹們的幫忙，尤其是黃乾綱老師、陳倩瑜老師、佩均、榮元和孜如。此外，也要感謝台北榮總皮膚部的劉漢南主任給予我的支持和包容，使我能夠沒有後顧之憂地將部分時間分配在研究所的課業上。

當然更要感謝恩重如山、始終對我不斷付出無怨無悔的雙親大人，妹妹、妹夫和外甥的加油打氣，還要謝謝祥雲、蕙敏、迺騰、珮如、趙亮、正捷和佩璇，你們都是我生命中的貴人。最後，我要嘉許我自己。我原本是一個資工的門外漢，竟然能有驚無險地通過了四科資格考，也順利修過了十八個學分的課程。儘管這個過程比念醫學系還要漫長，但我憑藉著「堅持到底」的信念終究還是修成了正果。這個學位對我而言，一方面彌補了大學時期沒能考上台大的遺憾，一方面也證明了自己除了醫學之外，還擁有跨領域的學習能力。

總而言之，無盡感謝。

中文摘要



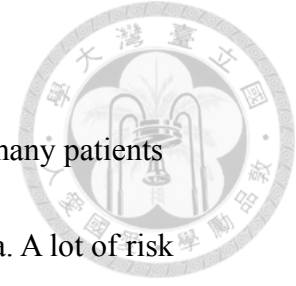
帶狀疱疹是台灣常見的疾病之一，許多患者還受到其併發症——疱疹後神經痛所苦。雖然已有許多帶狀疱疹的危險因子被報告過，例如紅斑性狼瘡，但大部份報告都只針對單一疾病。近年來機器學習演算法已被大量應用於大型醫學資料庫之分析且獲致極有價值之訊息。本研究針對台灣之全民健康保險研究資料庫，利用兩種特性相異之機器學習演算法——「決策樹」及「線性元件分析」，同時對多種危險因子進行全面性的分析。我們首先以線性鑑別分析法進行特徵選擇，並找出最具相關性的八種共病：心律不整、缺乏性貧血、腎衰竭、類風濕性關節炎、冠狀動脈心臟病、惡性腫瘤、慢性肺病及風濕性疾病。這八種疾病皆以線性元件分析方法加以整合分析，而後四種疾病則用於建構決策樹分析。除了一個末端節點外，所有決策樹之末端節點及線性元件分析之不等式皆得到有意義之勝算比及百分之九十五信賴區間，顯示這兩個方法皆有能識別出部分罹患帶狀疱疹之高危險群。其中兩個末端節點及四個不等式甚至得到大於三之勝算比，這些患者具有較多次因惡性腫瘤、風濕性疾病、冠狀動脈心臟病或腎衰竭的就診紀錄。我們發現有一群患者可以同時被決策樹及線性元件分析找到，特別是風濕性疾病或冠狀動脈心臟病。線性元件分析可以找到決策樹所無法鑑別的一群腎衰竭患者。雖然決策樹與線性元件分析所識別的帶狀疱疹患者有所重疊，卻也有部分互斥，對於不同族群具有相異之鑑別力，能夠以互補的方式識別出不同族群，因此聯集這兩個演算法之結果或可增加識別之靈敏度。

關鍵字：帶狀疱疹、危險因子、全民健康保險研究資料庫、機器學習演算法、決

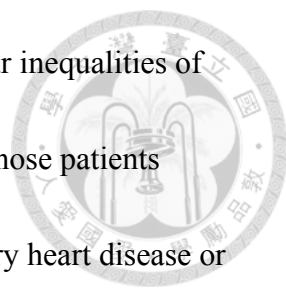
策樹、線性元件分析



Abstract



Herpes zoster is a relatively common disorder in Taiwan and many patients suffer from its most notorious complication-- postherpetic neuralgia. A lot of risk factors of herpes zoster have been reported, e.g., systemic lupus erythematosus, but most of the observations were based on one single disorder. In this study, rather than focusing on a specific risk factor, we conducted comprehensive analyses on the joint effects of multiple risk factors based on the National Health Insurance Research Database in Taiwan. In recent years, machine learning algorithms have been applied to the analyses of large medical databases and have yielded valuable information. In this respect, we employed two machine learning algorithms with distinctive characteristics, decision tree and linear component analysis. We first performed feature selection by linear discriminant analysis and identified the most relevant eight comorbidities of herpes zoster, i.e. arrhythmia, deficiency anemia, renal failure, rheumatoid arthritis, coronary heart disease, malignancy, chronic pulmonary disease, and rheumatologic disease; all of them were used in linear component analysis while the last four disorders were incorporated to generate the decision tree. All leaf nodes except for one of decision tree and all inequalities of linear component analysis corresponded to a significant odds ratio and 95% confidence interval, indicating that both methods were able to identify some sub-populations of patients with higher risk

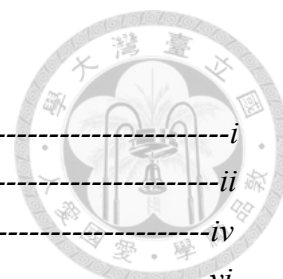


of developing herpes zoster. Two leaf nodes of decision tree and four inequalities of linear component analysis yielded odds ratios even higher than 3. Those patients showed more visits for malignancy, rheumatologic disorder, coronary heart disease or renal failure. We noticed a very specific subgroup of patients, especially those of rheumatologic disorder or coronary heart disease, could be identified by both decision tree and linear component analysis algorithms at the same time while linear component analysis discovered cases of renal failure that decision tree was unable to recognize. Though there were overlaps of cases identified by some leaf nodes of decision tree and inequalities of linear component analysis, some extent of exclusions also existed between the herpes zoster cases identified by both algorithms. It implied that decision tree and linear component analysis were able to recognize different populations of herpes zoster cases in a complementary way and therefore union of the results obtained by both methods could possibly increase the sensitivity.

Key Words: *herpes zoster; risk factor; National Health Insurance Research*

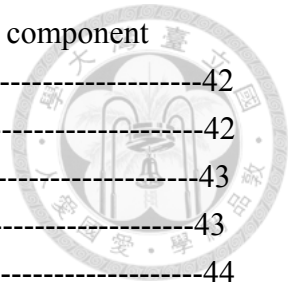
Database; machine learning algorithm; decision tree; linear component analysis

Table of Contents



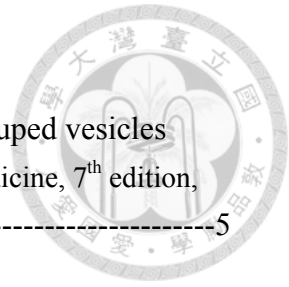
| | |
|---|------|
| 致謝 | i |
| 中文摘要 | ii |
| Abstract | iv |
| Table of Contents | vi |
| List of Figures | viii |
| List of Tables | ix |
| | |
| 1. Motivation | 1 |
| 2. Background | 3 |
| 2.1 Introduction of herpes zoster | 3 |
| 2.1.1 Etiology | 3 |
| 2.1.2 Complications | 9 |
| 2.1.3 Incidence | 12 |
| 2.1.4 Risk factors | 12 |
| 2.1.5 Prevention | 16 |
| 2.2 Introduction of National Health Insurance Research Database | 17 |
| 2.3 Introduction of machine learning algorithms | 18 |
| 2.3.1 Decision tree | 19 |
| 2.3.2 Linear component analysis | 20 |
| 2.3.3 Rank-based adaptive mutation evolutionary algorithm | 20 |
| 2.4 Application of machine learning algorithms to large medical databases | 24 |
| 3. Materials and Methods | 27 |
| 3.1 Case definition and control selection | 27 |
| 3.2 Diseases utilized as potential risk factors | 27 |
| 3.3 Feature selection by linear discriminant analysis | 30 |
| 3.4 Machine learning algorithms | 30 |
| 3.4.1 Decision tree | 31 |
| 3.4.2 Linear component analysis and RAME | 31 |
| 3.5 Statistical Analysis | 32 |
| 4. Results | 33 |
| 4.1 Demographic data | 33 |
| 4.2 Univariate analysis | 33 |
| 4.3 Linear discriminant analysis | 34 |
| 4.4 Decision tree | 36 |
| 4.5 Linear component analysis | 39 |

| | |
|--|----|
| 4.6 Analysis of cases identified by decision tree and/or by linear component analysis----- | 42 |
| 4.6.1 Leaf nodes 6 & 10 of DT----- | 42 |
| 4.6.2 Inequalities 4, 5, 7 and 9 of LCA----- | 43 |
| 4.6.3 DT and LCA validates each other----- | 43 |
| 4.6.4 DT and LCA is complementary to each other----- | 44 |
| 5. Discussion----- | 47 |
| 6. Conclusions----- | 53 |
| 7. Future Work----- | 55 |
| References----- | 57 |



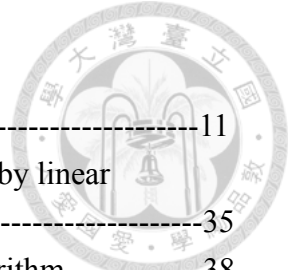
List of Figures

- Figure 1. Clinical pictures of herpes zoster showing unilaterally grouped vesicles on red base (from: Fitzpatrick's Dermatology in General Medicine, 7th edition, p1890)-----5
- Figure 2. Dermatome map (from: <http://neurobiography.info/teaching.php?mode=view&lectureid=37&slide=20>)---6
- Figure 3. Decrease in varicella-zoster virus-specific immunity with age in immunocompetent individuals (from: Vaccination: a new option to reduce the burden of herpes zoster. Mick G. Expert Rev. Vaccines. 2010;9 (3 Suppl.): p32)-----7
- Figure 4. Age- and sex-specific incidence rate (and 95% confidence intervals) of herpes zoster during 2000 to 2006 in Taiwan (from: Epidemiological features and costs of herpes zoster in Taiwan: a national study 2000 to 2006. Jih JS *et al.* Acta Derm Venereol. 2009;89 (6): p613) -----8
- Figure 5. A diagram to illustrate the concept of RAME (from: http://zoro.ee.ncku.edu.tw/bp/res/10-machine_learning.pdf)-----23
- Figure 6. Decision tree analysis to identify cases of herpes zoster-----37



List of Tables

| | |
|--|----|
| Table 1. Potential complications of herpes zoster----- | 11 |
| Table 2. Patient numbers, count of visits and correlation coefficient by linear discriminant analysis of 35 groups of diseases----- | 35 |
| Table 3. Characteristics of leaf nodes obtained by decision tree algorithm----- | 38 |
| Table 4. Results of linear component analysis----- | 40 |
| Table 5. Analyses of weights of inequalities 1 & 4 of linear component analysis----- | 41 |
| Table 6. Distribution of cases of herpes zoster in the intersections of decision tree leaf nodes and LCA inequalities----- | 45 |
| Table 7. Exclusions in intersections of decision tree leaf nodes and LCA inequalities----- | 46 |



Chapter 1

Motivation



Herpes zoster is a common disorder. In Taiwan, during 2000 to 2006, the incidence of herpes zoster (HZ) was 4.89/1000 person-years.¹ Patients of herpes zoster frequently suffer from its complications. The most common complication of HZ is post-herpetic neuralgia (PHN). Forty percents of individuals aged 65 years or older reported moderate to severe impairment of general activities as a result of PHN.² In Taiwan, 8.6% of HZ patients still suffered from PHN even 3 months after the initiation of zoster.¹ Even in the acute stage, pain could also be quite severe-- 42% of patients referred to their worst pain as “horrible” or “excruciating”.³

However, herpes zoster is preventable. A live attenuated zoster vaccine has been licensed by the US Food and Drug Administration in 2006, to boost the specific immunity of older adults and, via this mechanism, to protect the individual against HZ and its complications. Unfortunately, the vaccination is expensive and costs \$160 to \$195 per dose.⁴ Therefore, from the aspect of public health economics, identifying those individuals with higher risk of developing HZ as the candidates for vaccination is a significant and practical issue.

It has been shown that disease states such as HIV infection,⁵ psychiatric disorders,⁶ chronic obstructive pulmonary diseases,⁷ rheumatoid arthritis,⁸ chronic

kidney disease,⁹ systemic lupus erythematosus¹⁰ and under long-term hemodialysis,¹¹ increase the risk of developing HZ. In this study, rather than focusing on a specific risk factor, we conducted comprehensive analyses on the joint effects of multiple risk factors based on a national health database.

Machine learning is supposed to be able to describe the observed data in some meaningful way. A lot of machine learning algorithms have been developed for the purpose of classification or prediction and many of them have been applied to analyze large medical databases. In this respect, we employed two machine learning algorithms with distinctive characteristics. The two algorithms are the well-known decision tree (DT) and linear component analysis (LCA). The decision tree algorithm has been employed to develop medical informatics based diagnosis and prognosis.^{12,13} One favorite characteristics of the decision tree algorithm is that the user can easily comprehend the rules output by the algorithm. Nevertheless, the conventional decision tree algorithm evaluates the significance of possible factors one by one. In other words, the decision tree algorithm may miss to identify the significance of several factors combined. Accordingly, in our study, we also employed a linear component based multivariate analysis to obtain more comprehensive results.

Chapter 2



Background

In this chapter, introduction of three major elements of this thesis, i.e. herpes zoster, National Health Insurance Research Database (NHIRD) and machine learning algorithms, will be provided.

2.1 Introduction of herpes zoster

Both herpes zoster and varicella are induced by varicella-zoster virus (VZV) but they show totally different clinical pictures. Understanding the etiology, complications, incidence, risk factors and a possible way of prevention of HZ is helpful to establish the plans and goals of research on this disorder.

2.1.1 Etiology

Herpes zoster or shingles is a disorder caused by activation of the VZV that has remained latent in the sensory ganglia and dorsal nerve roots following varicella (chicken pox) infection. HZ is characterized by unilaterally grouped vesicles with radiating pain (Fig. 1), which is generally limited to a single dermatome (the region of the body surface that sends sensory nerve projections to a single segment of the spinal cord) (Fig. 2).¹

The disease most commonly occurs as a result of an age-related decline in cell-mediated immunity (CMI) (Fig. 3) and therefore the incidence increased with

age (Fig. 4). CMI keeps the virus under control. It is known that this immunity may be boosted by periodic exposure to external sources of VZV, such as children infected with chickenpox and by periodic minor reactivation of internal virus. It is believed that when the CMI to VZV is insufficient to keep it under control, the virus reemerges from the dorsal root ganglia. However, the precise mechanism of virus activation is poorly understood.¹⁴

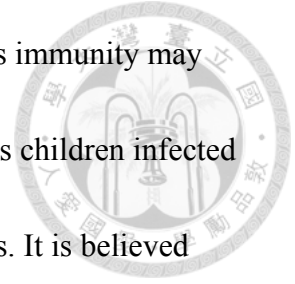




Figure 1. Clinical pictures of herpes zoster showing unilaterally grouped vesicles on red base.

(Fitzpatrick's Dermatology in General Medicine, 7th edition, p1890)

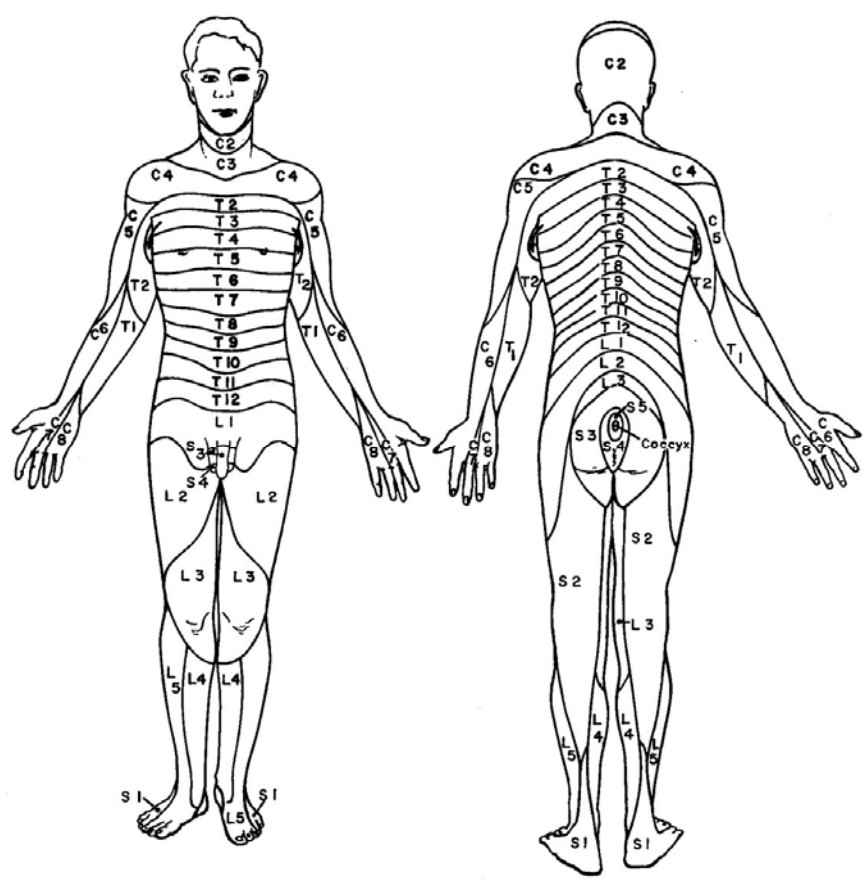


Figure 2. Dermatome map. (<http://neurobiography.info/teaching.php?mode=view&lectureid=37&slide=20>)

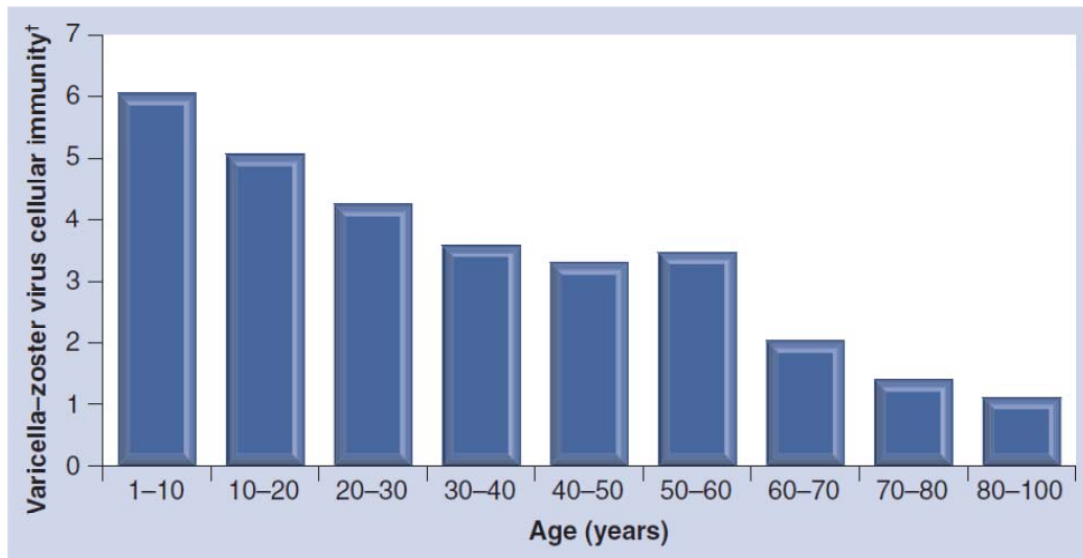


Figure 3. Decrease in varicella-zoster virus-specific immunity with age in immunocompetent individuals.¹⁵

†Measured by *in vitro* varicella-zoster virus-induced lymphocyte stimulation.

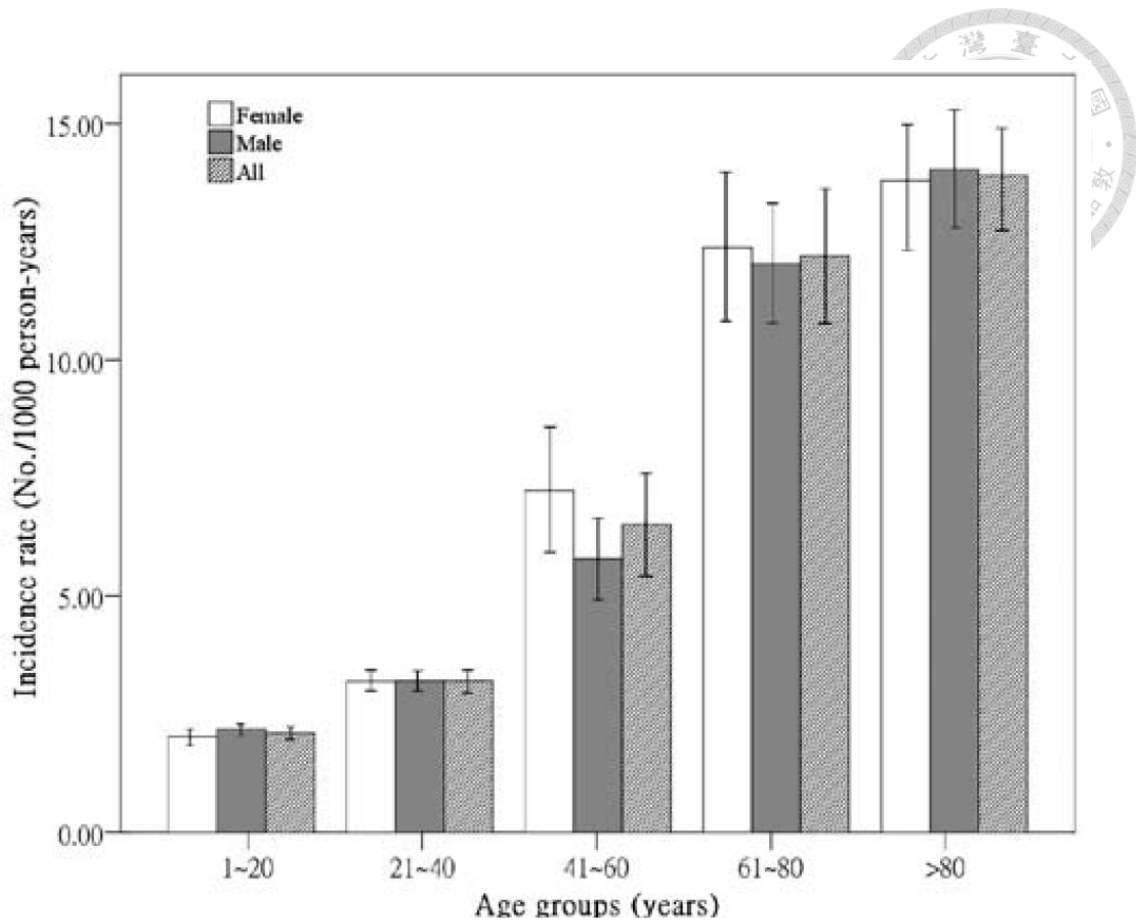
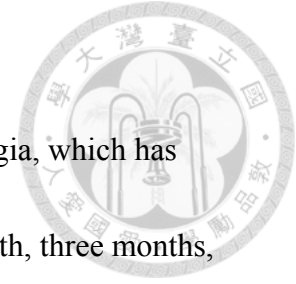


Figure 4. Age- and sex-specific incidence rate (and 95% confidence intervals) of herpes zoster during 2000 to 2006 in Taiwan.¹

2.1.2 Complications



The most common complication of HZ is post-herpetic neuralgia, which has been variably defined as any pain persisting for more than one month, three months, or four months after the rash appears. In one study of patients aged ≥ 65 years, the mean duration of pain was 3.3 years, and ranged from 3 months to more than 10 years.² It substantially reduces the day-to-day functioning and quality of life of affected individuals, particularly older adults. Katz *et al.* assessed the quality of pain in the acute stage and found that pain seemed to be the rule rather than the exception— only 4% of patients reported no pain.³ Most patients (58.9%) reported pain most days (14%) or every day (44.9%); individual pain episodes lasted from a few minutes for 25.6% to all day for 22.9% of patients. This pain was significantly correlated with impairment in physical functioning ($P < .001$), role functioning ($P < .001$), social functioning ($P < .001$) and with depressive symptoms ($P < .01$).³

Patients who have more severe pain during acute herpes zoster may be at increased risk of more prolonged PHN.³ In a survey of 385 individuals aged 65 years or older with PHN, 40% of individuals reported moderate to severe impairment of general activities, 45% reported a moderate to severe impairment in mood, and 48% reported moderate to severe impairment of their enjoyment of life as a result of PHN.² More than half (54%) had problems performing their usual activities, while 5% were

completely unable to do so.² PHN resulting in reduced quality of life and functional disability to a degree comparable to that experienced by patients with congestive heart failure, diabetes mellitus and major depression.¹⁶ Pain scores for PHN have been shown to be as high as those for chronic pain from osteoarthritis and rheumatoid arthritis.¹⁷

Table 1 showed other potential complications after HZ. Among them, herpes zoster ophthalmicus (HZ involving eyes) should be paid more attention to since permanent ocular damage, vision loss and even blindness can occur. Besides, epidemiological studies in Taiwan found patients with herpes zoster ophthalmicus have a 4.52-fold (95% confidence interval (CI), 2.45–8.33) higher risk of stroke than the matched comparison and the adjusted hazard ratio of stroke after herpes zoster during the 1-year follow-up period was 1.31 (95% CI, 1.00-1.60).¹⁸

Table 1. Potential complications of herpes zoster¹⁹

Neurologic

Postherpetic neuralgia

Motor paralysis

Meningoencephalitis

Transverse myelitis

Cerebral vasculitis

Cranial palsy

Ocular

Lid ulceration

Conjunctivitis, keratitis, uveitis

Optic neuritis

Retinal necrosis

Secondary glaucoma

Visceral

Pneumonitis

Myocarditis

Hepatitis

Esophagitis

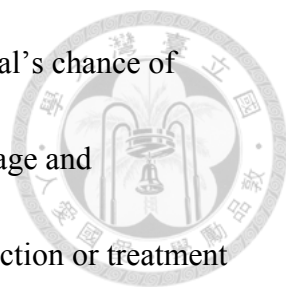


2.1.3 Incidence



The estimated lifetime risk of developing zoster in those exposed to varicella is 10–30%, while the incidence and severity of HZ increases with age. In USA, the incidence among individuals younger than 40 years ranges from 0.9 to 1.9 cases per 1000 patient-years, but it begins to climb thereafter: HZ occurs in 2.5 per 1000 patient-years among individuals aged 40 to 49 years; 3.8 cases among 50 to 59 years; 6.1 cases among 60 to 69 years; 8.5 cases among 70 to 79 years; and 9.4 cases among individuals aged 80 years or older.¹⁴ Sixty percent of cases occur in individuals aged 50 years or older. It affects up to 50% of people who live to 85 years. HZ results in 2.1 hospitalizations per 100,000 patient-years and the average cost per HZ-related hospitalization was \$15,583 in 1995.¹⁴ In Taiwan, during 2000 to 2006, the incidence of herpes zoster was 4.89/1000 person-years and 8.6% developed PHN 3 months after the initiation of the zoster (incidence 0.42/1000 person-years).¹ The hospitalization rate for HZ was 16.1 per 100,000 person-years. The largest proportion (59.5%) of hospitalizations was in adults older than 60 years of age. The costs for each home care case and hospitalized case were approximately NTD 1,652 and NTD 37,966, respectively.¹ Recurrence of a HZ episode is rare in immunocompetent patients, estimated at 1% to 6%.¹⁴

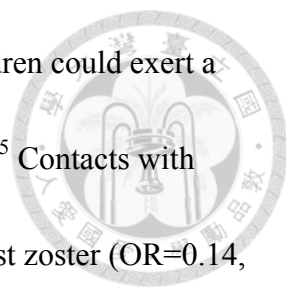
2.1.4 Risk factors



Factors that compromise the CMI to VZV increase an individual's chance of developing herpes zoster. Chief among these factors are advancing age and accompanying immune-senescence. Disease states such as HIV infection or treatment regimens that impair CMI may achieve similar results in any age group. Stress and physical trauma also appear to play a role in determining the timing and, possibly, location of HZ.¹⁴ In a case-control study of patients with HZ who were matched by age, sex, and race with HZ-free control subjects, individuals with HZ were significantly more likely to have had negative life events, e.g. death of a close family member or a personal accident, in the two (odds ratio (OR) 2.60, 95% CI 1.13-6.27), three (OR 2.64, 95% CI 1.20-6.04), or six months (OR 2.00, 95% CI 1.04-3.93) before the onset of zoster.²⁰

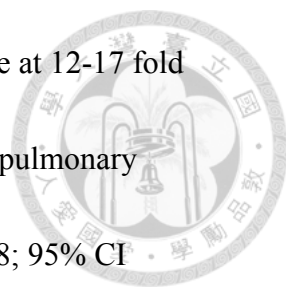
Studies on seasonality of zoster incidence were inconclusive. It was reported in Italy and Japan that the incidence of HZ was highest in summer and lowest in winter while Brisson et al. displayed no identifiable periodicity in the UK and Canada.²¹⁻²³ Urban/rural residence and population density/household crowding were investigated and none saw a significant effect.⁵ The risk of zoster also did not vary significantly with socioeconomic status (adjusted $p=0.935$).⁵

At the annual level, an increase in varicella incidence in children under 5 years old was accompanied by a significant decrease in zoster incidence among individuals



aged 15–44 years, suggesting that increased varicella in young children could exert a protective effect against zoster in the young adults exposed to them.⁵ Contacts with varicella cases were associated with a strong protective effect against zoster (OR=0.14, 95% CI 0.05–0.39 for those with ≥ 5 contacts). Contact with multiple ill children (e.g. general practitioners) was associated with significantly lower zoster risk (adjusted OR=0.20, 95% CI 0.06-0.73) for those with occupational contacts in the past 10 years but contact with multiple well children (e.g. teachers) was not associated with significant protection.⁵ On the contrary, a study in Taiwan based on NHIRD concluded that the incidence of HZ is higher among health-care workers and no protective effect against HZ exists for them in Taiwan.²⁴

As far as disease states were concerned, patients with inflammatory bowel diseases, chronic conditions characterized by altered regulation of the immune system, including Crohn’s disease and ulcerative colitis, were at higher risk for HZ (Crohn’s disease, incidence rate ratio, 1.61; 95% CI, 1.35–1.92; ulcerative colitis, incidence rate ratio, 1.21; 95% CI, 1.05–1.40), especially those on immunosuppressive medications.²⁵ After adjusting for potential confounders, patients with psychiatric disorders, including affective psychoses, neurotic illness and personality disorders, were more likely to have an episode of HZ than the control population (adjusted hazard ratio (HR), 1.29; 95% CI, 1.18–1.38).⁶ In three studies that compared zoster



incidence in HIV-positive and HIV negative people, the former were at 12-17 fold greater risk of developing zoster.⁵ Patients with chronic obstructive pulmonary diseases were more likely to have incidents of HZ (adjusted HR 1.68; 95% CI 1.45–1.95). The adjusted HR of HZ was 1.67 (95% CI 1.43–1.96) for patients with chronic obstructive pulmonary diseases not taking steroid medications, 2.09 for patients using inhaled corticosteroids only (95% CI 1.38–3.16) and 3.00 for patients using oral steroids (95% CI 2.40–3.75).⁷ The adjusted HR of HZ for patients with rheumatoid arthritis (RA) compared with non-RA patients were 1.91 (95% CI 1.80–2.03) in the US PharMetrics database and 1.65 (95% CI 1.57–1.75) in the UK General Practice Research Database.⁸ Unexpectedly, people in the North Carolina cohort with a history of cancer were not at increased risk of zoster after 8 years of follow-up (adjusted relative risk=1.03; 95% CI 0.58-1.80).²⁶

In Taiwan, after adjusting for potential confounders, the adjusted HR of HZ was 1.98 (95% CI, 1.72-2.27) in patients treated with long-term hemodialysis, 1.6 (95% CI, 1.41-1.81) in patients with chronic kidney disease and 2.45 (95% CI, 1.77-3.40) in patients with systemic lupus erythematosus (SLE).⁹⁻¹¹

Only a few studies focused on the possible risk factors of PHN. Jung et al. reported the risk indicators for the occurrence of PHN included older age, female sex, presence of a prodrome, greater rash severity, and greater acute pain. A clear elevation

in risk existed in patients with multiple risk factors. In addition, PHN developed in only 5% to 10% of patients who had none of these risk factors.²⁷



2.1.5 Prevention

A live attenuated zoster vaccine has been licensed by the US Food and Drug Administration in 2006. The HZ vaccinated group had a 51% lower incidence of HZ, a 67% reduction in PHN (defined as pain rated at three or more on a scale ranging from zero [no pain] to ten [pain as bad as you can imagine], persisting or occurring 3 months or more after rash onset), and a 61% lower burden of illness (a composite measure of the incidence, severity and duration of pain and discomfort caused by HZ), indicating that the vaccine decreased both the incidence of HZ and the average severity of HZ in vaccinees who developed HZ.¹⁵ Moreover, there was a 73% reduction in the number of cases of HZ with severe and long-lasting pain. Overall, the vaccine was well-tolerated with the most common adverse events being mild injection site reactions and headache.¹⁵

The vaccine is recommended by the Advisory Committee on Immunization Practices (October 25-26, 2006) of the Centers for Disease Control and Prevention of USA for the prevention of HZ in individuals aged 60 years or older.¹⁴ In Europe, this vaccine is indicated for the prevention of HZ and PHN in immunocompetent individuals 50 years of age or older. However, the vaccination costs \$160 to \$195 per

dose.⁴ The policy of vaccination for HZ in Taiwan has not been established or announced.



2.2 Introduction of National Health Insurance Research Database

Taiwan initiated a single-payer National Health Insurance (NHI) Program in 1995. Currently, there are more than 23 million enrollees in the program, representing approximately 99.9 % of Taiwan's entire population. NHI claims have been collected into the National Health Insurance Research Database which has been released to researchers in an electronically encrypted form since 1999. The database includes all claims from ambulatory care and inpatient care and comprises comprehensive information on insured registrants, such as demographic data, dates of clinical visits, diagnostic codes, details of prescriptions, and expenditure amounts. The diagnostic codes of the patients, in the format of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), were established by the board-certified physicians in the corresponding specialties. The NHI Bureau regularly audits claims and imposes fines for false claims and this database has been used extensively in many epidemiologic studies in Taiwan.²⁸ The major merit of NHIRD is that it's nationwide and population-based and therefore it provides good demographic diversity. Researchers are able to test their hypotheses expeditiously without spending too much time to recruit cases. In this study, we identified patients who were

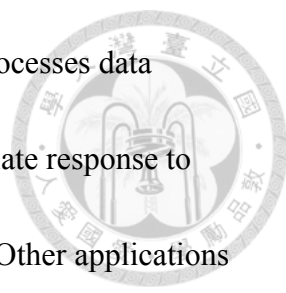
diagnosed as herpes zoster in outpatient and/or inpatient records from the NHIRD during 2004 to 2008 as “cases” and for each case, three age and sex-matched insurants without any record of HZ were randomly selected as “controls”.



However, there are some limitations of NHIRD-based studies. First, the disease diagnoses were coded according to the ICD9-CM and obtained from administrative claims reported by hospitals or clinics, which are considered less accurate than clinical diagnoses made by standard criteria and than those on official medical records. A second limitation of this medical claims database is patient compliance: inpatient or ambulatory care orders for drugs do not guarantee drug adherence. This may therefore overestimate or underestimate the effects of medication on the patients. Third, the administrative claims data from the NHIRD did not include detailed personal information (e.g., family history, body mass index, living habits such as smoking and alcohol use, or laboratory test results), which may be important confounding factors for the studies.

2.3 Introduction of machine learning algorithms

In 1959, Arthur Samuel defined machine learning as “a field of study that gives computers the ability to learn without being explicitly programmed”. The field of machine learning studies the design of computer algorithms capable of inducing patterns or rules from past experiences, e.g., inducing medical knowledge from

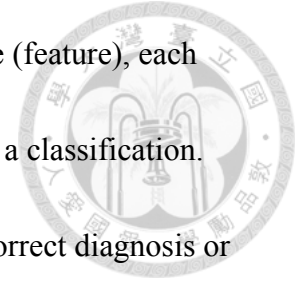


medical records by data mining. A “learner” (computer program) processes data representing past experiences and tries to either develop an appropriate response to future data, or describe the observed data in some meaningful way. Other applications include autonomous driving, speech recognition and self customizing programs, e.g., a newsreader that learns user interests. In this project, we tried to apply some well-established machine learning algorithms, including decision tree and linear component analysis to analyze and organize features of HZ patients obtained from NHIRD and to identify high risk patients who were more likely to suffer from HZ than others.

2.3.1 Decision tree

The technique of decision trees reaches a classification through consecutive question and answer sessions. It is for predicting the class of an object from the values of its discriminating variables. The tree is constructed by recursively partitioning a learning sample of data in which the values of the discriminating variables and the class label for each case are known. Each partition is represented by a node in the tree. It tries to reject the null hypothesis of independence between any of the discriminating variables and the class by a p-value corresponding to a test, e.g. student’s t-test.²⁹ This examines all possible splits along respective discriminating variables and selects the split that most reduces measure of node impurity and gives the largest information

gain. In brief, in a decision tree, each internal node tests an attribute (feature), each branch corresponds to an attribute value and each leaf node assigns a classification.



Decision tree has been used in the medical field to help making a correct diagnosis or

predicting prognosis. A decision tree has been constructed according to clinical

symptoms, medical history, laboratory or biopsy tests etc., to guide diagnostic

interpretation and therapeutic options for temporal arteritis.¹² Markey *et al.*

differentiated lung cancer cases from non-cancer controls by construction of the

decision tree according to proteins identified by mass spectrometry of blood serum

samples.¹³

2.3.2 Linear component analysis

The second machine learning algorithm employed in our study to identify groups

of subjects with high risk of developing HZ was a linear component analysis. It was

aimed to identify a hyperplane in the vector space defined by the features so that the

subjects on one side of the hyperplane suffered a higher risk of developing HZ than

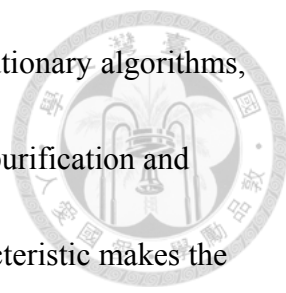
the subjects on another side of the hyperplane. An optimization algorithm was

invoked while implementing LCA. In this respect, we adopted the rank-based

adaptive mutation evolutionary (RAME) algorithm developed by our research team in

recent years as the optimization algorithm.³⁰

2.3.3 Rank-based adaptive mutation evolutionary algorithm³⁰



The RAME algorithm belongs to the general category of evolutionary algorithms, and it is able to generate populations with a good balance between purification and diversification even after a large number of generations. This characteristic makes the RAME algorithm quickly concentrate on a few optima that have been identified, including the global optimum and local optima, while still keeping some power to search for other optima. In each generation of RAME, n individuals, represented by s_1, s_2, \dots, s_n , are generated. Each s_i is in fact a vector corresponding to a point in the domain of the objective function. Let $t_1, t_2 \dots t_n$ denote the ordered individuals according to descending optimization score. In the RAME algorithm, n Gaussian distributions, denoted by $G_1, G_2 \dots G_n$, which govern the creation of the next generation of population are then identified. The center of each Gaussian distribution is selected randomly and independently from $t_1, t_2 \dots t_n$, where the probability is not uniform but instead follows a discrete diminishing scale, $n : n - 1 : \dots : 1$. That is, the probability of picking up t_1 for creating a Gaussian distribution is n times larger than that of picking up t_n , the probability of picking up t_2 is $n - 1$ times larger than that of picking up t_n , and so on, i.e., the centers of the n Gaussian distributions $G_1, G_2 \dots G_n$ are selected based on a probability distribution determined by how the individuals are ranked. One can expect that any individual, t_k , may be selected a number of times. Therefore, among $G_1, G_2 \dots G_n$, there may be multiple of them having an identical

mean and variance. After these centers have been selected, the next generation of individuals, denoted by s'_1, s'_2, \dots, s'_n respectively, is then created by taking one random sample from each of the Gaussian distributions. The probability distribution of G_i is governed by:

$$\left(\frac{1}{\sqrt{2\pi} \cdot \sigma_i}\right)^d \exp\left(-\frac{(s_i - t_k)^2}{2\sigma_i^2}\right),$$

where d is the dimension of the vector space, t_k is the center of G_i , and

$$\sigma_i^2 = \alpha + \frac{(\beta - \alpha)k}{n-1},$$

where σ_i is the standard deviation of Gaussian distribution G_i , α and β are two tuning parameters that control the width of the Gaussian distribution. The idea behind the RAME algorithm is that sharper Gaussians will be employed to generate next-generation individuals surrounding the current-generation individuals with better scores, while shallower Gaussians will be employed to generate next-generation individuals surrounding the current-generation with worse scores (Fig. 5). Owing to this design there will always be a finite probability of sampling regions surrounding previously identified local optima while the distribution becomes denser in the proximity of global optima as the evolution progresses.

RAME (Rank-based Adaptive Mutation Evolutionary) algorithm

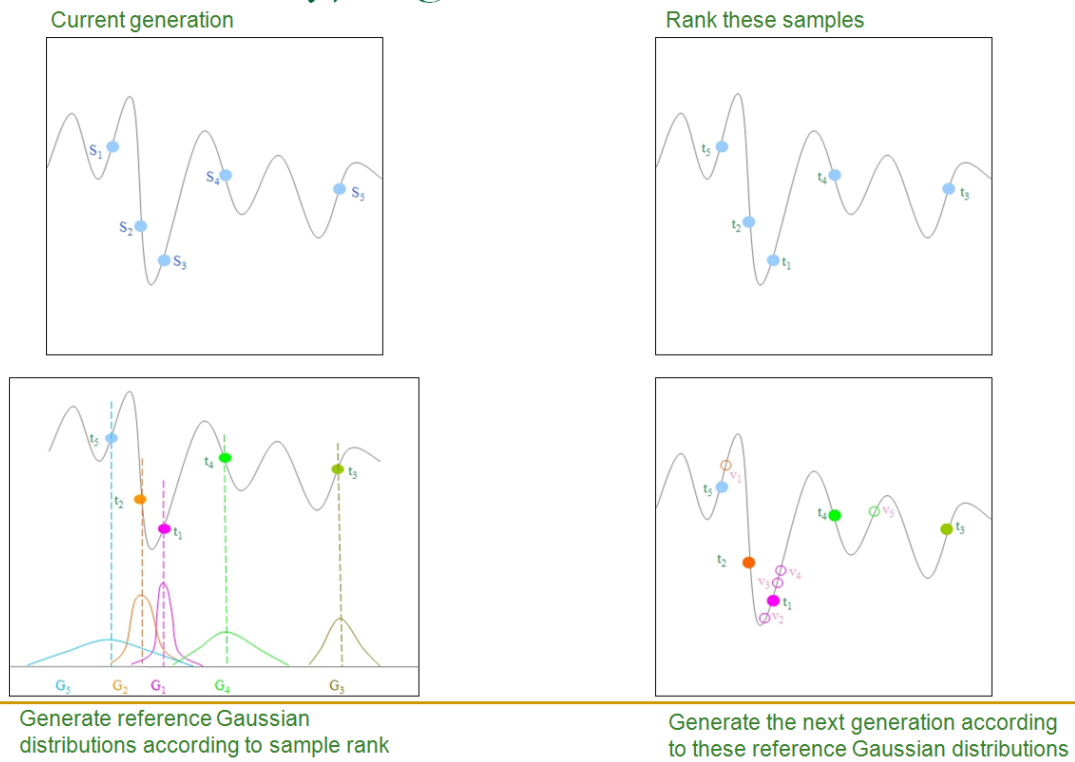
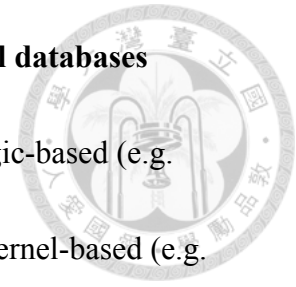


Figure 5. A diagram to illustrate the concept of RAME

(http://zoro.ee.ncku.edu.tw/bp/res/10-machine_learning.pdf)

2.4 Application of machine learning algorithms to large medical databases



Machine learning techniques fall into two main categories: logic-based (e.g. decision tree, association rule mining and Bayesian network) and kernel-based (e.g. artificial neural network and support vector machine). The former methods provide an overall picture regarding the distributions of data, while the latter deliver better evaluation performance but at the cost of higher complexities and longer computation time.

Scientists have tried using machine learning approaches to extract precious information and to discover valuable knowledge from large medical databases.³¹ The following are some applications based on different algorithms.

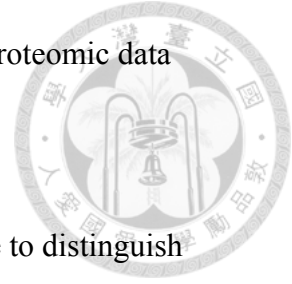
A. Decision tree: See section 2.3.1.

B. Association rule mining: This method has been applied to analyze co-prescription patterns for antacids and to uncover psychiatric comorbidities associated with attention deficit/hyperactivity disorder.^{32,33}

C. Bayesian network: It has helped predict heart disease and identify comorbidity between chronic obstructive pulmonary disease and asthma.^{34,35}

D. Artificial neural network: One recent study collected clinical information and genomic data as features, and then applied the artificial neural network method to construct a prognostic prediction model for diffuse large B-cell lymphoma.³⁶

Lancashire *et al.* utilized the same technique for the analysis of proteomic data related to breast cancer cell lines obtained from patients.³⁷



E. Support vector machine: Support vector machine has been able to distinguish cases of ovarian cancer from non-cancer controls using serum proteomic patterns as biomarkers,³⁸ discriminate malignant melanoma from dysplastic nevus by digital images of skin lesions,³⁹ determine the efficacy of interferon treatment for chronic hepatitis C according to clinical markers,⁴⁰ and detect the movements of patients as context-aware agents.⁴¹



Chapter 3

Materials and Methods



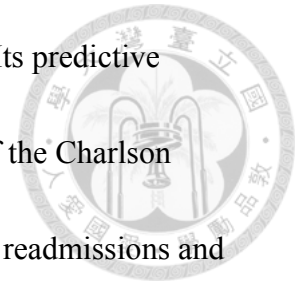
3.1 Case definition and control selection

The version of the NHIRD used in our study contains the claim records of 1 million randomly selected enrollees from all enrollees in the year 2005. Patients who were older than 18 and diagnosed as herpes zoster in outpatient and/or inpatient records from the NHIRD during 2004 to 2008 were defined as “cases”. The ICD-9 CM codes used for diagnosis of HZ include 053.xx. In this case-control study, for each HZ case, three insurants without any record of HZ during 1996 to 2010 were randomly selected from the NHIRD and the gender and age were matched with those of the corresponding case. The index date for a case was defined to be the date when the patient was diagnosed with herpes zoster for the first time and the same index date was assigned to its corresponding controls. In this study, we attempted to identify the risk factors that occurred during the period up to 24 months prior to the index date.

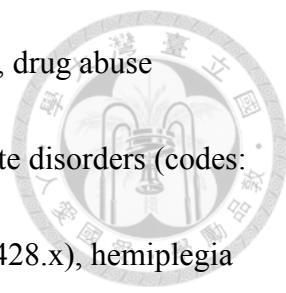
3.2 Diseases utilized as potential risk factors

In this study, we included a total of 35 disorder conditions in the Charlson Index and in the Elixhauser Index as potential risk factors.^{42,43} The Charlson index was developed originally as a prognostic indicator for estimating risk of death on the basis of patients with a variety of conditions admitted to a general medical service and then

validated in an independent cohort of women with breast cancer.⁴² Its predictive validity was confirmed by finding many significant relationships of the Charlson index with various criterion outcomes, such as mortality, disability, readmissions and length of stay.⁴⁴ It has been used subsequently to account for the impact of comorbid conditions on studies of diseases such as ischemic stroke and intracerebral hemorrhage.^{45,46} The Elixhauser index was a comprehensive set of 30 comorbidity measures and it was associated with substantial increases in length of stay, hospital charges, and mortality, both for heterogeneous and homogeneous disease groups.⁴³



This index has been used as a predictor of early postoperative outcomes after laparoscopic Roux-en-Y gastric bypass.⁴⁷ Altogether, the 35 disorder conditions considered as potential risk factors included AIDS (ICD-9 CM code: 042), alcohol abuse (codes: 265.2, 291.xx, 303.xx, 305.0x, 357.5, 425.5, 535.3x, 571.0, 571.1, 571.3, 980.x, and V11.3), blood loss anemia (code: 280.0), cardiac arrhythmias (codes: 426.0, 426.1x, 426.7, 426.9, 427.0-427.4x, 427.6x, 427.8x, 427.9, 785.0, 996.01, 996.04, V45.0x, and V53.3x), cerebrovascular diseases (codes: 430-438.xx), chronic pulmonary disease (CPD) (codes: 490-496, 500-505, and 506.4), coagulopathy (codes: 286.x, 287.1, and 287.3-287.5), coronary heart disease (CHD) (codes: 411.1, 411.8x, and 413.xx -414.xx), deficiency anemia (codes: 280.1-280.9, and 281.x), dementia (codes: 290.xx and 331.0), depression (codes: 296.2x, 296.3x, 296.5x, 300.4, and 311),



diabetes mellitus and chronic complications (codes: 250.0x-250.9x), drug abuse (codes: 292.xx, 304.xx, 305.2x-305.9x, and V65.42), fluid electrolyte disorders (codes: 253.6 and 276.x), heart failure (codes: 402.x1, 404.x1, 404.x3, and 428.x), hemiplegia and/or paraplegia (codes: 342.xx and 344.1), hyperlipidemia (codes: 272.x), hypertension (codes: 401.x), hypothyroidism (codes: 243 and 244.x), leukemia and/or lymphoma (codes: 200.xx-208.xx), malignancy (codes: 140.x-172.x, and 174.x-195.x) (including primary malignant neoplasm of all organ systems except skin), metastatic solid tumor (codes: 196.x-199.x), mild/moderate/severe liver disease (codes: 571.2, 571.4-571.9, 573, 572.2-572.8, and 456.0-456.2), myocardial infarction (codes: 410.xx and 412), obesity (codes: 278.0x), peptic ulcer disease (codes: 531.xx-534.xx), peripheral vascular disease (codes: 443.xx, 785.4, and V43.4), psychoses (codes: 293.8x, 295.xx, 297.x, and 298.x), renal disease (codes: 582.xx, 583.xx, and 588.x), renal failure (codes: 403.x1, 404.x2, 404.x3, 585, 586, V42.0, V45.1, and V56.xx), rheumatologic disease (RD) (codes: 446.x, 701.0, 710.x, 711.2x, and 720.xx), rheumatoid arthritis (codes: 714.0-714.2, 714.3x, 714.4, 714.8x, 714.9, and 719.3x), and weight loss (codes: 260-263.x, 783.2, 783.21, and 799.4).

For each subject, all the outpatient and inpatient claims during the study period were examined and the counts of outpatient and/or inpatient visits during the study period were used as feature values. As a result, each cohort subject was associated

with a 35-dimensional feature vector.



3.3 Feature selection by linear discriminant analysis

With 35 potential risk factors, we then applied linear discriminant analysis (LDA),⁴⁸ which is a well-known multivariate analysis in statistics, to identify the most relevant comorbidities of HZ to be included in the following in-depth analyses. In this respect, each subject was associated with a 35-dimensional feature vector composed of the counts of outpatient or inpatient visits due to the 35 disorder conditions during the 24-month study period. LDA aims to identify a hyperplane in the feature vector space that can separate the two classes of subjects, i.e. the cases and the controls. The hyperplane identified by the LDA algorithm is the one that yields the maximal ANOVA score when the linear discriminant function corresponding to the normal vector of the hyperplane is applied to the subjects.

With the separating hyperplane identified, each coefficient in the linear discriminant function then represents the effectiveness of the corresponding feature with respect to discriminating the two classes of subjects.

3.4 Machine learning algorithms

In our study, we applied two machine learning algorithms to identify groups of subjects with high risk of developing HZ. The two machine learning algorithms employed were decision tree and linear component analysis. The motivation to apply

two different machine learning algorithms was to exploit the powers featured by alternative machine learning algorithms.



3.4.1 Decision tree

In our study, the QUEST package (IBM SPSS Statistics 18.0) was employed to construct the decision tree.²⁹ The QUEST algorithm uses the Pearson contingency table χ^2 -test or F-test for the independence between the class variable and each discriminate variable. If the smallest p-value is less than a predefined threshold, the corresponding discriminate variable is selected. Then this is followed by the CRIMCOORD transformation for categorical variables to remove the bias in variable selection. In addition, it utilizes the linear combination split to yield a shorter tree for faster computation. For our research, we set the minimal size for the parent node as 100, and the minimal size for the child node as 50. The maximal depth for the tree was set to 5, and 10-fold cross-validation was used for optimization. F-test was issued for the independence between the class variable and each discriminate variable; p-value < 0.05 represents statistical significance.

3.4.2 Linear component analysis and RAME

The linear component analysis invoked an optimization algorithm (RAME) to figure out a set of weights, w_1, w_2, \dots, w_k , for the k features selected by LDA along with a threshold θ that maximize

$$\frac{(\# \text{ of cases with } \sum_{i=1}^k w_i x_i \geq \theta) / (\# \text{ of controls with } \sum_{i=1}^k w_i x_i \geq \theta)}{(\# \text{ of cases with } \sum_{i=1}^k w_i x_i < \theta) / (\# \text{ of controls with } \sum_{i=1}^k w_i x_i < \theta)},$$



where x_1, x_2, \dots, x_k are the values of the k features associated with a subject.

3.5 Statistical analysis

The statistical significance of differences between HZ patients and the controls in prevalence and counts of visits for each comorbid disorder was evaluated by p-value using the chi-square test or student T-test when appropriate. $P < 0.05$ was considered statistically significant. We used OR and 95% CI to denote the statistical significance of decision tree and LCA by the logistic regression. In each leaf node of the decision tree, we defined OR as the ratio of (*samples of HZ/ samples of control*) in that leaf node and (*all the other samples of HZ/ all the other samples of control*) not in that particular node.

Chapter 4

Results



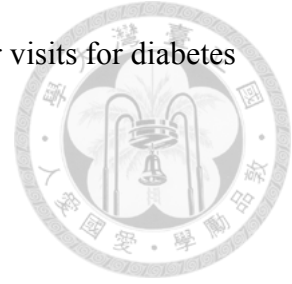
4.1 Demographic data

Based on the criteria provided in the previous section, a total, 25,863 herpes zoster patients were identified. Among them, 53.9% were female and the remaining 46.1% were male. The mean age of HZ patients was 51.7 years with a standard deviation of 17.3. With age and sex matched, 77,589 controls were randomly selected.

4.2 Univariate analysis

Our investigation began with the conventional univariate analysis to compare the prevalence of 35 comorbid disorders between the cases and the controls during the 24 months prior to the index date. As shown in Table 2, the case and control groups were significantly different in the prevalence of almost all comorbid disorders except the following 6: alcohol abuse, blood loss anemia, dementia, hemiplegia/ paraplegia, moderate to severe liver disease, and obesity. The most common comorbidities of HZ patients were hypertension (24.8%), followed by peptic ulcer disease (18.6%) and hyperlipidemia (17%). We also compared the counts of visits during the study period for management of the individual comorbid disorders. Within the 35 disorders, HZ patients made significant more visits than controls for CPD, CHD, deficiency anemia, hypertension, leukemia/ lymphoma, malignancy, metastatic solid tumor, renal disease,

renal failure, RA, and RD. On the contrary, HZ patients made fewer visits for diabetes chronic complications and fluid/ electrolyte disorders (Table 2).



4.3 Linear discriminant analysis

Before carrying out further in-depth analyses, we applied the LDA algorithm to eliminate the less relevant comorbid disorders. The last column in Table 2 lists the corresponding LDA coefficients of the 35 comorbid disorders. We then included only the following eight disorders with largest coefficients in the subsequent analyses:

arrhythmia, CPD, CHD, deficiency anemia, malignancy, renal failure, RA, and RD.

LDA took both patient numbers and counts of visits into account and probably that was the reason why arrhythmia got a high correlation coefficient though its counts of visits was not significantly different between cases and controls.

Table 2. Patient numbers, counts of visits and correlation coefficient by linear discriminant analysis of 35 groups of diseases

| Diseases | Patient Numbers | | | Counts of Visits | | | LDA Correlation coefficient |
|--------------------------------|--------------------------|-----------------------------|---------|------------------|-----------------|---------|--------------------------------|
| | Case (%) (n = 25,863) | Control (%) (n = 77,589) | P-value | Case (SD) | Control (SD) | P-value | |
| AIDS | 20 (0.1) | 22 (0) | <0.001 | 9.273 (10.484) | 14.364 (9.54) | 0.099 | .062 |
| Alcohol Abuse | 201 (0.8) | 557 (0.7) | 0.333 | 3.734 (5.161) | 4.164 (7.211) | 0.424 | .002 |
| Blood Loss Anemia | 100 (0.4) | 253 (0.3) | 0.148 | 2.5 (4.564) | 2.783 (4.321) | 0.573 | .033 |
| Cardiac Arrhythmias | 1659 (6.4) | 3863 (5) | <0.001 | 7.569 (10.166) | 7.378 (9.863) | 0.504 | .465 |
| Cerebrovascular Disease | 1826 (7.1) | 5153 (6.6) | 0.02 | 14.205 (20.497) | 14.873 (21.296) | 0.236 | .029 |
| Chronic Pulmonary Disease | 4292 (16.6) | 10267 (13.2) | <0.001 | 7.931 (13.882) | 7.007 (11.993) | <0.001 | .398 |
| Coagulopathy | 113 (0.4) | 256 (0.3) | 0.012 | 3.566 (6.501) | 3.896 (9.096) | 0.718 | .061 |
| Coronary Heart Disease | 2903 (11.2) | 7072 (9.1) | <0.001 | 10.732 (12.56) | 9.77 (11.057) | <0.001 | .387 |
| Deficiency Anemia | 395 (1.5) | 896 (1.2) | <0.001 | 5.014 (7.597) | 3.848 (5.563) | 0.002 | .406 |
| Dementia | 332 (1.3) | 1017 (1.3) | 0.74 | 10.463 (14.368) | 11.235 (14.654) | 0.391 | -.025 |
| Depression | 718 (2.8) | 1749 (2.3) | <0.001 | 9.631 (11.058) | 9.866 (10.984) | 0.625 | .218 |
| Diabetes Chronic Complications | 861 (3.3) | 2079 (2.7) | <0.001 | 11.46 (12.893) | 12.549 (13.736) | 0.043 | .084 |
| Diabetes Mellitus | 3364 (13) | 8453 (10.9) | <0.001 | 15.265 (14.564) | 14.734 (14.385) | 0.068 | .309 |
| Drug Abuse | 52 (0.2) | 100 (0.1) | 0.009 | 2.885 (3.869) | 3.529 (4.785) | 0.401 | .050 |
| Fluid Electrolyte Disorders | 583 (2.3) | 1448 (1.9) | <0.001 | 1.7 (1.928) | 2.054 (3.041) | 0.007 | .026 |
| Heart Failure | 817 (3.2) | 2006 (2.6) | <0.001 | 7.788 (9.812) | 7.569 (9.576) | 0.576 | .136 |
| Hemiplegia/ Paraplegia | 147 (0.6) | 452 (0.6) | 0.795 | 8.316 (14.907) | 9.34 (16.494) | 0.494 | -.021 |
| Hyperlipidemia | 4408 (17) | 10729 (13.8) | <0.001 | 8.782 (9.939) | 8.676 (9.905) | 0.543 | .332 |
| Hypertension | 6416 (24.8) | 17255 (22.2) | <0.001 | 12.744 (11.948) | 12.39 (11.613) | 0.036 | .308 |
| Hypothyroidism | 414 (1.6) | 848 (1.1) | <0.001 | 6.386 (8.542) | 6.559 (8.703) | 0.734 | .269 |
| Leukemia/ Lymphoma | 119 (0.5) | 107 (0.1) | <0.001 | 28.524 (32.061) | 18.046 (21.77) | 0.004 | .338 |
| Malignancy | 1353 (5.2) | 2860 (3.7) | <0.001 | 22.827 (26.991) | 17.047 (20.053) | <0.001 | .511 |
| Metastatic Solid Tumor | 299 (1.2) | 472 (0.6) | <0.001 | 8.263 (12.342) | 5.911 (9.833) | 0.003 | .291 |
| Mild Liver Disease | 2655 (10.3) | 6790 (8.8) | <0.001 | 5.944 (9.261) | 6.039 (9.515) | 0.656 | .142 |
| Moderate/ Severe Liver Disease | 72 (0.3) | 229 (0.3) | 0.665 | 5.273 (7.837) | 4.397 (7.515) | 0.379 | .021 |
| Myocardial Infarction | 210 (0.8) | 513 (0.7) | 0.012 | 7.652 (10.025) | 6.472 (9.124) | 0.117 | .109 |
| Obesity | 81 (0.3) | 206 (0.3) | 0.207 | 4.552 (7.023) | 5.535 (7.434) | 0.292 | .002 |
| Peptic Ulcer Disease | 4804 (18.6) | 11486 (14.8) | <0.001 | 6.093 (8.018) | 6.004 (8.21) | 0.514 | .370 |
| Peripheral Vascular Disease | 345 (1.3) | 778 (1) | <0.001 | 3.958 (5.784) | 4.196 (6.162) | 0.536 | .070 |
| Psychoses | 147 (0.6) | 575 (0.7) | 0.004 | 17.047 (15.771) | 18.686 (17.566) | 0.299 | -.199 |
| Renal Disease | 960 (3.7) | 2212 (2.9) | <0.001 | 14.511 (21.147) | 12.23 (18.632) | 0.002 | .262 |
| Renal Failure | 171 (0.7) | 347 (0.4) | <0.001 | 8.821 (13.53) | 4.675 (7.487) | <0.001 | .488 |
| Rheumatoid Arthritis | 1013 (3.9) | 2298 (3) | <0.001 | 5.271 (9.035) | 4.525 (8.463) | 0.018 | .469 |
| Rheumatologic Disease | 584 (2.3) | 1029 (1.3) | <0.001 | 12.656 (20.117) | 7.91 (13.549) | <0.001 | .410 |
| Weight Loss | 184 (0.7) | 440 (0.6) | 0.009 | 2.431 (3.87) | 2.36 (3.612) | 0.818 | .154 |

LDA, linear discriminant analysis; SD, standard deviation

4.4 Decision tree

As mentioned earlier, the first machine learning algorithm exploited to carry out in-depth analyses in our study was the decision tree algorithm. Accordingly, the counts of visits of the above eight disorders were input to the QUEST software to construct the decision tree. Figure 6 shows the decision tree output by the QUEST software. The decision tree contains five levels with ten leaf nodes. Among the eight disorders, only malignancy, RD, CHD and CPD were incorporated to generate the tree by the program.

The overall observation about the decision tree shown in Figure 6 is that people with a relatively healthier condition are less likely to suffer HZ than those with a relatively poorer condition, which is consistent with the general impression. If we compare the ORs between any two branches, then we will find that the OR of the branch corresponding to the higher visit count of the disorder is larger. Furthermore, leaf node 17 in the decision tree represents those people with the healthiest condition and node 17 is the only leaf node with an OR smaller than 1 while leaf node 10 (OR 5.13, 95% CI 3.10-8.49) and node 6 (OR 3.29, 95% CI 2.45-4.42) contain two subgroups of cases who are extremely vulnerable to HZ. Table 3 ranks the leaf nodes in Figure 6 by their ORs and presents the characteristics of the counts of visits of each leaf node.

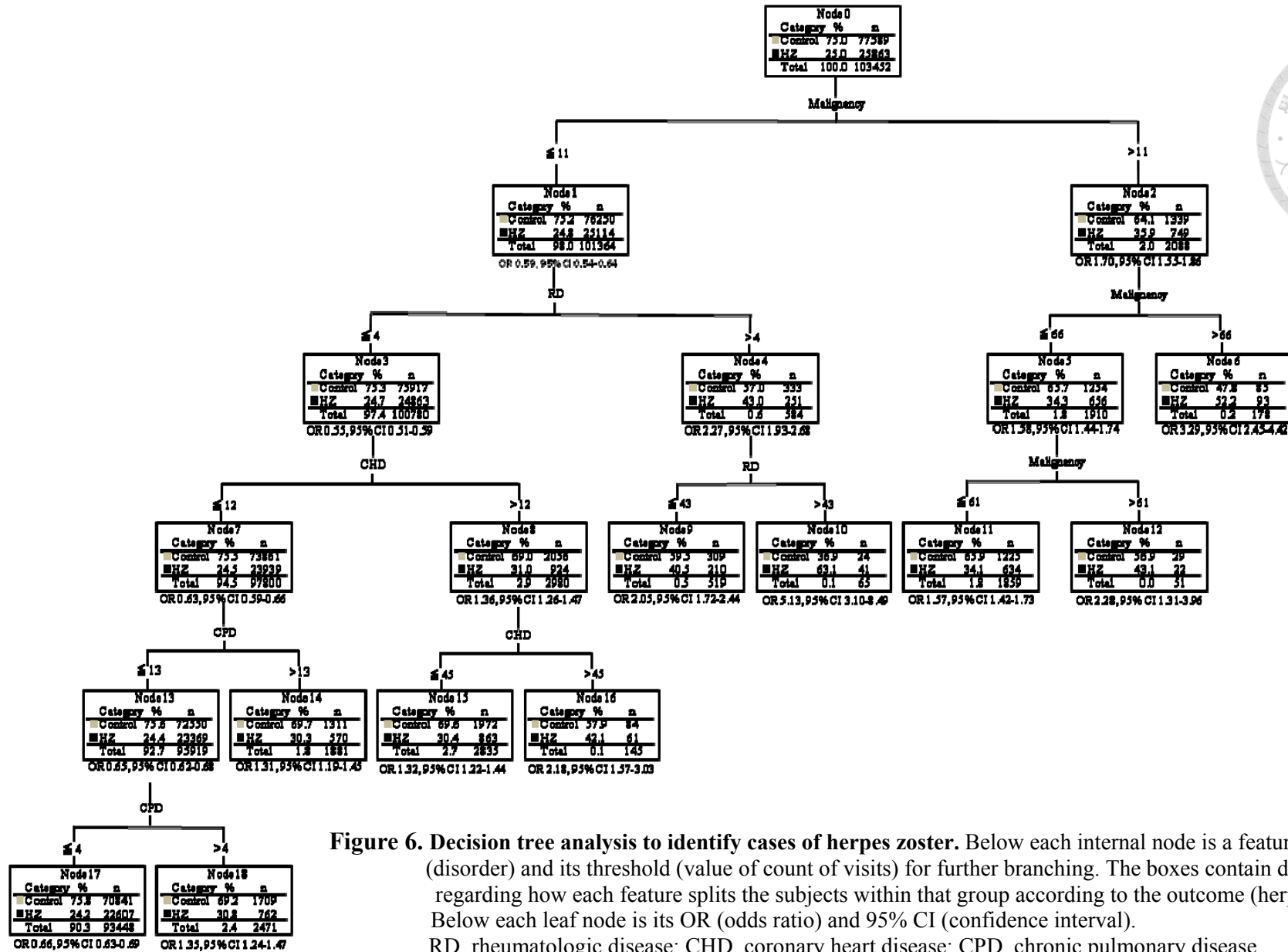


Figure 6. Decision tree analysis to identify cases of herpes zoster. Below each internal node is a feature (disorder) and its threshold (value of count of visits) for further branching. The boxes contain data regarding how each feature splits the subjects within that group according to the outcome (herpes zoster). Below each leaf node is its OR (odds ratio) and 95% CI (confidence interval). RD, rheumatologic disease; CHD, coronary heart disease; CPD, chronic pulmonary disease

Table 3. Characteristics of leaf nodes obtained by decision tree algorithm*

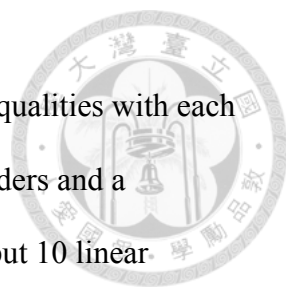
| Node No.# | HZ (n) | Control (n) | Counts of Visits | | | | OR | 95% CI |
|-----------|--------|-------------|------------------|----------|-----------|----------|------|-----------|
| | | | Malignancy | RD | CHD | CPD | | |
| 10 | 41 | 24 | ≤ 11 | >43 | - | - | 5.13 | 3.10-8.49 |
| 6 | 93 | 85 | >66 | - | - | - | 3.29 | 2.45-4.42 |
| 12 | 22 | 29 | ≤66 & >61 | - | - | - | 2.28 | 1.31-3.96 |
| 16 | 61 | 84 | ≤ 11 | ≤4 | >45 | - | 2.18 | 1.57-3.03 |
| 9 | 210 | 309 | ≤ 11 | ≤43 & >4 | - | - | 2.05 | 1.72-2.44 |
| 11 | 634 | 1,225 | ≤61 & >11 | - | - | - | 1.57 | 1.42-1.73 |
| 18 | 762 | 1,709 | ≤ 11 | ≤4 | ≤ 12 | ≤13 & >4 | 1.35 | 1.24-1.47 |
| 15 | 863 | 1,972 | ≤ 11 | ≤4 | ≤45 & >12 | - | 1.32 | 1.22-1.44 |
| 14 | 570 | 1,311 | ≤ 11 | ≤4 | ≤ 12 | >13 | 1.31 | 1.19-1.45 |
| 17 | 22,607 | 70,841 | ≤ 11 | ≤4 | ≤ 12 | ≤4 | 0.66 | 0.63-0.69 |

* In the order of ranking of odds ratio

The node number is the same as that assigned in Fig. 6.

RD, rheumatologic disease; CHD, coronary heart disease; CPD, chronic pulmonary disease; OR, odds ratio; CI, confidence interval

4.5 Linear component analysis



The results of LCA were represented by a number of linear inequalities with each inequality defined by eight weights corresponding to the eight disorders and a threshold θ . In our experiment, the software package was set to output 10 linear inequalities. Table 4 shows the 10 sets of weights along with the thresholds corresponding to the 10 inequalities obtained. Table 4 reveals that each of the 10 inequalities identifies a group of subjects (CaseLE θ) that suffered higher risk of developing HZ. CaseLE θ represented herpes zoster samples whose sum of $w_i x_i \geq \theta$, where w_i was the weight and x_i was the count of visits of the corresponding disorder. Similarly, CaseST θ represented herpes zoster samples whose sum of $w_i x_i$ was smaller than θ . ControlLE θ and ControlST θ were defined in the same way. In particular, the first inequality identifies the largest group of subjects (n=2,241) and the fifth inequality identifies a group of subjects with the highest OR (7.31 with 95% CI 4.08-13.08).

We noticed that negative weights were quite common in all inequalities. Four inequalities had only one positive weight (Table 4, in bold), i.e. the weight for RD in the fourth inequality, for CHD in the fifth, for renal failure in the seventh and ninth. We analyzed the relationships between the signs of the weights, the number of subjects with the comorbid disease and the counts of visits for that disorder. Table 5 showed two examples of the analyses (inequalities 1 and 4). We found that only those disorders with significantly more victims and/or more counts of visits for the subjects of (CaseLE θ + ControlLE θ) than for (CaseST θ + ControlST θ) had positive weights. Interestingly, we also found that these four inequalities had higher ORs than the other five inequalities.



Table 4. Results of linear component analysis

| No. | Weights* | θ | CaseLE θ | CaseST θ | ControlLE θ | ControlST θ | OR | 95% CI |
|-----|---|----------|-----------------|-----------------|--------------------|--------------------|------|------------|
| 1 | -0.663076 0.62682 0.484115 0.0218818 0.209204 -0.144505 -0.539232 0.505844 | 0.308634 | 2,241 | 23,622 | 4,573 | 70,762 | 1.47 | 1.39-1.55 |
| 2 | -0.566942 0.0630207 0.185095 -0.661184 -0.348979 -0.605029 -0.881405 -0.910154 | 0.722892 | 672 | 25,191 | 1,049 | 75,503 | 1.92 | 1.74-2.12 |
| 3 | -0.460677 -0.143162 0.860286 -0.040315 -0.600146 0.661916 -0.265908 0.123447 | 0.595752 | 1,627 | 24,236 | 3,387 | 72,571 | 1.44 | 1.35-1.53 |
| 4 | -0.27604 -0.573412 0.16007 -0.106113 -0.484359 -0.697317 -0.381939 -0.259072 | 0.86993 | 157 | 25,706 | 149 | 77,056 | 3.16 | 2.52-3.95 |
| 5 | 0.0178533 -0.0358592 -0.379131 -0.702567 -0.0739463 -0.730277 -0.0621662 -0.32725 | 0.739616 | 39 | 25,824 | 16 | 77,445 | 7.31 | 4.08-13.08 |
| 6 | -0.347209 0.338969 0.176 -0.961913 0.153172 0.239845 -0.924131 0.104587 | 0.932615 | 1,439 | 24,424 | 2,737 | 73,144 | 1.57 | 1.47-1.68 |
| 7 | -0.1948 -0.326151 -0.487472 -0.148839 -0.294412 -0.772027 -0.684927 0.357341 | 0.520249 | 73 | 25,790 | 49 | 77,316 | 4.47 | 3.11-6.42 |
| 8 | -0.99414 0.319437 -0.401349 -0.772759 -0.435286 -0.98352 -0.132786 0.900998 | 0.42967 | 977 | 24,886 | 1,861 | 74,552 | 1.57 | 1.45-1.7 |
| 9 | -0.984741 -0.484115 -0.644032 -0.146214 -0.825739 -0.833247 -0.859249 0.923643 | 0.782037 | 89 | 25,774 | 87 | 77,255 | 3.07 | 2.28-4.12 |
| 10 | -0.0912198 -0.70159 0.337016 -0.228248 0.297769 -0.215552 0.463179 -0.98822 | 0.608386 | 856 | 25,007 | 1,654 | 74,883 | 1.55 | 1.43-1.69 |

*The order of the weights: CHD_Malignancy_RD_CPD_Deficiency Anemia_Arrhythmias_RA_Renal Failure
 CHD, coronary heart disease; RD, rheumatologic disease; CPD, chronic pulmonary disease; RA, rheumatoid arthritis; OR, odds ratio; CI, confidence interval
 CaseLE θ and CaseST θ represented the number of herpes zoster samples whose sum of $w_i x_i \geq \theta$ and $< \theta$ respectively. ControlLE θ and ControlST θ represented
 number of control samples whose sum of $w_i x_i \geq \theta$ and $< \theta$ respectively, where w_i was the weight and x_i was the count of visits of the corresponding disorder.



Table 5. Analyses of weights of inequalities 1 & 4 of linear component analysis

| Inequality 1 | | CaseLE θ (n=2,241)+ ControlLE θ (n=4,573) (n=6,814) | | CaseST θ (n=23,622)+ ControlST θ (n=70,762) (n=94,384) | | CaseLE θ + ControlLE θ | | CaseST θ + ControlST θ | |
|---------------------|---------------|---|----------|---|----------|---|--------|---|--------|
| Comorbidity | Weight | Number with comorbidity | % | Number with comorbidity | % | Counts of visits | | | |
| | | | | | | Mean | SD | Mean | SD |
| CHD | -0.663076 | 513 | 7.5 | 9,780 | 10.4 | 4.54 | 6.047 | 10.34 | 11.675 |
| Malignancy | 0.62682 | 3,808 | 55.9 | 508 | 0.5 | 20.61 | 23.399 | 6.19 | 9.147 |
| RD | 0.484115 | 1,271 | 18.7 | 394 | 0.4 | 11.16 | 17.952 | 4.75 | 8.185 |
| CPD | 0.0218818 | 2,520 | 37.0 | 12,452 | 13.2 | 18.83 | 19.67 | 4.94 | 8.939 |
| Deficiency anemia | 0.209204 | 663 | 9.7 | 682 | 0.7 | 6.38 | 7.618 | 2.09 | 3.516 |
| Arrhythmias | -0.144505 | 440 | 6.5 | 5,254 | 5.6 | 4.82 | 7.337 | 7.66 | 10.114 |
| RA | -0.539232 | 242 | 3.6 | 3,220 | 3.4 | 5.39 | 8.854 | 4.71 | 8.636 |
| Renal Failure | 0.505844 | 359 | 5.3 | 185 | 0.2 | 8.02 | 11.887 | 2.3 | 2.482 |
| Inequality 4 | | CaseLE θ (n=157)+ ControlLE θ (n=149) (n=306) | | CaseST θ (n=25,706)+ ControlST θ (n=77,056) (n=102,762) | | CaseLE θ + ControlLE θ | | CaseST θ + ControlST θ | |
| Comorbidity | Weight | Number with comorbidity | % | Number with comorbidity | % | Counts of visits | | | |
| | | | | | | Mean | SD | Mean | SD |
| CHD | -0.27604 | 29 | 9.5 | 10,264 | 10 | 4.66 | 6.493 | 10.07 | 11.537 |
| Malignancy | -0.573412 | 4 | 1.3 | 4,312 | 4.2 | 1.5 | 1 | 18.93 | 22.686 |
| RD | 0.16007 | 306 | 100 | 1,359 | 1.3 | 28.82 | 23.5 | 5.32 | 10.21 |
| CPD | -0.106113 | 55 | 18 | 14,917 | 14.5 | 5.31 | 6.968 | 7.29 | 12.606 |
| Deficiency anemia | -0.484359 | 7 | 2.3 | 1,338 | 1.3 | 2.43 | 1.813 | 4.22 | 6.295 |
| Arrhythmias | -0.697317 | 11 | 3.6 | 5,683 | 5.5 | 4.55 | 7.076 | 7.44 | 9.96 |
| RA | -0.381939 | 41 | 13.4 | 3,421 | 3.3 | 3.93 | 5.091 | 4.77 | 8.686 |
| Renal Failure | -0.259072 | 6 | 2 | 538 | 0.5 | 4.17 | 4.708 | 6.1 | 10.174 |

CHD, coronary heart disease; RD, rheumatologic disease; CPD, chronic pulmonary disease; RA, rheumatoid arthritis; SD, standard deviation . CaseLE θ and CaseST θ represented the number of herpes zoster samples whose sum of $w_i x_i \geq \theta$ and $< \theta$ respectively. ControlLE θ and ControlST θ represented number of control samples whose sum of $w_i x_i \geq \theta$ and $< \theta$ respectively, where w_i was the weight and x_i was the count of visits of the corresponding disorder.

4.6 Analysis of cases identified by decision tree and/or by linear component analysis



Since the motivation to invoke two different machine learning algorithms in our study was to exploit the distinctive characteristics of alternative algorithms, it is of interest to investigate whether this objective has been achieved or not. We analyzed the intersections of cases in each leaf node of DT and cases in “CaseLE θ ” identified in each inequality of LCA (Table 6). We defined that “exclusion” existed in that intersection if there was less than 3% overlap with those cases identified by the corresponding DT leaf node and LCA inequality (Table 7). In the following discussion, we will examine whether both algorithms identified groups of subjects with distinctive characteristics. In particular, our discussion will focus on the groups of subjects with an OR higher than 3.

4.6.1 Leaf nodes 6 & 10 of DT

Firstly, if we examine the decision tree shown in Fig. 6, we can find that node 10 and node 6 are the only two leaf nodes with an OR higher than 3. Further analyses showed that the group of subjects identified by leaf node 10 is actually a subset of the groups of subjects identified by inequalities 3 and 10 shown in Table 6. Furthermore, most subjects identified by leaf node 10 are also included in the groups of subjects identified by inequalities 1, 2, 4 and 6. Though most subjects identified by leaf node 10 are also included in the groups of subjects identified by inequality 4, the corresponding OR of leaf node 10 (=5.13) is higher than that of inequality 4 (=3.16). Nevertheless, the ORs corresponding to the groups of subjects identified by inequalities 1, 2, 3, 6 and 10 are all lower than 2 (Table 4). It means that leaf node 10 identifies a group of subjects with a higher risk of developing HZ than any groups of subjects identified by LCA.

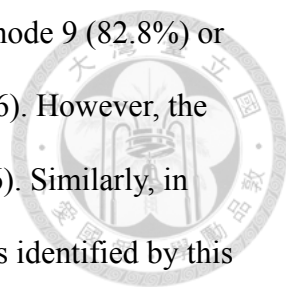
A detailed analysis of leaf node 6 showed that this group of subjects is actually a subset of the group of subjects identified by inequality 1 (Table 6). Furthermore, most subjects identified by leaf node 6 are also included in the groups of subjects identified by inequalities 2, 6 and 8. Nevertheless, the ORs corresponding to the groups of subjects identified by inequalities 1, 2, 6, and 8 are all lower than 2 (Table 4), while the OR corresponding to leaf node 6 is 3.29 (Table 3). It means that leaf node 6 really identifies a group of subjects sharing some distinctive characteristics that led to high risk of developing HZ.

4.6.2 Inequalities 4, 5, 7 and 9 of LCA

Table 4 shows that inequalities 4, 5, 7, and 9 identified groups of subjects with a corresponding OR higher than 3. An in-depth analysis showed that most of the subjects identified by inequality 4 are included in the group of subjects identified by leaf node 9 and most of the subjects identified by inequality 5 are included in the group of subjects identified by leaf node 16 (Table 6). Nevertheless, both the ORs corresponding to leaf nodes 9 and 16 are lower than 3 (Table 3), while the ORs corresponding to the groups of subjects identified by inequalities 4 and 5 are 3.16 and 7.31 (Table 4), respectively. Concerning those groups of subjects identified by inequalities 7 (n=73) and 9 (n=89), our analysis do not intersect with any group of subjects identified by the leaf nodes in the decision tree (Table 7), i.e. at least 89 cases of HZ recognized by LCA are unable to be identified by DT method. After further analysis, we notice that the only one positive weight in inequalities 7 and 9 is for renal failure and subjects recognized by both inequalities are mainly patients of renal failure, a disorder not adopted for constructing the decision tree.

4.6.3 DT and LCA validates each other

In inequality 4, only the weight for RD is positive and all the subjects identified



by this inequality ($n=157$) are also recognized by DT, either in leaf node 9 (82.8%) or 10 (17.2%) (Table 6), whose discriminator is the count of RD (Fig. 6). However, the OR of leaf node 10 ($=5.13$) is higher than that of inequality 4 ($=3.16$). Similarly, in inequality 5, only the weight for CHD is positive and all the subjects identified by this inequality ($n=39$) are also recognized by DT, either in leaf node 15 (28.2%) or 16 (71.8%) (Table 6), whose discriminator is the count of CHD (Fig. 6). On the contrary, the OR of inequality 5 ($=7.31$) is much higher than that of either leaf node 15 ($=1.32$) or 16 ($=2.18$). It means a very specific subgroup of patients, especially with RD or CHD, could be identified by both DT and LCA algorithms at the same time. Besides, as many as 98.2% of cases identified by inequality 2 could be found in leaf nodes of DT.

4.6.4 DT and LCA is complementary to each other

As mentioned in section 4.6.2, LCA discovered cases of renal failure that DT was unable to identify. Besides, for example, in the first inequality of LCA with the largest group of subjects ($n=2,241$), only 61.67% of them could be identified by DT and in the second largest group (inequality 3, $n=1,627$), only 29.62% could be identified by DT (Table 6). Table 7 also showed exclusions were not uncommon between the HZ cases identified by DT and LCA. It means that some HZ cases could be identified only by either DT or by LCA and applying these two algorithms together could increase the total number of cases recognized.

Table 6. Distribution of cases of herpes zoster in the intersections of decision tree leaf nodes and LCA inequalities

| n (% in LCA; % in DT) | | No. of Leaf Node and Its Case Number by Decision Tree Algorithm | | | | | | | | |
|---|-------------|---|--------------------|-------------------|--------------------|------------------|--------------------|------------------|-------------------|-----------------|
| | | 6 (n=93) | 9 (n=210) | 10 (n=41) | 11 (n=634) | 12 (n=22) | 14 (n=570) | 15 (n=863) | 16 (n=61) | 18 (n=762) |
| No. of inequality and its case number of CaseLE θ^* by LCA | 1 (n=2,241) | 93 (4.1%; 100%) | 181 (8.1%; 86.2%) | 38 (1.7%; 92.7%) | 617 (27.5%; 97.3%) | 22 (1%; 100%) | 377 (16.8%; 66.1%) | 0 (0%; 0%) | 0 (0%; 0%) | 54 (2.4%; 7.1%) |
| | 2 (n=672) | 79 (11.8%; 84.9%) | 125 (18.6%; 59.5%) | 24 (3.6%; 58.5%) | 412 (61.3%; 65%) | 20 (3%; 90.9%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) |
| | 3 (n=1,627) | 4 (0.2%; 4.3%) | 207 (12.7%; 98.6%) | 41 (2.5%; 100%) | 19 (1.2%; 3%) | 0 (0%; 0%) | 56 (3.4%; 9.8%) | 86 (5.3%; 10%) | 2 (0.1%; 3.3%) | 67 (4.1%; 8.8%) |
| | 4 (n=157) | 0 (0%; 0%) | 130 (82.8%; 61.9%) | 27 (17.2%; 65.9%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) |
| | 5 (n=39) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 11 (28.2%; 1.3%) | 28 (71.8%; 45.9%) | 0 (0%; 0%) |
| | 6 (n=1,439) | 88 (6.1%; 94.6%) | 135 (9.4%; 64.3%) | 26 (1.8%; 63.4%) | 557 (38.7%; 87.9%) | 21 (1.5%; 95.5%) | 0 (0%; 0%) | 7 (0.5%; 0.8%) | 0 (0%; 0%) | 1 (0.1%; 0.1%) |
| | 7 (n=73) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 1 (1.4%; 0.2%) | 0 (0%; 0%) | 1 (1.4%; 0.2%) | 1 (1.4%; 0.1%) | 0 (0%; 0%) | 1 (1.4%; 0.1%) |
| | 8 (n=977) | 85 (8.7%; 91.4%) | 0 (0%; 0%) | 0 (0%; 0%) | 531 (54.4%; 83.8%) | 21 (2.1%; 95.5%) | 1 (0.1%; 0.2%) | 0 (0%; 0%) | 0 (0%; 0%) | 1 (0.1%; 0.1%) |
| | 9 (n=89) | 0 (0%; 0%) | 0 (0%; 0%) | 0 (0%; 0%) | 2 (2.2%; 0.3%) | 0 (0%; 0%) | 1 (1.1%; 0.2%) | 0 (0%; 0%) | 0 (0%; 0%) | 2 (2.2%; 0.3%) |
| | 10 (n=856) | 0 (0%; 0%) | 198 (23.1%; 94.3%) | 41 (4.8%; 100%) | 1 (0.1%; 0.2%) | 0 (0%; 0%) | 5 (0.6%; 0.9%) | 10 (1.2%; 1.2%) | 0 (0%; 0%) | 16 (1.9%; 2.1%) |

*CaseLE θ represented the number of herpes zoster samples whose sum of $W_i X_i \geq \theta$, where W_i was the weight and X_i was the count of visits of the corresponding disorder in the inequality

Table 7. Exclusions in intersections of decision tree leaf nodes and LCA inequalities

| | | No. of Leaf Node of Decision Tree | | | | | | | | |
|---------------------------------|-----------|-----------------------------------|---|----|----|----|----|----|----|----|
| | | 6 | 9 | 10 | 11 | 12 | 14 | 15 | 16 | 18 |
| No. of Inequality of LCA | 1 | | | | | | | X | X | |
| | 2 | | | | | | X | X | X | X |
| | 3 | | | | | X | | | | |
| | 4 | X | | | X | X | X | X | X | X |
| | 5 | X | X | X | X | X | X | | | X |
| | 6 | | | | | | X | X | X | X |
| | 7 | X | X | X | X | X | X | X | X | X |
| | 8 | | X | X | | | X | X | X | X |
| | 9 | X | X | X | X | X | X | X | X | X |
| | 10 | X | | | X | X | X | X | X | X |

X: exclusion existed in that intersection if there was less than 3% overlap with those cases identified by DT leaf node and the corresponding LCA inequality
 LCA, linear component analysis

Chapter 5

Discussion

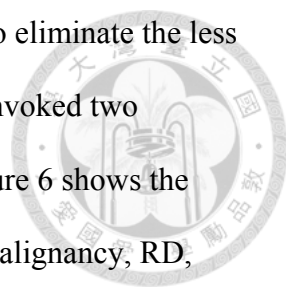


The importance of herpes zoster lies in its high prevalence, frequent intolerable neuralgia, either during or after the acute stage, and other complications like keratitis or encephalitis. Besides, a recent study reported that herpes zoster ophthalmicus patients had a 4.52-fold (95% CI, 2.45–8.33) higher risk of stroke than the matched comparison.¹⁸ However, HZ might be preventable by a vaccine licensed in 2006 but the cost of vaccination probably is not affordable for all potential victims, especially for those in developing countries.⁴ Accordingly, from the viewpoint of medical economics, identifying those patients with higher risk of developing HZ as the candidates for vaccination is an important and practical issue.

In our study, we firstly conducted a comprehensive univariate analysis to figure out whether HZ patients were generally in poorer health condition than the controls. In this respect, we considered 35 major disorders extracted from two popular health indexes, the Charlson Index and the Elixhauser Index. The results shown in Table 2 are consistent with the common impression that HZ patients were generally in a relatively “less healthy” condition. Furthermore, the risk factors of HZ which have been reported in recent articles, including HIV infection, psychiatric disorders, chronic obstructive pulmonary diseases, rheumatoid arthritis, chronic kidney disease, systemic lupus erythematosus and under hemodialysis, are all among the 29 disorders in Table 2, from which the HZ patients suffered with higher prevalence.⁵⁻¹¹ A recent retrospective hospital-based cohort study in Japan showed that patients who had one of the 14 underlying diseases, i.e., brain tumor, lung cancer, breast cancer, esophageal cancer, gastric cancer, colorectal cancer, gynecologic cancer, malignant lymphoma,

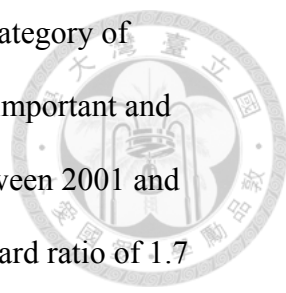
systemic lupus erythematosus, RA, diabetes mellitus, hypertension, renal failure, and disk hernia, displayed a 1.8–8.4-fold increased risk of HZ events compared to controls, which was compatible with our findings though we didn't specify the organ of malignancy and disk hernia was not surveyed.⁴⁹ Concerning the correlations between HZ and various types of cancers, quite the contrary, people in the North Carolina cohort with a history of cancer were not at increased risk of HZ after 8 years of follow-up.²⁶ Nevertheless, our analysis showed significantly higher prevalence of lymphoma/ leukemia, malignancy and metastatic solid tumor among HZ patients than the controls.

In our analysis, we further compared the counts of visits of various comorbid disorders between the HZ patients and the controls during the study period. As shown in Table 2, as expected, HZ patients had more counts of visits for cancers (including lymphoma/ leukemia, malignancy and metastatic solid tumor) and diseases possibly controlled with immunosuppressants or systemic corticosteroids (CPD, RA and RD). Besides, HZ patients visited doctors significantly more often than controls for renal disease/ renal failure, hypertension/ CHD and deficiency anemia. Those findings are also generally consistent with some reports. It has been shown that end-stage renal disease is associated with immune deficiency.⁵⁰ Danesh *et al.* reviewed a large number of studies reporting associations of CHD and certain persistent bacterial and viral infections though available evidence was still sparse.⁵¹ A matched case-control study in Spain demonstrated high herpetic burden (aggregate number of antibody seropositivities (IgG) for herpesvirus) was associated with ischemic heart disease.⁵² Ekiz *et al.* revealed that humoral, cell-mediated and nonspecific immunity and the activity of cytokines which have an important role in various steps of immunogenic mechanisms are influenced by iron deficiency anemia.⁵³



To conduct in-depth analyses, we applied the LDA algorithm to eliminate the less relevant comorbid disorders (dimensionality reduction). Then, we invoked two machine learning algorithms to carry out multivariate analyses. Figure 6 shows the decision tree output by the QUEST package, which contains only malignancy, RD, CHD and CPD, four discriminating variables out of the eight variables selected by the LDA algorithm. Except node 17, all leaf nodes corresponded to a significant OR and 95% CI. We also found that the one node containing cases with more counts of visits yielded higher OR than the other node with lower counts of the same disorder. It seemed that the counts of visits were correlated to the disease activity and subsequently to the risk of having HZ. Leaf node 10 and node 6 obtained highest ORs by DT and contained two subgroups of cases with much higher risk of HZ. Though LCA did identify some subgroups of cases of leaf nodes 6 and 10, the ORs of those inequalities were all smaller than that of leaf node 6 or 10. We found that cases in leaf node 10 may or may not suffer from malignancy but visited doctors for rheumatologic disease more than 43 times before onset of HZ and cases in leaf node 6 more than 66 visits for malignancy (Table 3). It's reasonable to postulate that those frequent-visiting patients were either under an unstable disease activity or receiving intensive treatment and therefore immunosuppressive medication for treating RD or chemotherapy/ radiotherapy for cancers made them more vulnerable to HZ.

We invoked RAME, specifically designed for simulation of protein-ligand docking initially,⁵⁴ as the optimization algorithm while implementing LCA, which identified many groups of subjects who were more likely to develop HZ, though their features could not be easily interpreted according to the output in the format of linear inequalities. Four inequalities yielded ORs higher than 3 and all of them had only one positive weight in the inequalities, either for CHD, RD or renal failure, implying the



importance of these disorders on the risk of developing HZ. In the category of rheumatologic disorders, systemic lupus erythematosus is the most important and devastating one. A prospective cohort study following 10 years between 2001 and 2010 reported that SLE patients had more HZ at all ages, with a hazard ratio of 1.7 (95% CI 1.08–2.71).⁵⁵ Increasing age and reduced functional status were other independent predictors of HZ. In SLE, prednisone and mycophenolate mofetil use conferred additional risk.⁵⁵ Though renal failure was not adopted in constructing the decision tree by the QUEST program, the results of LCA pointed out that it was worthwhile to pay attention to such patients. It has been shown that end stage renal disease is coupled with immune deficiency caused by depletion of dendritic cells, naïve and central memory T cells and B cells and impaired phagocytic function of neutrophils and monocytes.⁵⁰

A very specific subgroup of patients could be identified by both DT and LCA algorithms at the same time and therefore both algorithms were able to validate each other in some way. Though there were overlaps of cases identified by some DT leaf nodes and LCA inequalities, both algorithms demonstrated different differentiating powers by showing high ORs for various subgroups of patients (see Sections 4.6.1 and 4.6.2). On the contrary, we also noted that some extent of exclusions existed between the HZ cases identified by both algorithms. It implied that DT and LCA were able to recognize different populations of HZ cases in a complementary way and therefore union of the results obtained by both methods could possibly increase the sensitivity.

The initial motivation of this thesis was to identify individuals with higher risk of developing HZ for consideration of vaccination for them. Efforts have been devoted to the study of cost-effectiveness of vaccination against herpes zoster in Canada,

Netherlands and Belgium etc.⁵⁶⁻⁵⁸ However, such researches required the input of, in addition to outpatient and inpatient visits information, mortality data, vaccine efficacy data, costs for management of HZ and PHN, mean duration of PHN, vaccine costs and costs for execution of vaccine program (education, broadcasting, administration etc.), indirect costs like work loss, and estimated Quality-Adjusted Life-Years loss. Most of the above-mentioned information was not available from NHIRD. More investigations are needed to make a policy about zoster vaccination in Taiwan.

However, in addition to the limitations inherent to the NHIRD-based studies mentioned in section 2.2, there are some others in this study. First, the counts of visits used as features could be influenced by some factors other than the disease activity, ex. the availability of medical institutions, the amount of medicine dispensed etc. Second, 8,114 (31.4%) of HZ cases and 31,125 (40.1%) of controls didn't suffer from any of the 35 diseases we utilized as potential risk factors and therefore these two populations were unable to be differentiated with any features we used in this study. Third, though DT did recognized subgroups of cases with higher risk of developing HZ, however, only 3,256 cases were identified and 87.4% of cases either didn't suffer from any of the four diseases used in constructing the tree or were not severe enough. Fourth, as for the eight disorders selected by LDA for further analysis, only 0.48% of 25,863 HZ cases suffered from four comorbidities of them, 2.44% from three, 8.46% from two and 23.41% suffered from only one disorder. About 65% of cases were not affected by any of these diseases used in LCA analysis. Finally, we didn't take patients' medications into account while the treatment regimens might also have an effect on the risk of HZ. For example, for patients with chronic obstructive pulmonary diseases, the adjusted HR of HZ was 1.67 (95% CI 1.43–1.96) for those not taking steroid medications, 2.09 for patients using inhaled corticosteroids only (95% CI

1.38–3.16) and 3.00 for patients using oral steroids (95% CI 2.40–3.75).⁷



Chapter 6

Conclusions



By applying DT and LCA, we were able to identify a group of patients with high risk of developing HZ, by using the counts of visits for some major comorbid disorders as features. All leaf nodes except for one of DT and all inequalities of LCA corresponded to a significant OR and 95% CI, indicating that both methods were capable of identifying some higher risk sub-populations. Two leaf nodes of DT and four inequalities of LCA yielded ORs even higher than 3. Those patients were mainly victims of malignancy, rheumatologic disorder, coronary heart disorder or renal failure. We noticed a very specific subgroup of patients could be identified by both DT and LCA algorithms at the same time and LCA discovered cases of renal failure that DT was unable to recognize. Though there were overlaps of cases identified by some DT leaf nodes and LCA inequalities, some extent of exclusions also existed between the HZ cases identified by both algorithms. It implied that DT and LCA were able to recognize different populations of HZ cases in a complementary way and therefore combination of the results obtained by both methods could possibly increase the sensitivity.



Chapter 7

Future Work



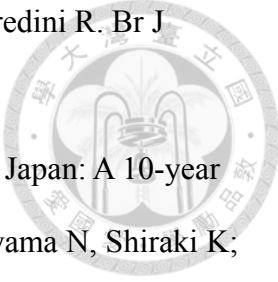
- A. Studying the influence of age on the risk of HZ, e.g., to categorize the study population by age
- B. Focusing on the patients developing PHN and specifying their risk factors
- C. Analyzing the trend of change of the general health condition during some proper time periods before and after affecting HZ, by scoring Charlson Index or other suitable indexes
- D. Observing if HZ can serve as a maker for subsequent malignancy or other major disorders by analyzing onset of new comorbidities after developing HZ
- E. Investigating the effects of medications on the risk of developing HZ
- F. Using other machine learning algorithms to analyze risk factors of HZ, e.g., association rule mining or Bayesian network
- G. Finding other "better" features, e.g., disorders other than those extracted from the Charlson Index and in the Elixhauser Index, or the time gap between the onset of the comorbidity and the onset of HZ



References

1. Epidemiological features and costs of herpes zoster in Taiwan: a national study 2000 to 2006. Jih JS, Chen YJ, Lin MW, Chen YC, Chen TJ, Huang YL, Chen CC, Lee DD, Chang YT, Wang WJ, Liu HN. *Acta Derm Venereol.* 2009;89(6):612-6.
2. Pain, medication use, and health related quality of life in older persons with postherpetic neuralgia: results from a population based survey. Oster G, Harding G, Dukes E, Edelsberg J, Cleary PD. *J Pain.* 2005;6:356-63.
3. Acute pain in herpes zoster and its impact on health-related quality of life. Katz J, Cooper EM, Walther RR, Sweeney EW, Dworkin RH. *Clin Infect Dis.* 2004;39:342-8.
4. Efficacy of live zoster vaccine in preventing zoster and postherpetic neuralgia. Gilden D. *J Intern Med.* 2011;269(5):496–506.
5. What does epidemiology tell us about risk factors for herpes zoster? Thomas SL, Hall AJ. *Lancet Infect Dis.* 2004;4:26–33.
6. Risk of herpes zoster among patients with psychiatric diseases: a population-based study. Yang YW, Chen YH, Lin HW. *J Eur Acad Dermatol Venereol.* 2011;25(4):447-53.
7. Risk of herpes zoster among patients with chronic obstructive pulmonary disease: a population-based study. Yang YW, Chen YH, Wang KH, Wang CY, Lin HW. *CMAJ.* 2011;183(5):E275-80.
8. The risk of herpes zoster in patients with rheumatoid arthritis in the United States and the United Kingdom. Smitten AL, Choi HK, Hochberg MC, Suissa S, Simon TA, Testa MA, Chan KA. *Arthritis Rheum.* 2007;57(8):1431-8.
9. Risk of Herpes Zoster in CKD: A Matched-Cohort Study Based on Administrative Data. Wu MY, Hsu YH, Su CL, Lin YF, Lin HW. *Am J Kidney Dis.*

- 2012;60(4):548-52.
10. Risk of herpes zoster in patients with systemic lupus erythematosus: a three-year follow-up study using a nationwide population-based cohort. Chen HH, Chen YM, Chen TJ, Lan JL, Lin CH, Chen DY. *Clinics (Sao Paulo)*. 2011;66(7):1177-82.
11. Risk of herpes zoster in patients treated with long-term hemodialysis: a matched cohort study. Kuo CC, Lee CT, Lee IM, Ho SC, Yang CY. *Am J Kidney Dis*. 2012;59(3):428-33.
12. Management of the patient with suspected temporal arteritis: a decision-analytic approach. Niederkoher RD, Levin LA. *Ophthalmology*. 2005;112:744-56.
13. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. Markey MK, Tourassi GD, Floyd CE, Jr. *Proteomics*. 2003;3:1678-9.
14. The burden of herpes zoster and postherpetic neuralgia in the United States. Weaver BA. *J Am Osteopath Assoc*. 2007;107(3 Suppl 1):S2-7.
15. Vaccination: a new option to reduce the burden of herpes zoster. Mick G. *Expert Rev. Vaccines*. 2010; 9(3 Suppl.):31-5.
16. Herpes zoster and quality of life: a self-limited disease with severe impact. Lydick E, Epstein RS, Himmelberger D, et al. *Neurology*. 1995;45:S52-3.
17. Measurement of pain. Katz J, Melzack R. *Surg Clin North Am* 1999;79:231-52.
18. Herpes zoster ophthalmicus and the risk of stroke: a population-based follow-up study. Lin HC, Chien CW, Ho JD. *Neurology*. 2010;74(10):792-7.
19. Herpes zoster: epidemiology, natural history, and common complications. Weinberg JM. *J Am Acad Dermatol*. 2007;57(6 Suppl):S130-5.
20. Schmader K, Studenski S, MacMillan J, Grufferman S, Cohen HJ. Are stressful life events risk factors for herpes zoster? *J Am Geriatr Soc*. 1990;38:1188-94.

- 
21. Seasonal variation in herpes zoster infection. Gallerani M, Manfredini R. *Br J Dermatol.* 2000;142(3):588-9.
 22. Epidemiology of herpes zoster and its relationship to varicella in Japan: A 10-year survey of 48,388 herpes zoster cases in Miyazaki prefecture. Toyama N, Shiraki K; Society of the Miyazaki Prefecture Dermatologists. *J Med Virol.* 2009;81(12):2053-8.
 23. Epidemiology of varicella zoster virus infection in Canada and the United Kingdom. Brisson M, Edmunds WJ, Law B, Gay NJ, Walld R, Brownell M, Roos L, De Serres G. *Epidemiol Infect.* 2001;127(2):305-14.
 24. Do the health-care workers gain protection against herpes zoster infection? A 6-year population-based study in Taiwan. Wu CY, Hu HY, Huang N, Pu CY, Shen HC, Chou YJ. *J Dermatol.* 2010;37(5):463-70.
 25. Incidence and risk factors for herpes zoster among patients with inflammatory bowel disease. Gupta G, Lautenbach E, Lewis JD. *Clin Gastroenterol Hepatol.* 2006;4(12):1483-90.
 26. Racial and psychosocial risk factors for herpes zoster in the elderly. Schmader K, George LK, Burchett BM, Pieper CF. *J Infect Dis.* 1998;178(Suppl 1):S67-70.
 27. Risk factors for postherpetic neuralgia in patients with herpes zoster. Jung BF, Johnson RW, Griffin DR, Dworkin RH. *Neurology.* 2004;62:1545-51.
 28. Association of primary cutaneous amyloidosis with atopic dermatitis: a nationwide population-based study in Taiwan. Lee DD, Huang CK, Ko PC, Chang YT, Sun WZ, Oyang YJ. *Br J Dermatol.* 2011;164(1):148-53.
 29. Split Selection Methods for Classification Tree. Loh WY, Shih YS. *Statistica Sinica* 1997;7:815-40.
 30. MEdock: a web server for efficient prediction of ligand binding sites based on a

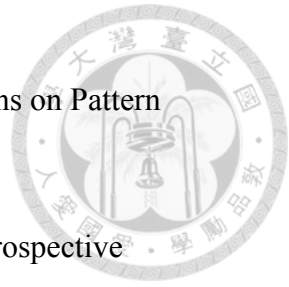
novel optimization algorithm. Chang DT, Oyang YJ, Lin JH. *Nucleic acids research* 2005;33:W233-8.



31. Application of data mining techniques to healthcare data. Obenshain MK. *Infect Control Hosp Epidemiol.* 2004;25:690-5.
32. Application of a data-mining technique to analyze co-prescription patterns for antacids in Taiwan. Chen TJ, Chou LF, Hwang SJ. *Clinical therapeutics.*2003;25:2453-63.
33. Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. Tai YM, Chiu HW. *International journal of medical informatics.* 2009;78:e75-83.
34. Intelligent heart disease prediction system using data mining techniques. Palaniappan S, Awang R. In *IEEE/ACS International Conference on Computer Systems and Applications*; March 31-April 4. 2008:108-15.
35. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF. *J Am Med Inform Assoc* 2009;16:371-9.
36. Prediction of clinical behaviour and treatment for cancers. Futschik ME, Sullivan M, Reeve A, Kasabov N. *Applied bioinformatics* 2003;2:S53-8.
37. Current developments in the analysis of proteomic data: Artificial neural network data mining techniques for the identification of proteomic biomarkers related to breast cancer. Lancashire LJ, Mian S, Ellis IO, Rees RC, Ball GR. *Current Proteomics* 2005;2:15-29.
38. Data mining techniques for cancer detection using serum proteomic profiling. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA. *Artificial intelligence in medicine* 2004;32:71-83.

39. Characterization of digital medical images utilizing support vector machines. Maglogiannis IG, Zafiropoulos EP. BMC medical informatics and decision making 2004;4:4.
40. Efficacy of interferon treatment for chronic hepatitis C predicted by feature subset selection and support vector machine. Yang J, Nugroho AS, Yamauchi K, Yoshioka K, Zheng J, Wang K, Kato K, Kuroyanagi S, Iwata A. Journal of medical systems 2007;31:117-23.
41. Patient Fall Detection using Support Vector Machines. Doukas C, Maglogiannis I, Tragas P, Liapis D, Yovanof G. In Artificial Intelligence and Innovations: from Theory to Applications. Volume 247. Edited by Boukis C, Pnevmatikakis A, Polymenakos L: Springer Boston; 2007:147-56.
42. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Charlson ME, Pompei P, Ales KL, MacKenzie CR. J Chron Dis 1987;40:373-83.
43. Comorbidity measures for use with administrative data. Elixhauser A, Steiner C, Harris DR, Coffey RM. Medical care 1998;36:8-27.
44. How to measure comorbidity. a critical review of available methods. de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. J Clin Epidemiol. 2003;56(3):221-9.
45. Charlson Index comorbidity adjustment for ischemic stroke outcome studies. Goldstein LB, Samsa GP, Matchar DB, Horner RD. Stroke. 2004;35(8):1941-5.
46. Charlson comorbidity index adjustment in intracerebral hemorrhage. Bar B, Hemphill JC 3rd. Stroke. 2011;42(10):2944-6.
47. The application of comorbidity indices to predict early postoperative outcomes after laparoscopic Roux-en-Y gastric bypass: a nationwide comparative analysis of over 70,000 cases. Shin JH, Wormi M, Castleberry AW, Pietrobon R, Omotosho

- PA, Silberberg M, Østbye T. *Obes Surg.* 2013;23(5):638-49.
48. PCA versus LDA. Martinez, A. M.; Kak, A. C. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2001;23:228–33.
49. Risk of Herpes zoster in patients with underlying diseases: a retrospective hospital-based cohort study. A. Hata, M. Kuniyoshi, Y. Ohkusa. *Infection.* 2011;39(6):537–44.
50. Effect of uremia on structure and function of immune system. Vaziri ND, Pahl MV, Crum A, Norris K. *J Ren Nutr.* 2012;22(1):149-56.
51. Chronic infections and coronary heart disease: is there a link? Danesh J, Collins R, Peto R. *Lancet.* 1997;350:430-6.
52. Association between herpetic burden and chronic ischemic heart disease: matched case-control study. Esteban-Hernández J, San Román Montero J, Gil R, Anegón M, Gil A. *Med Clin (Barc).* 2011;137(4):157-60.
53. The effect of iron deficiency anemia on the function of the immune system. Ekiz C, Agaoglu L, Karakas Z, Gurel N, Yalcin I. *Hematol J.* 2005;5(7):579-83.
54. On the design of optimization algorithms for prediction of molecular interactions. Chang DTH, Lin JH, Hsieh CH, Oyang YJ. *Int J Artif Intell Tools.* 2010;19(3):267–80.
55. Increased incidence of herpes zoster among patients with systemic lupus erythematosus. Chakravarty EF, Michaud K, Katz R, Wolfe F. *Lupus.* 2013;22(3):238-44.
56. Cost effectiveness of herpes zoster vaccine in Canada. Najafzadeh M, Marra CA, Galanis E, Patrick DM. *Pharmacoeconomics.* 2009;27(12):991-1004.
57. Assessing the potential effects and cost-effectiveness of programmatic herpes zoster vaccination of elderly in the Netherlands. van Lier A, van Hoek AJ,



Opstelten W, Boot HJ, de Melker HE. BMC Health Serv Res. 2010;10:237.

58. Cost-effectiveness of vaccination against herpes zoster in adults aged over 60 years in Belgium. Bilcke J, Marais C, Ogunjimi B, Willem L, Hens N, Beutels P. Vaccine. 2012;30(3):675-84.

