運用資料探勘技術建構健康趨勢及疾病關聯性之分析預測系統

藍國誠 [a], 李昭輝 [a], 李語嫣 [b], 吳晉祥 [c], 黎煥中 [d], 曾新穆 [a,b*]
Guo-Cheng Lan [a], Chao-Hui Lee [a], Yu-Yen Lee [b], Jin-Shang Wu [c],
Huan-Chung Li [d], Vincent S. Tseng [a,b*]
[a] *國立成功大學資訊工程研究所*
[b] *國立成功大學醫學資訊研究所*
[c] *國立成功大學醫學院醫學系家庭醫學科*
[d] *財團法人資訊工業策進會-創新應用服務研究所*
[*] *通訊作者: 曾新穆, tsengsm@mail.ncku.edu.tw*

摘要

近年來，健康相關議題日漸受到重視，民眾藉由定期健康檢查以瞭解自己的健康狀況，以便及早發現疾病並儘速就醫，避免錯過最佳治療時機，因此，健康檢查的角色也就更加重要。然而，在每次健檢後，受測者卻僅能得知當次的健檢結果報告，並無法知道自己健康的趨勢是否為高風險的趨勢。因此，本研究提出一個健康資料分析系統架構，所提的系統利用資料探勘技術來分析受測者的健檢及檢驗等歷史記錄，以獲得各項檢測項目之健康風險樣式，同時，也探討這些健康風險樣式與慢性疾病之間的關聯性，以建構更完整的預測模型。經由本系統所提供的資訊將可作為醫護人員的輔助資訊，進而對受測者提出健康警訊或適當的醫療決策，以達到及早預防及早治療的成效。

關鍵字：資料探勘、健康檢查、健康風險樣式、疾病分析、預測模型

Abstract

In recent years, people have paid much attention to the health-related issues. In order to understand clearly their health statuses, people should regularly carry out the health examination. We could early diagnose the disease and take medical treatment via the health examination, and it is possible that we could avoid missing the best treatment time. Hence, the health examination is a more important role. After completing the health examination, however, examinants only know about the result report of currently completing health examination, and they do not further know whether their health trends are high-risk or not. In this study, therefore, we proposed a novel framework for discovering health risk patterns and the relation between the health patterns and the target disease from health examination history data. The information can build effective prediction model for target disease diagnoses. According to these information provided, the physicians could early provide the health alerting and the medical treatment for people.

Keywords: Data mining, health examination, health risk pattern, disease analysis, prediction model.

1、前言

近年來，由於經濟發展，國民所得提高與醫療技術的進步，使得健康相關議題日漸受到重視，一般民眾也更加關心自身的健康狀況。由中華民國行政院衛生署的統計資料得知，在西元 2008 年台灣地區的十大死因[1]裡，將近一半以上都是屬於可以早期發現早期治療的慢性疾病，除此之外，也隨著疾病型態的轉變與罹病的年齡層也正在逐年下降[4]，因此，健康檢查的角色也就更加重要[9][11]。

由於只有透過定期健康檢查，才能瞭解自己的健康狀況，也才能及早發現疾病並儘速就醫，避免錯過最佳治療時機[3][5]，如高血壓、糖尿病、腎臟病、心臟病---等多元慢性疾病。在這樣的健檢需求的發展大前景之下，針對健檢醫療相關產品或技術的研究與發展，更是有其重要的意義與利基。

在每次健檢後，受測者僅能得知當次的健檢結果，若能可追蹤該受測者歷年來的健檢資料記錄，也許可從這些健檢結果得知其風險狀態。例如，在某受測者的

健檢記錄裡，其膽固醇、三酸甘油脂與血糖等健檢指標的數據雖未高於警戒範圍，但從該受測者的歷史健檢記錄裡，卻可觀察出其這些健檢指標的數據評估值皆具有逐年向上升的趨勢。所以，觀察受測者的歷史健檢資料，以得知其健康趨勢變化是有必要的。

隨著資料探勘(Data mining)的技術蓬勃發展，促使相關醫療系統也以轉型為智慧型系統為目標，如生理訊號自動判讀，慢性病患的基本照護等智慧型系統[16]。因此，本研究提出以資料探勘為發展技術，針對健檢資料進行分析，藉以獲得各個健檢指標之健康風險樣式；同時，也將進一步分析健康風險樣式與某些疾病之間的相關性，這是過去較少被探討的部份。Figure 1 為我們所提出的個人健康管理系統的基本架構圖。

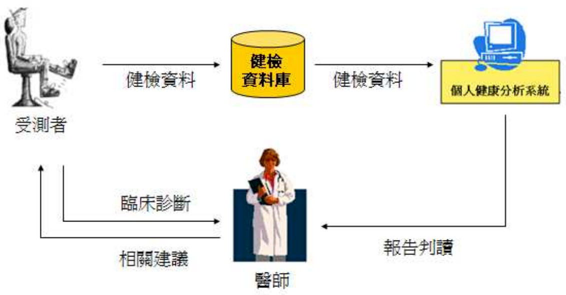

Figure 1 個人健康管理系統環境示意圖

由於相關研究[2][8][13]大多為提供健檢資料查詢、統計分析，並且由臨床方式來做偵測與診斷，但這些相關研究較缺乏以健檢資料為基礎的健康趨勢變化型樣分析、或是從單一與多個健檢指標探勘與疾病之相關性的預測。因此，我們的系統針對這些不足部份提出一個較完整的健檢資料探勘框架。其程可分為兩個部份，第一部份先將受測者的各個健康檢查指標資料進行前置處理，亦即將健檢資料轉換為可處理型態後，再以資料探勘技術產生規則庫；第二部份再以規則庫分析受測者單一或多個健檢指標的健康風險樣式與疾病之間的相關性。故本論文之貢獻如下兩點：

1. 本研究提出一個以健檢資料為基礎的健康資料分析系統之架構，可分析出各個健檢指標的健康風險樣式。

2. 本研究也進一步探勘健康風險樣式與慢性疾病之間的相關性，並建置慢性疾病的預測模型。

2、相關文獻

在本章節將針對問題的定義、健康檢查之重要性加以討論，以及關聯規則、時序樣式與分類器等常被運用的資料探勘技術的內容加以說明。

2.1 問題定義

健康檢查的資料是由多種檢測項目所集合而成的報告。受測者透過健康檢查可取得目前身體狀況的各種檢測項目的測量值，促使許多民眾以定期做健康檢查來確認自己的身體狀況，Table 1 為單一受測者多次健康檢查資料表。

Table 1 單一受測者多次健康檢查資料表

	檢測項目A1	檢測項目A2	…	檢測項目(N-1)	檢測項目N
第一次健檢	3.2	27	…	97	1.3
第二次健檢	3.8	30	…	110	0.95
第三次健檢	5.1	28	…	130	1
第四次健檢	4.8	30	…	133	0.95
第五次健檢	5	35	…	157	1.3

而多次的健康檢查除了能提供每一個時間點受測者的健康狀況外，也可將每一個檢測項目視為一個隨著時間變動的時間序列值。因此對於連續受測的眾多受測者資料，其各個屬性我們可以將其變化視為一時間序列，如 Figure 2 所示。

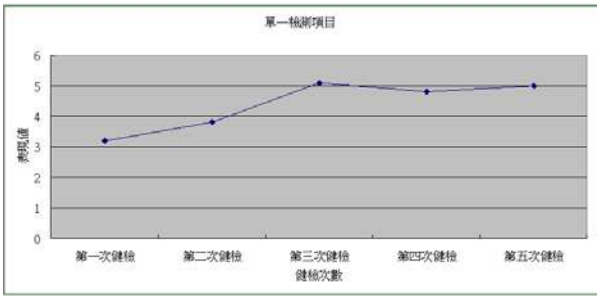

Figure 2 單一受測者單一屬性多次健康檢查趨勢圖

本研究主要針對健檢資料的趨勢分析與疾病的預防，因此，本研究目標可分為以下兩個研究主題，茲探討如下。

(1)如何從單個屬性的連續健檢測資料中運用資料探勘技術找尋出可能危及健康的風險樣式。在多次的健檢資料裡，每個檢測項目的值可在多次測量後得到健康風險樣式。

(2)如何從健檢資料中預測受測者罹患疾病的可能性。若能藉由健檢資料來針對慢性疾病的分析及預測，將可有效提高國人健康安全的品質。因此，本研究將利用受測者的單次健檢資料與健康風險樣式的資料，建構慢性疾病的預測模組。

2.2 健康檢查之重要性

過去最早提倡定期性及綜合性健檢觀念的是英國的 Dr. Horace Dobell(1861)，亦即是把多項疾病篩選檢查組合

在一起便成為「健康檢查」，並認為受測者定期的檢查可以預防罹患疾病及死亡，之後的醫師也皆認同定期健檢是預防醫學的基礎[9][11]。藉由定期的健康檢查，民眾才能了解自己目前的健康狀況，且當某健檢指標為異常時，受測者可在最短時間內就醫，並進行合適的病症處置，才不致錯過最佳治療時機。因此，我們可清楚瞭解健康檢查的重要性，亦即可預先獲得個人健康狀況，進而提供慢性疾病的早期預防機會，如高血壓、糖尿病、腎臟病、心臟病等慢性疾病。

但是受測者往往只得知當次的健檢報告結果，因此，若能以受測者的”歷史健檢記錄之健康變化趨勢”取代每次只有”當次”的健檢報告，也許可讓醫療人員更瞭解受測者的健康狀況，進而給予更詳細的診斷建議，以減少病症所造成的健康損害。例如，雖然受測者的當次健檢結果並未出現健康警訊，但若其健檢指標的歷年趨勢若與 Figure 2 的健康趨勢相似時，則醫師可及早給予受測者相關健康預警與診斷建議。因此，發展可實際應用於健檢資料的分析技術，就有其需要與效益，以達到「及早發現儘速治療」的實質功效。

2.3 關聯規則

關聯規則的主要目的是欲從龐大的資料中發現資料項目之間的相關性，最早是由 Agrawal [6]等學者所提出。假設在資料庫中存在著許多筆交易記錄，且每筆記錄皆記錄消費者所購買的商品項目。而關聯法則就是用來表示項目之間的關係，其形式為”{X}→{Y}”，其中 X 和 Y 代表項目的集合，例如，最為人知的真實關聯規則為{尿布}→{啤酒}，此規則表示出大部份消費者在購買尿布的時候，通常也會購買啤酒。而本研究亦將利用關聯規則的技術來發現在不同檢測項目內屬性值之間的相關性，作為預測模型的輔助資訊。

2.4 時序樣式

時序樣式探勘[7]是由 Agrawal 等學者在 1995 年所發表的論文 Mining Sequential Patterns 所提出的方法。在此論文裡，其主要目的是針對交易資料庫作分析並且探勘其交易的行為模式。交易資料庫中包含顧客身份辨識碼(customer_id)、交易發生時間(transaction time)、以及該筆交易內顧客所購買的項目(items bought in the transaction)。交易資料庫好比一般超級市場中，每一次的消費記錄都存在顧客編號、交易發生時間、以及顧客所購買的商品組合。為了深入瞭解顧客的消費行為模式，期待找出有順序性的資料，利用時序樣式探勘方法來找出其中隱含且有用的資訊，例如顧客在買了電腦後，在下次消費時，會購買印表機，最後，也會在購買墨水匣等循序購買行為，如 Figure 3 所示。然而，這並非是要一定連續發生的事件，而是循序發生的事件，亦即是顧客可能在這當中，會購買其它產品的消費行為，僅是依照時間發展的順序去表現此行為模式，這就稱為時序樣式探勘。

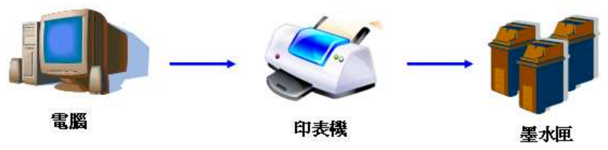

Figure 3 時序樣式示意圖

在本研究將循序樣式概念應用於各個檢測項目的多次測量值資料裡，是否存在某些循序樣式而與慢性疾病是具有風險上的關係，亦即是尋找在健檢資料的各個檢測項目資料裡的循序樣式，藉以建構更完善的慢性疾病預測機制。

2.5 分類器

分類器主要的目的在建立一個分類模組(Classification model)，目前有許多文獻在研究建立分類模組[12][14]，雖然在分類器建置方法上不盡相同，但主要分類概念如 Figure 4 所示。

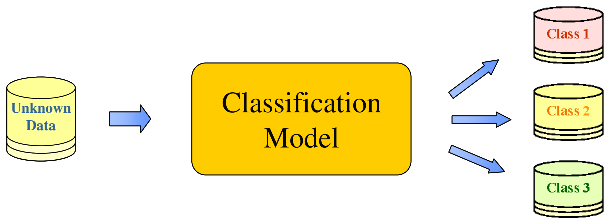

Figure 4 分類器分類概念圖

Figure 4 中未知的資料可以透過事先訓練好的分類模組進行分類動作。舉例而言，當一位新病患到診所求診時，有經驗的醫生透過自己的專業知識告知病患罹患的是哪一種疾病；分類器所建立的分類模組就好比有經驗的醫師一樣，可以根據訓練模組進行判斷的工作。

2.5.1 關聯規則分類探勘

關聯規則分類方法[10]是 Liu 等學者於 1998 年所提出的方法，其整個方法執行過程分為二個階段，第一個

階段是關聯規則的產生，第二個階段是關聯規則分類器的建立。在第一階段，藉由關聯規則探勘方法可取得屬性與屬性之間的關聯性，進而建立可讀性高的分類規則，關聯規則的表示方式為 X→Y，其中，X 為探勘所獲得的頻繁項目集(Frequent Itemset)，而 Y 為一分類類別(Class)。

在研究[14]中，其採用 CBA 方法分析關於氣喘的實際資料，而得到關於氣喘發作高風險判斷之關聯規則，如{溫差大，懸浮微粒增加，發燒}→{潛在會發生氣喘}，該規則的意思是指當過去幾天發生過溫差大，空氣中的懸浮微粒增加且病患本身出現發燒的症狀時，則病患有可能在幾天裡會出現氣喘發作的狀況。

由於完成關聯規則探勘後，會取得大量的關聯規則，因此，第二階段主要目的是為了建立有效的關聯分類器，所以，必須針對所獲得的關聯規則進行排序處理，主要是依據信賴度由高至低排序，若有相同值者，則以在資料庫中，出現次數的比例高者為優先，接著，再以規則長度短者為優先來進行分類規則的排序程序。此分類方法之優勢為在進行分類時，不需將整個資料庫皆讀入記憶體中，僅需比對經過修剪與排序後的規則作為判定的標準即可，此方式不僅能提昇記憶體的使用率，亦可增加執行處理的效率。在本研究裡，關聯規則的概念可應用於尋找各個檢測項目之間的相關性，以找出對於慢性疾病的影響風險因子的資訊，藉以建置更完善的疾病預測機制。

3、研究方法

在本章節，本研究所提的系統架構將說明如下。

3.1 系統建置程序

本章節將介紹本研究所提出的個人健康分析系統的方法設計。系統的方法設計主要分為系統建置和系統運作兩種程序。在系統建置程序，此程序的目的是在於系統的建立。因此，方法設計以健康檢查為輸入的資料，並分為兩階段對資料做不同的處理，分別輸出本計劃提出的健康風險樣式和疾病預測模組。系統運作程序則是系統的應用方式，將本計劃的系統應用在健檢的機制上用以加強對受測者健康的分析和慢性疾病的預測。因此，在方法的設計上則是以如何針對新的健康檢查資料給予未來健康趨勢分析以及可能之慢性疾病的準確預測為主。

本程序的方法主要由兩個階段所構成。此兩階段分別為健康風險樣式探勘階段和疾病預測模組建立階段。第一階段的健康風險樣式探勘，找出健檢歷史資料中各種項目的健康風險樣式。而健康風險樣式則是各項檢測值中隱藏的導致受測者有健康風險的常見樣式。第二階段的疾病預測模組建立，則是利用健康風險樣式和靜態的健檢資料針對慢性疾病的發病建立分類模組。Figure 5 顯示了本程序的方法流程圖。

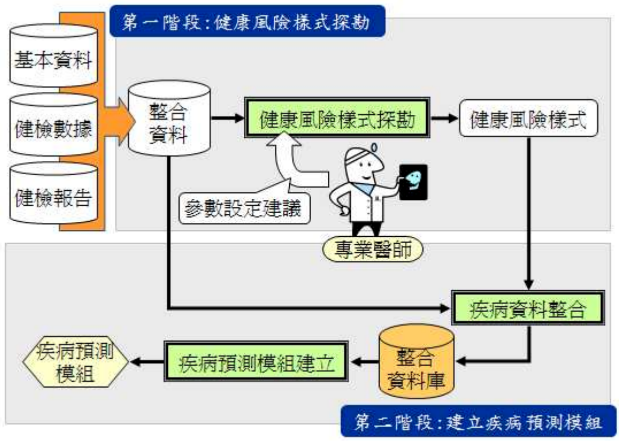

Figure 5 系統建置程序圖

3.1.1 第一階段：健康風險樣式探勘

本程序的第一階段主要為健康風險樣式的探勘。利用健康檢查的歷史數據和報告探勘出各種檢測值常見的健康風險樣式。為了能探勘出各項檢查值的隨時間變動的健康趨勢，我們匯整所有健康檢查的歷史記錄。將每位受測者各項的健檢項目依照檢驗的次序做排列。因此，每一位受測者的各項健檢記錄都能整理成一條時序資料。根據醫師的專業建議將每一個受測項目的健康區間記錄下來，藉此我們能了解每一個檢測值其健康的程度。

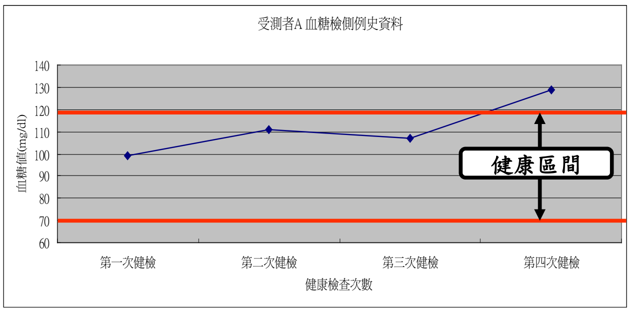

Figure 6 單一血糖檢測歷史資料及血糖值健康區間圖

Figure 6 顯示了一個受測者的血糖資料，受測者擁有多次的檢測記錄，因此以時序資料的型態顯示多次檢測值的變化。而 Figure 6 中的紅色線所夾的區間則是由

醫學資料或是醫師經驗所定義的健康區間。

在健康風險樣式探勘方法的設計上，並無預設哪檢測項目與目標疾病相關，因此樣式的探勘著重於找出單一檢測項目的特徵。我們將健康風險樣式定義為每一個檢測中，能代表著走向高風險趨勢的高頻序列樣式和代表著走向無風險趨勢的高頻序列樣式。

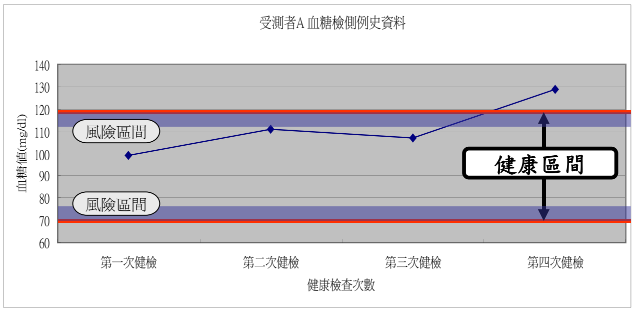

Figure 7 高風險區間

藉由健康區間診斷健檢資料是以往傳統健檢報告的作法。然而，本研究的目的是要提供受測者更有預防作用的健康風險樣式分析，因此定義了比健康區間更嚴格的高風險區間，並設定一個參數 R，其數值代表健康檢查的檢測值有多接近健康區間的臨界點，用來界定該項目檢測值是否為高風險。R 即為決定高風險區間大小的參數，其代表的大小為健康區間的百分比。Figure 7 中，R 的值被設定為 0.3(30%)，代表健康區間的 30%將被規劃為高風險區間，其餘的 70%則為無風險區間。若有一受測者的血糖檢測項目的時序曲線處於高風險區間或是已經處於健康區間之外，則代表該受測者的血糖時序曲線屬於高風險資料。反之，處於健康區間內且不在高風險區間中，則該血糖的時序曲線屬於無風險資料。

將這些時序資料依照被標示的風險程度(高風險/無風險)區分為兩群資料。利用時序樣式探勘(sequential pattern mining)的技術針對這兩群資料裡每一個受測項目所構成的時序資料集做探勘。找出所有與風險相關的高頻率序列樣式(sequential pattern)，即可得到高風險和無風險狀態下各項檢測項目的健康風險樣式，但有些健康風險樣式會同時存在高風險及無風險狀態下，因此，將這些樣式選擇移除，讓留下來的健康風險樣式更有參考價值。這些健康風險樣式與健康的程度有高度的相關，可以提供醫療人員做診斷的參考特徵。在系統建置程序的第二階段中將充份的利用這些健康風險樣式建立疾病預測模組。

3.1.2 第二階段：疾病預測模組建立階段

在第二階段中，我們將利用第一階段所找到的健康風險樣式建立慢性疾病的預測模組。必須要指定要預測的目標疾病，並且針對現有資料中所有受測者做疾病的判斷。

糖尿病	飯前血糖	血壓(收縮)	…	飯前血糖 P1	飯前血糖 P2	…	有風險?
受測者P1	99.2	130	…	T	F	…	是
受測者P2	84.8	140	…	T	F	…	是
受測者P3	101.1	125	…	F	F	…	否
受測者P4	89.8	128	…	F	T	…	否
受測者P5	95.0	150	…	T	F	…	是

Figure 8 整合靜態資料與健康風險樣式的資料表

健康風險樣式可算是身體狀況變化的趨勢，因此無論是高風險或無風險的樣式皆可視為是身體某種狀態的變化。為了建立準確的慢性疾病預測模組，考慮健康風險樣式是必需的。健康檢查多數受測者是屬於自發性的檢查，因此資料的記錄時間與記錄次數也不一致，更多受測者是只有檢查過一次的資料。這些受測者本身的健檢記錄無法構成連續的時序資料，因此在健康風險樣式上無法為這些資料做分析。為了能建構更全面性的慢性疾病預測模組，所有健檢資料的最後一次健檢結果也會被加入考慮。整合靜態資料後的資料表將會如 Figure 8 所示，每一筆包含一位受測者最後一次接受健檢的數據和報告，若該資料所屬的受測者有多次的歷史健檢記錄，則該筆資料也會包含健康趨勢的特徵屬性。Figure 8 的最後一欄則是針對系統所要建立的目標疾病做標示。由健檢資料或是門診資料的記錄看出受測者是否有目標疾病。在健檢資料上，根據健檢的健康參考值來判斷受測者是否有目標疾病，亦或是由受測者的門診記錄中的數值或是診斷報告中判斷是否有目標疾病。

最後，將以關聯規則分類探勘(Classification based on Association Rules)的方式建構慢性疾病分類預測模組。

3.2 系統運作程序

在系統建置程序完成後，系統可提供健檢受測者更多關於健康的建議和慢性疾病相關的預測。

當一般受測者接受了健檢後，若該受測者之前有過健檢的記錄，那麼本研究的系統將會整合各檢測項目的歷史資料。將這些資料輸入本研究的系統中，受測者

除了一般健檢的檢驗報告外，也可以根據過往的檢查記錄獲知各項健檢項目的健康風險樣式，以了解是否有哪些健康風險樣式是有風險且需注意的。此外，慢性疾病的預測功能也在同一時間給與關於慢性疾病未來發病機率的預測，醫師或健檢人員可以根據這些資料告知受測者更多的資訊或建議，若有發現慢性疾病潛伏也能盡早給予追蹤治療。

4、討論與結論

本研究針對健檢資料提出一個個人健康資料分析系統架構，此系統運用資料探勘來發展相關分析技術，藉以擷取各個健檢項目的健康趨勢型樣，同時，本研究分析健康趨勢型樣與疾病之間的相關性，並進一步建構預測模型。依據系統分析後的資訊可作為輔助資訊，使醫護人員可對受測者提出健康警訊或作出適當的醫療決策，進而達到及早預防及早治療的成效。

未來我們將依據此描述架構實作此系統，並透過南部某醫學中心之健檢資料進行各項實驗評估與分析，藉以驗證所提系統之實用性。

誌謝

本研究依經濟部補助財團法人資訊工業策進會「 98 年度資訊應用與整合技術開發第二期計畫(1/4) 」辦理。

參考文獻

[1] 中華民國衛生署，「2008 年國人十大死因」http://www.doh.gov.tw/CHT2006/DM/DM2_2.aspx?now_fod_list_no=10642&class_no=440&level_no=3.

[2] 李博智、邱昭彰、邱文科、劉祖華、莊逸洲、黃崇哲、許光宏。三維人體測值及資料探勘技術在高血脂症預測模式之應用，台灣醫療管理科學學會研討會，2002。

[3] 林惠美等，美兆觀點-e 世紀健康檢查，台北市：美兆文化，1999。

[4] 黃惠鈴，當年輕人得癌症，健康雜誌，第十期，1999。

[5] 詹長權，全球「社區整合式疾病篩檢模式」的先驅者-陳秀熙教授，臺大校友雙月刊，第四十期，2005。

[6] R. Agrawal, T. Imielinksi and A. Swami, “Mining association rules between sets of items in large database,” *The ACM SIGMOD Conference*, pp. 207-216, 1993.

[7] R Agrawal. and R. Srikant, “Mining Sequential Patterns,” *The 11th International Conference on Data Engineering*, Taiwan, 1995.

[8] I. Kononenko, “Machine Learning for Medical Diagnosis: History, State of the Art and Perspective,” Artificial Intelligence in Medicine, Vol. 23, Issue:1, pp. 89-109, 2001.

[9] R. Lawrence and A. Mickalide. “Preventive service in clinical practice: Designing the periodic health examination.” JAMA. 1987;257:2205-7.

[10] B. Liu, W. Hsu and Y. Ma, “Integrating classification and association rule mining,” *The Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 80-86, 1998.

[11] Medical Practice Committee, American College of Physician. Periodic Health Examination: A guide for designing individualized preventive health care in the asymmetric patient. Annals of Internal Medicine. 1981;95:729-32.

[12] J.R. Quinlan, “C4.5: Programs for machine learning,” Morgan Kaufman, 1993.

[13] K. Steven, W. Dennis and K. Matthew, ”Artificial Neural Networks for Early Detection and Diagnosis of Cancer,” Cancer Letters, Vol. 77, pp. 79-83, 1994.

[14] J.R. Quinlan, “Discovering rules from large collections of examples: a case study,” In Michie, D., editor, Expert Systems in the Microelectronic Age. Edinburgh University Press, Edinburgh Scotland, 1979.

[15] Vincent S. Tseng, C.H. Lee and Jessie C.Y. Chen, “An integrated data mining system for patient monitoring with applications on asthma care,” *The 21th IEEE International Symposium on Computer-Based Medical Systems*, 2008.

[16] M. Zamora, "The Study of Micro-Arousals Using Neural," IEE Conference Publication, pp. 625-630, 1999.