比較三種資料探勘演算法預測子宮頸癌五年存活的外部通用性效能

張語恬[1]　朱基銘[1]　簡戊鑑[1]　周雨青[1]　楊　燦[5]
盧瑜芬[1]　白健佑[2]　白　璐[1]　Thomas Wetter[4]
孫建安[1]　羅慶徽[3]

本研究比較類神經網路、邏輯斯迴歸及決策樹三種資料探勘演算法，使用不同診斷年份的樣本作模型訓練，對預測子宮頸癌五年存活情形的效能，並進行外部通用性（External Generalization）驗證。

本研究採用美國國家癌症研究所（NCI: National Cancer Institute）所提供的流行病學調查（SEER: the Surveillance, Epidemiology, and End Results）數據中的癌症登記資料庫（CIPUD，Cancer Incidence Public-Use Database），從西元1973年至西元2000年間選取156,502筆資料記錄及72個變項，經過資料清理後，留下與預測子宮頸癌五年存活較相關的18個變項，與子宮頸癌診斷年份爲1988-1996年的資料共2,022筆，依診斷年份將樣本，分成8組不同的模型訓練樣本與測試樣本，帶入類神經網路（artificial neural network）、決策樹（decision tree）以及邏輯斯迴歸（logistic regression）三種演算法造出模型，以AUC（area under the ROC curve）、準確率（accuracy），作爲演算法預測能力評估，並找出可以得到良好預測結果的模型設計。

結果顯示：內部驗證的模型預測力最好的爲類神經網路的模型1，其AUC與準確率值分別爲0.9392、0.9474。外部驗證的AUC結果，以類神經網路的模式7表現最好，其值分別爲0.6455。在內部驗證（internal validation）的AUC與準確率結果表現，類神經網路與決策樹都較邏輯斯迴歸佳。在外部驗證（external validation）的AUC結果表現，類神經網路與邏輯斯迴歸都較決策樹好。

類神經網路與邏輯斯迴歸建造的模型，有較好的外部通用性，而類神經網路與決策樹建造的模型，有較好的模型準確率。若想要得到較好的外部驗證結果，訓練樣本可以取過去的2-3年以上的資料。

（台灣家醫誌 *2007; 17: 222-38*）

關鍵詞：cervical cancer survivability, logistic regression, artificial neural network, decision tree, AUC (Area Under the ROC Curve)

[1]國防醫學院公共衛生學系暨研究所、三軍總醫院病理部[2]、家庭暨社區醫學部[3]、[4]Department of Medical Informatics, University of Heidelberg, Heidelberg, Germany、[5]美和技術學院 健康事業管理系

受理日期：95年8月7日　　同意刊登：96年10月28日

通訊作者：朱基銘　　通訊地址：台北市114內湖區民權東路六段161號　公共衛生學系

前　言

子宮頸癌為女性主要癌症之一，世界衛生組織資料顯示，子宮頸癌為全球女性常見癌症第三位，為我國女性癌症死亡的第五位，在2004年有926名女性因子宮頸癌死亡（每十萬女性人口死亡率8.33）。全球每年約有50萬新發生病例。2002年台灣子宮頸癌共有5,984例發生病例，包括原位癌3,618例及侵襲癌2,107例個案被診斷出來，粗發生率為每十萬人口51.88人，子宮頸癌之發生率高居女性癌症首位。

邏輯斯迴歸模型是傳統最常用於建立預測二元類別依變項的統計方法之一，過去疾病的預測，統計分析上大都以邏輯斯迴歸（logistic regression）作疾病預後存活情形預測，可以同時處理類別與連續變項，並可算出存活機率；但邏輯斯迴歸分析會受有效樣本數目、統計分析假設限制和交互作用項設定的影響，增加進行建置存活預測模型的困難。目前新興的人工智慧領域中的資料探勘技術，如類神經網路（artificial neural networks）與決策樹（decision tree）等，也如同邏輯斯迴歸模型可以建立預測二元類別依變項的資訊學技術，而且他們藉由對遺漏值較高的容忍度，有對少量有效樣本數的苛刻情況下，仍可以進行模型建立及分析的特性，對於在醫學上獲取完整自變項數據的有效樣本取得不易的情形，有較邏輯斯迴歸進行建置存活預測模型上有其潛在的能力與優點，可彌補邏輯斯迴歸演算法不足之處。

在臨床醫學上，如果腫瘤經治療後五年仍無復發，就被認為已經治癒，因此本研究用「五年存活情形」作為子宮頸癌預測的依變項。本篇研究目的在於探究以三種資料探勘技術在子宮頸癌預測模式的運用，分別應用類神經網路、邏輯斯迴歸以及決策樹三種演算法，與不同訓練樣本作預測模型訓練，並進行內外部驗證的比較；來探討已建置預測模型在預測未來樣本是的可用性，藉以推估不同演算法的外部通用性能力。

最後模型效能評估利用ROC（receiver operation characteristic）的AUC（area under the curve）與準確率（accuracy），作為評估演算法預測能力的指標，希望從中找出最佳預測模型，作為臨床上輔助鑑別診斷以提高診斷正確率，甚至更進一步改善治療提高存活率。

材料與方法

（一）資料來源

本研究使用美國SEER（The Surveillance, Epidemiology, and End Results）資料庫（CIPUD，Cancer Incidence Public-Use Database），1973年至2000年筆子宮頸癌資料記錄、72個變項進行資料清理，將包含遺漏值（missing value）、極端值、登錄錯誤、不合常理的個案數去除；此SEER子宮頸癌資料庫的追蹤終止日為2001/12/31日，為了去除無法判斷五年後是否為存活或死亡的個案，會刪去診斷日期為1996/12/31之後的個案，最後參與分析的資料個案數為2,022筆，留下17個較相關的自變項與1個依變項（表1），其中變項“罹病程度”（extent of disease）包含腫瘤大小（size of tumor）、腫瘤臨床分級（clinical extension of tumor）、淋巴轉移（lymph node involvement）、陽性淋巴節數目（＃of positive nodes

表1 變項內容

變項名稱（原英文變項名稱）	分類（類）	變項尺度
存活狀態(SURVIVAL)	2	類別
種族(RACE_ETH)	28	類別
婚姻狀態(MARITAL)	6	類別
腫瘤型態(PRIMARY_SITE)	4	類別
組織學型態(HISTOLOGY)	51	類別
腫瘤病變程度(BEHAVIOR)	2	類別
病理分級(GRADE)	5	類別
腫瘤進行外科手術類型(SITE_SPE)	18	類別
放射治療類型(RADIATION)	9	類別
臨床分期(SEER_MOD)	25	類別
首次惡性腫瘤(FIRST_MA)	2	類別
腫瘤擴散程度(EE)	19	類別
淋巴結侵犯狀態(L)	6	類別
淋巴結個數(EX)	—	集合
陽性淋巴結個數(PN)	—	集合
腫瘤個數(PRIMARY#)	—	集合
診斷年齡(AGE@DIAGNOSIS)	—	連續
腫瘤大小(SSS)	—	連續

註：詳細譯碼請見SEER網站

examined）、檢查的淋巴節數目（number nodes examined）、1995病理分級（pathological extension for 1995+ prostate cases only），本研究依該變項內容的特性轉換為上述六個變項，研究分析對象篩選流程如圖1。

將分析資料進行除錯，並將有包含遺漏值的個案數去除，最後參與分析的資料個案數為2,022筆，17個自變項與1個依變項，依變項為二分類：存活五年以上（“0”），存活不到五年（“1”）。訓練樣本、及測試樣本的建構情形如表2，依子宮頸癌診斷年份（1988-1996年）分成9份資料，並利用這9份資料的組合，建出8個不同的模型，模型1由1988年資料組成；模型2由1988年，加上1989年資料組成；模型3由1988年，加上1989年，加上1990年資料組成，以此類推，再利用每個模型資料裡，最高診斷年數的後一年資料做測試組，測試模型的外部通用性。

根據回顧文獻，小樣本的表現類神經網路優於邏輯斯迴歸，因此本研究設計建置上述模型1-8的理由是訓練樣本數目逐年增加；如表2顯示模型1-8的訓練樣本數目分別逐年遞增為190、414、618、824、1046、1270、1500和1779，並依照訓練資料集的下一年度則作為外

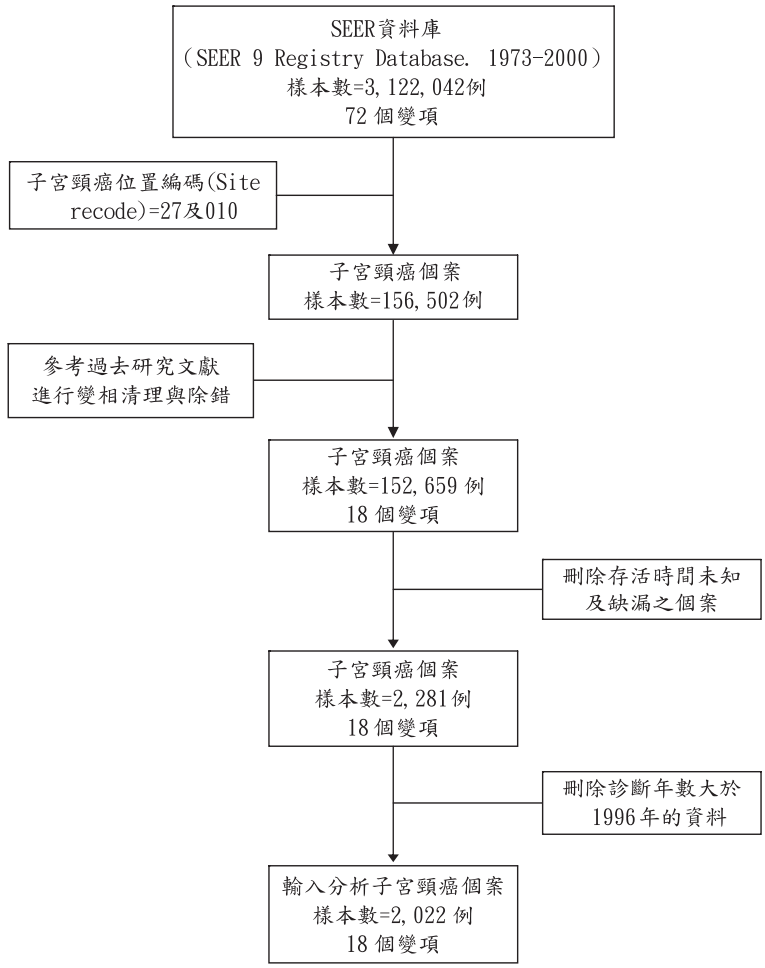


圖1　研究對象篩選流程

部驗證用的測試樣本；數目相近，分別為204-279不等。這樣的用意乃為了排除publication bias，驗證訓練樣本數目的多少和累積的年代長短，對模型預測能力的影響究竟為何？

此外，由於不同年代組合的訓練樣本，邏輯斯迴歸以向前逐步法選取變項，所形成的預測模型自變項可能自然（nature course）會有差異，所以邏輯斯迴歸模型自變項的選擇乃依照預測準確率高者為條件；而不以相同預測自變項組合為條件，但仍根據表1列出的自變項為模型初始納入的自變項，因此模型1-8納入的自變項組合變異情形對於準確率是為保守估計。

（二）內、外部驗證

三種預測模式內部設定的決定，是以得到較好的外部通用性（external generalization）為依據。

外部驗證

所謂外部驗證即外部通用性，意指

表2 模型建構及其訓練資料、及測試資料（外部驗證）設計

模型	訓練資料	個案數	測試資料（外部驗證）	個案數
模型1	1988	190	1989	224
模型2	1988+1989	414	1990	204
模型3	1988+1989+1990	618	1991	206
模型4	1988+1989+1990+1991	824	1992	222
模型5	1988+1989+1990+1991+1992	1,046	1993	224
模型6	1988+1989+1990+1991+1992+1993	1,270	1994	230
模型7	1988+1989+1990+1991+1992+1993+1994	1,500	1995	279
模型8	1988+1989+1990+1991+1992+1993+1994+1995	1,779	1996	243

將訓練好的預測模型，帶入不屬於模型訓練個案集合的測試樣本的資料，所得到的預測結果，看其預測結果（AUC、準確率…等），此驗證的過程稱為外部驗證（external validation）。若外部通用性好的話，表示此模型可以被廣泛的使用於其他非訓練樣本的資料。

內部驗證

意指將訓練好的模型，帶入仍屬於模型建造時的訓練個案集合的樣本，看其預測結果（AUC、準確率…等），此過程為內部驗證，目的是為了檢視模型本身訓練結果的好壞，若內部驗證AUC與準確率值高的話，表示模型能準確預測，所放入訓練樣本的情形。

但是有好的內部驗證結果，不一定外部驗證結果也好，所以一個模型的建造，除了考慮內部驗證的結果，同時也要考慮其外部驗證的結果。本研究三種預測模式的內部設定，是以得到較好的外部驗證結果為依據，所以本研究中內部驗證的結果並非三種預測模式的最佳情形。

（三）預測模式

本研究的資料處理是用SPSS 12.0，模型的建造與內、外部驗證則是用Clementine 7.2。以下為三種演算法的介紹：

類神經網路

類神經網路是仿造人類大腦組織及運作方式所發展而成，必須透過訓練（training）的方式，讓類神經網路反覆學習，直到每個輸入都能正確對應到所需要的輸出，讓輸出值越接近目標值。類神經網路訓練過程-首先網路會隨機分派初始加權值給神經元，網路會記住每個學習循環（epoch）的加權值，學習循環完畢後，驗證組（validation set）會被帶入不同加權值的網路模型，選出MSE（mean squared error）最小的模型，作為最終確定模型。類神經網路在學習前，必須建立出一個訓練樣本（training set）使類神經網路在學習的過程中有一個參考。類神經網路可以建構複雜且非線性的模型，亦可以接受不同種類的變數作為輸入。因為類神經網路模式所用的變

項可為時間相依性（time dependent），或是存在交互作用變項，且適用於複雜非線性的分析。類神經網路預測模式已經成功地被運用在臨床上的存活分析，為目前醫學界，用於二分類變項預測上最常用的方法。

本研究採用多層神經元（MLP: multi-layer perceptron）的倒傳遞法類神經網路，有81個輸入神經元（Inputs，輸入變項含虛擬後的變項）、1層隱藏層（hidden layer，40個隱藏層神經元）、1個輸出層（output，預測變項為存活或死亡）、及70%訓練樣本（dample to prevent overtraining，預防過度訓練的取樣百分比），隨機種子為“1”（tandom seed，取樣用的亂數基點），持續力200（persistence，停止訓練條件為平均錯誤率為改變的學系循環數），訓練方法（method for topology）為“快速”（根據使用者或預先設定隱藏層數及網路拓墣），訓練停止於“預設”（學習停止條件，達到準確率90%或學習循環數250或超過5分鐘），學習率（learning rates）為“預設”（alpha=0.9, Eta=0.3），資料檢查處理為強制（coerce，強制只納入變項數值在本研究所設定檢查範圍的資料）（表3）。

邏輯斯迴歸

邏輯斯迴歸，屬線性迴歸的一種，主要適用在預測二分類或是多項式的變項，普遍使用在依變項為類別變項的時候。本研究將依變項分成二類別的變項，分別是0和1（存活五年以上、未活過五年），並採用向前邏輯斯迴歸。邏輯斯迴歸的常見前提假設為，自變項要跟對數轉換後的依變項呈線性關係，通常這種模式的方法會使用逐步（向前、向後），來對變項進行篩修（pruning），但是在不同情況下，各種方法（輸入、向前、向後）都有其優點與缺點。邏輯斯迴歸當輸入的變項愈多的時候，會增加其內部驗證的準確率，但外部驗證的結果不一定會增加，有時甚至會下降許多。但在本實驗中用的是

表3 三種預測在Clementine 7.2分析條件

	類神經網路	邏輯斯迴歸	C5.0決策樹
模型內容設定	持續力：200	方法：向前	輸出類型：決策樹（使用增加功能：100）
	隱藏層神經元數：40	模型類型：主要效果	模式：簡單
	訓練樣本：70%	偏愛：專家	偏愛：準確度
	隨機種子：1	奇異容忍度1.0E-5	預期雜訊：5%
	方法：快速		
	停止於：預設		
	專家：		
	隱藏圖層1個		
	學習率：預設		
資料處理	檢查：強制	檢查：強制	檢查：強制

向前邏輯斯迴歸，納入的自變項最多只有9個，所以不會有此方面的問題。

本次研究邏輯斯迴歸的方法（method，自變項置入模型的規則）為“向前逐步迴歸”（forward，逐一置入具統計顯著性的自變項），模型類型為“主要效果”（main effect，自變項不含交互作用項），奇異容忍度為1.0E-5（singularity tolerance，共變矩陣的歧異度，若正交向量特徵值小於設定的容忍度即令其轉置矩陣為退化性矩陣），資料檢查處理為強制（同類神經網路參數說明，表3）。邏輯斯迴歸雖然方法在輸入（Enter，不論具統計顯著性與否置入所有自變項），即將所有變項都放入，會得到模型本身內部驗證時較好的準確率，但是外部通用性表現不佳，最後決定以“向前”為此次設定，因為此設定可以得到最好的外部通用性。模型類型，有兩種設定，一為“主要效果”，另一個為“飽和（full factorial）”，“主要效果”設定意為，變項的放入，不討論其交互作用的影響，但是飽和在變項放入時，會考慮變項間的交互作用，此方法可以用來處理較複雜的問題，但是因為此次研究放入的自變項共有17個，對飽和的模型類型來說，自變項太多，以致於無法執行，所以模型類型設為“主要效果”。

決策樹

決策樹是一建立分類模式（classification models）的方式之一，針對給定的資料利用歸納的方式產生樹狀結構的模式。為了要將輸入的資料分類，決策樹的每一個節點即為一個判斷式，判斷式針對一個變數去判斷輸入的資料大於或等於或小於某個數值，每一個節點因而可以將輸入的資料分成若干類。決策樹是通過遞迴分割（recursive partitioning）建立而成，遞迴分割是一種把資料分割成不同小的部分的疊代過程，如果有以下情況發生，決策樹將停止分割：該群資料的每一筆資料都已經歸類到同一類別；該群資料已經沒有辦法再找到新的屬性來進行節點分割；該群資料已經沒有任何尚未處理的資料 。基本的決策樹構造演算並沒有將雜訊考慮進去，此時模型很有可能會過度訓練（over-fitting或說overtraining），對訓練數據完全擬合，但不具很好的預測能力，樹枝的篩修可以解決過度訓練的問題，而刪修的方法又可分為，向前（forward pruning）與向後（backward pruning）刪修，一般情況下樹愈小，其預測能力愈強。此次研究決策樹的樹狀圖採用C5決策樹分析系統，輸出類型為“決策樹”，“使用增益功能”（use boosting，藉由重複選取及組合樣本訓練決策模型，用以增加其準確率的同時也預防過度訓練）為100組重組樣本，“模式”（model）為簡單，“偏愛”為準確度（favor accuracy，以準確度為建立決策樹的優先考量），“預期雜訊”為5%（expected noise，用以預防過度訓練），資料檢查處理為強制（同類神經網路說明，表3）。

（四）結果分析評估指標

AUC是由敏感度（x軸）與1-特異度（y軸）所畫出的ROC圖其曲線下的面積，是用來判定模型分類（discrimination）能力的好壞，AUC值愈大表示模型的分類能力愈好。圖2為模型1的AUC圖形。AUC值與模型分類能力好壞的一般大略可分為如下。

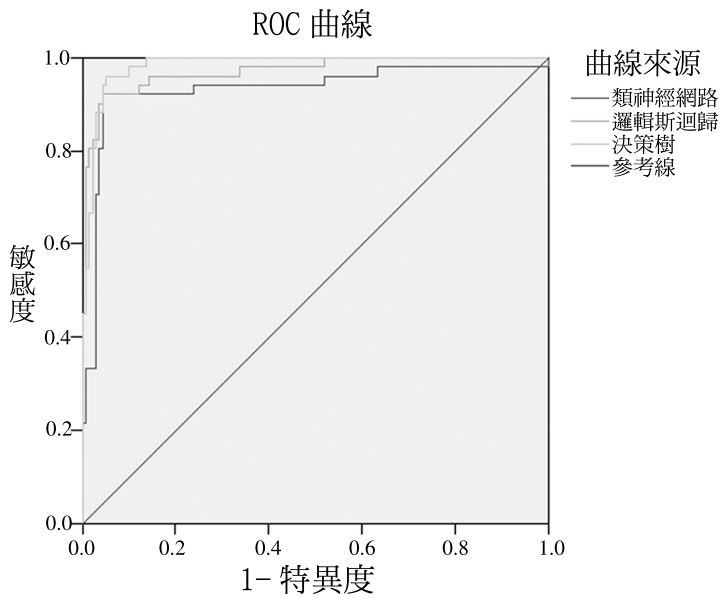
AUC值	模型分類能力
0.9~1.0	很好（excellent）
0.8~0.9	好（good）
0.7~0.8	普通（fair）
0.6~0.7	差（poor）
0.5~0.6	很差（fail）

準確率在流行病學上，常用來表示篩檢儀器篩檢能力好壞評估。在本研究中利用準確率（accuracy）來評估預測模型其預測能力的好壞，準確率愈高表示模型預測能力愈好，反之則反；計算方法如下：

準確率＝正確預測五年後存活與不存活的個案數／所有個案數

結　果

Delen 2005進行乳癌患者存活預測模型建立的研究中，將約40萬筆乳癌患者登記資料，在進行清理後剩約20萬筆資料進入分析模型。本研究與Delen 2005



預測模型AUC (Area Under the Curve)效能檢定

預測模型	AUC值	標準誤	*p*值[a]	95% 信賴區間	
				下限	上限
類神經網路	.933	.026	<0.001	.882	.984
邏輯氏回歸	.971	.013	<0.001	.945	.997
決策樹	.984	.007	<0.001	.971	.998

[a]: 比率檢定虛無假設AUC＝0.5

圖2　由SPSS 12.0繪出的模型的AUC圖形

不同的3點是；1.探討類似以女性為主的子宮頸癌，其患者的存活預測模型建立是否類似；本研究在SEER資料庫中，根據疾病分類碼共選取156,502筆子宮頸癌個案登記資料。2.在進行變項清理及除錯資料的步驟（詳見本文圖1），除了刪除資料中變項明顯的空值外，為了避免極端異常的個案影響分析，相較於Delen 2005刪減乳癌腫瘤大於200 mm的患者樣本數目差異；是由於本研究考慮子宮頸部位相對於腫瘤的大小及分布將子宮頸癌腫瘤大於50 mm （釐米）的個案剔除，剩餘152,659筆子宮頸癌個案登記資料。3.在刪除存活時間未知及闕漏值個案的步驟，本研究以5年含以上為乳癌所致存活或死亡判定標準，所以將觀察未滿5年的存活個案約（45,259筆）刪除為107,400筆；此外，也根據SEER譯碼本[16]將變項編譯為闕漏和未知及不明等數值的資料刪除（如腫瘤大小[17]的999, AJCC stage 3rd edition的98, 大部分變項的9或99），最後納入分析的資料由10萬筆左右減為2千筆；比較刪減前後的資料樣本年齡分布仍然相似[18]，可以推論刪減前後樣本仍同質和具代表性。這主要步驟相較於Delen 2005並未進行此一步驟，應該比較合理，但是使得資料筆數刪減量最大的主因。

模型的預測能力，可以用準確率與ROC曲線下面積（AUC）來做判斷，當準確率越大、AUC值愈大時，表示模型的預測效果愈好。如圖2所示，類神經網路、邏輯斯迴歸、及決策樹等分析所建置的存活預測模型整體的內部驗證AUC分別為0.933、0.971及0.984，根據其95%信賴區間（CI，Confidence Interval）這3種預測分析在訓練時的效能沒有統計顯著差異。

內外部驗證AUC結果如圖3所示，大體而言，在這八個模型中，類神經網路與決策樹模式的內部驗證表現結果比邏輯斯迴歸好，且三種預測模式其內部驗證的AUC值，大體都有由模型1往模型8下降的趨勢，此趨勢尤其以類神經網路與邏輯斯迴歸較明顯。而三種預測模式的外部驗證AUC結果，整體而言，由模型1往模型8攀升，但是幅度並不是很明顯。從模型4-8，類神經網路與邏輯斯迴歸的外部驗證AUC值較決策樹高。

內外部驗證準確率結果如圖5所示，在內部驗證準確率表現上，整體而言，決策樹會優於類神經網路與邏輯斯迴歸，且類神經網路與邏輯斯迴歸的曲線相似。三種預測模式的外部驗證準確率在八種模型間的表現並沒有明顯規則可循，但是可以看出在模型7的時候，三種預測模式都有較好的外部驗證值準確率。圖4、6為內外部驗證的AUC與準確率差值，當此差值小時，表示此模型的用於預測非訓練樣本的資料時穩定性高，即出來的預測結果與內部驗證結果相似。類神經網路與邏輯斯迴歸的AUC與準確率差值，都有由模型1往模型8下降的趨勢，且都在模型7達到最小差值，而決策樹的AUC與準確率差值在模型7達到最小，模型2為次小，整體上看來模型3~模型8決策樹的AUC與準確率差最大，邏輯斯迴歸的差最小。

不同的預測模式/模型，所納入的變項與變項相對重要性也不同，見表4。在類神經網路中，將8個＞病理分級＞組織學型態＞淋巴結侵犯狀態＞腫瘤進行外科手術類型＞腫瘤大小＞種族＞放射治療類型＞婚姻狀態＞腫瘤個數＞腫瘤型態惡性與否＞診斷年齡＞腫瘤病變程度＞首次惡性腫瘤＞淋巴結個數＞陽性淋

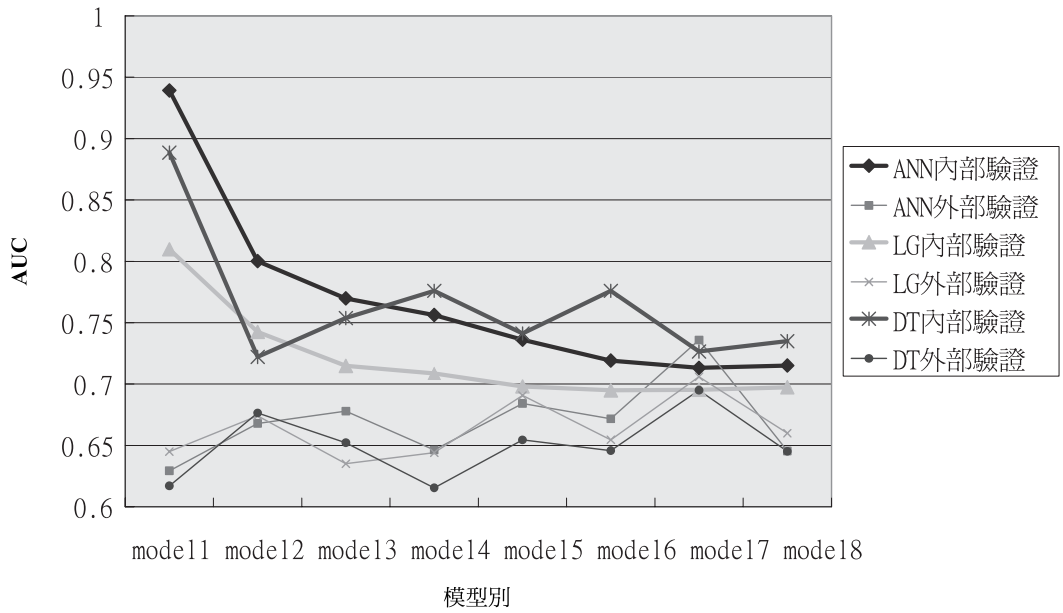


圖3　內、外部驗證AUC值折線圖比較。右邊圖例中的文字，ANN：類神經網路，DT：決策樹，LG：邏輯斯迴歸，model：模型。

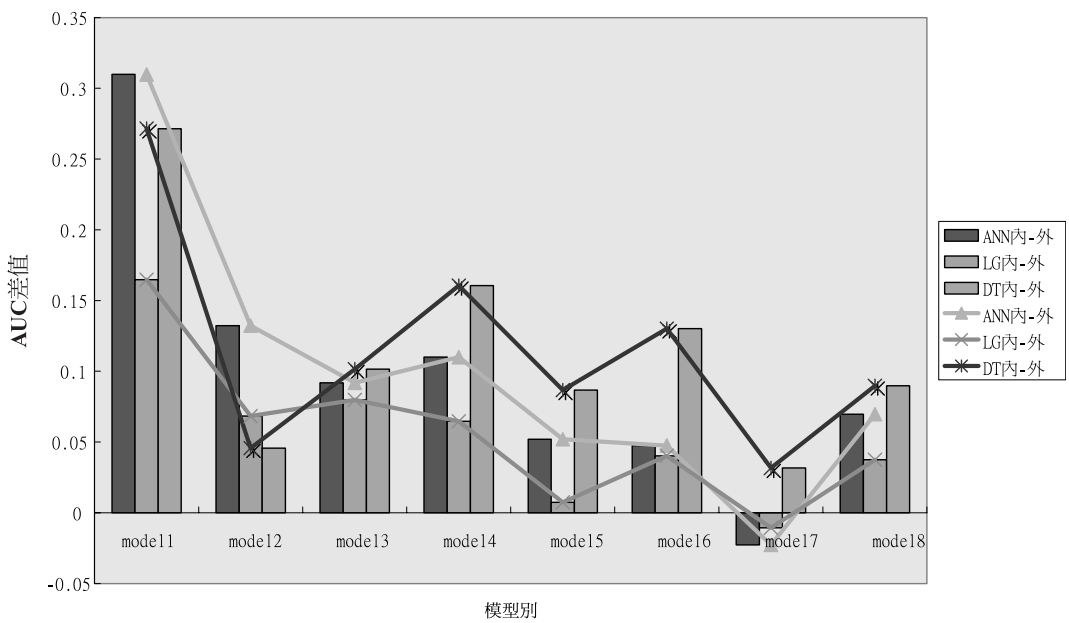


圖4　內、外部驗證AUC差值圖。右邊圖例中的文字，ANN：類神經網路，DT：決策樹，LG：邏輯斯迴歸，內－外：內部驗證AUC－外部驗證AUC。

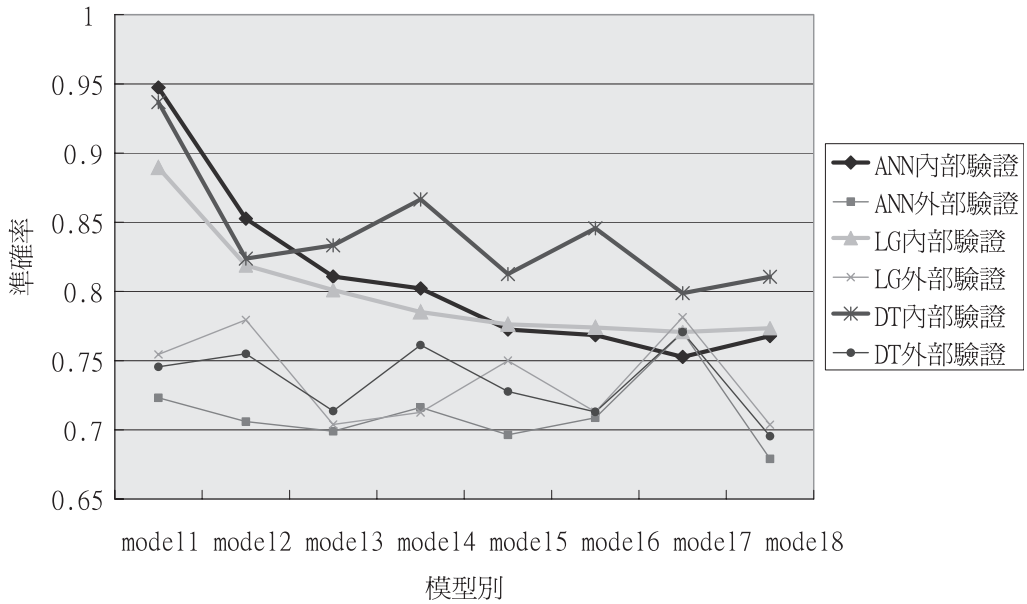


圖5　內、外部驗證準確率值折線圖比較。右邊圖例中的文字，ANN：類神經網路，DT：決策樹，LG：邏輯斯迴歸，測驗前：內部驗證，測驗後：外部驗證，model：模型。

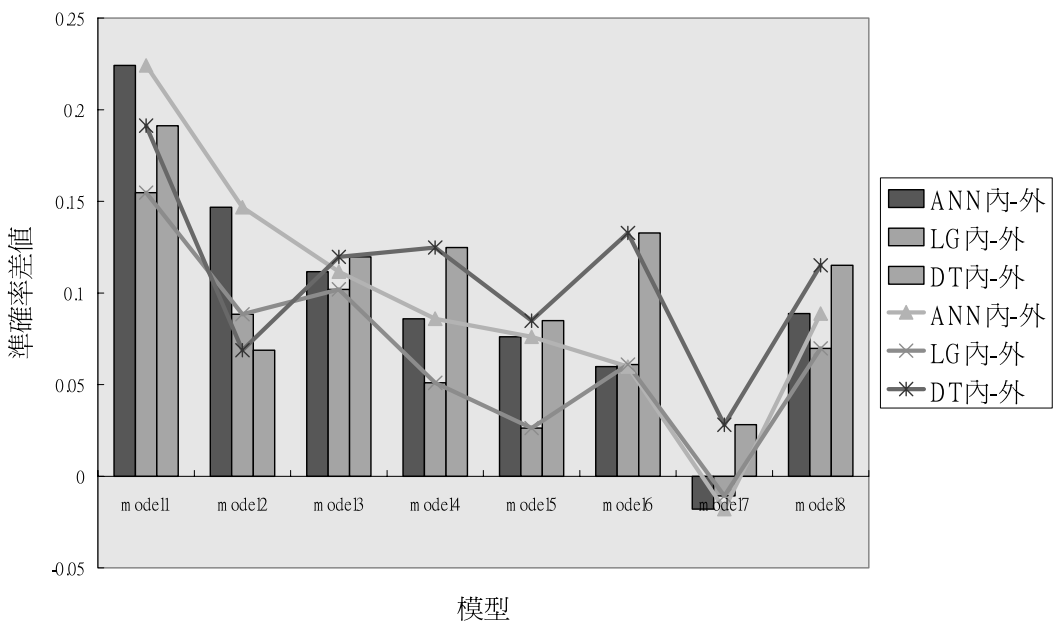


圖6　內、外部驗證準確率差值圖。右邊圖例中的文字，ANN：類神經網路，DT：決策樹，LG：邏輯斯迴歸，內－外：內部驗證準確率－外部驗證準確率。

表4　三種模式的預測

預測模式	變項重要性排名	較不重要變項
類神經網路	臨床分期>腫瘤擴散程度>病理分級>組織學型態>淋巴結侵犯狀態>腫瘤進行外科手術類型> 腫瘤大小	種族＞放射治療類型＞婚姻狀態＞腫瘤個數＞腫瘤型態惡性與否＞診斷年齡＞腫瘤病變程度＞首次惡性腫瘤＞淋巴結個數＞陽性淋巴結個數
向前法邏輯斯迴歸	腫瘤擴散程度＞淋巴結個數＝腫瘤大小＞病理分級＝淋巴結侵犯狀態	首次惡性腫瘤、婚姻狀態
決策樹	腫瘤大小＞腫瘤進行外科手術類型＞腫瘤個數＝陽性淋巴結個數＝婚姻狀態＝淋巴結侵犯狀態＝腫瘤擴散程度＝診斷年齡	腫瘤病變程度
綜合以上所論	腫瘤擴散程度、淋巴結侵犯狀態、腫瘤大小	首次惡性腫瘤、腫瘤個數

巴結個數等。

在邏輯斯迴歸預測模式中，因為本研究用的是向前邏輯斯迴歸，所以會選比較重要的變項納入，模型1～8所納入的的總變項數分別為，5、6、6、7、8、8、9、9，可以發現納入變項的數量從模型1往模型8增加；推測為訓練模型的樣本數增多時，向前法邏輯斯迴歸得增加納入的變項，以達良好的模型訓練效果。變項在八種模型中被使用的總頻率看來，變項使用頻率最高前五名變項為；腫瘤擴散程度＞淋巴結個數＝腫瘤大小＞病理分級＝淋巴結侵犯狀態，在八個模式中都沒被納入的變項為，首次惡性腫瘤、婚姻狀態。

在決策樹預測模式中，與向前法邏輯斯迴歸在變項納入數目有相似的趨勢，納入變項的數量從模型1往模型8增加。在此八個模型中，變項使用頻率最高前八名變項為，腫瘤大小＞腫瘤進行外科手術類型＞腫瘤個數＝陽性淋巴結個數＝婚姻狀態＝淋巴結侵犯狀態＝腫瘤擴散程度＝診斷年齡，在八個模式中都沒被納入的變項為腫瘤病變程度。

綜合以上所論，在所有三種預測模式中預測子宮頸癌患者存活的共同因子較重要的相同變項為；腫瘤擴散程度、淋巴結侵犯狀態、腫瘤大小，較不重要的變項為首次惡性腫瘤、腫瘤個數。

本研究所用的3個資料探勘模型，各有其有缺點。類神經網路模型可以容易進行非線性的輸入變項模型建構，較為有彈性，但過程各個輸入變項對輸出變項的關係為「黑盒子」，不易解釋。邏輯斯迴歸則相反，結果解釋容易，但處理變項間的交互作用和非線性模型較為

麻煩，且必須符合傳統的有統計的參數分配假設，限制較多。決策樹在產生判斷規則的能力上，是最好的，但該模型在處理連續數值資料，均必須化為類別或序位型態，不同的數值切點，對結果和其效能影響很大。未來的研究，必須考量資料特性及研究目的來決定採用的分析模型。

討　論

Snow 2001的研究，比較類神經網路與邏輯斯迴歸用在預測大腸癌治療後，五年存活的情形（活或是不活），類神經網路的內部驗證AUC結果為87.6%；而邏輯斯迴歸的內部驗證AUC為82%，類神經網路內部驗證AUC結果優於邏輯斯迴歸，與本研究相符。在Eftekhar 2005研究也有相同結果，他的研究利用臨床資料，比較類神經網路與邏輯斯迴歸在預測頭部創傷病人（n＝1,271）五年後的死亡情形預測結果，類神經網路與邏輯斯迴歸的內部驗證AUC與內部驗證準確率結果分別為，0.9646、0.9538；0.9509、0.9637，內部驗證AUC結果優於邏輯斯迴歸。

Delen 2005研究中，比較三種資料探勘模式：類神經網路、邏輯斯迴歸、決策樹，在預測乳癌病患五年存活情形比較，研究的內部驗證結果顯示決策樹表現最好（準確率：93.6%），類神經網路次之（準確率：91.2%），邏輯斯迴歸最後（準確率：89.2%）。本研究許多地方有效法Delen的研究，樣本都是來自SEER資料庫，重要變項與預測模式的選擇大都相同，但研究的癌症種類不同，Delen為乳癌，本研究為子宮頸癌，且三種預測模式的設定也不相同。在模型3～8中，其三種預測模型的內部驗證準確率結果趨勢與Delen結果相符，決策樹表現最好，類神經網路次之，邏輯斯迴歸最後，推論為樣本數多寡所致，當樣本數愈大時，決策樹的內部驗證準確率會最好。

Terrin 2003的研究中，利用同一筆資料，探討分別在五個不同資料分佈（線性、曲線、非線性非曲線、二類別、兩個二類別交互項）的情形下利用類神經網路、邏輯斯迴歸（standard logistic regression-LR1）、具次方項邏輯斯迴歸（piecewise-linear and quadratic terms logistic regression-LR2）、分類樹（Classification trees）建造出模型，並帶入同一筆新資料（非訓練模型的資料）對這些模型做外部驗證的測試；來探討三種演算法的表現情形。研究結果為LR2的外部驗證結果最好，其AUC最高可達0.769，外部驗證整體表現依序為LR2＞LR1＞類神經網路＞分類樹，其外部驗證AUC分別為0.734~0.769、0.713~0.741、0.703~0.724、0.667~0.682。在Terrin研究中，兩種不同類型的邏輯斯迴歸其外部驗證AUC結果都較類神經網路好，分類樹最差，但是之間的外部驗證AUC值並沒有很大差異，本研究三種預測模式的外部驗證AUC值很難看出其大小趨勢，但是大都是以決策樹表現較差，且三者之間的差值也不大。

根據Matheny 2005的研究中，採用七篇不同論文有關預測皮冠狀動脈介入治療（PCI: percutaneous coronary intervention）的病人其在院死亡情形的邏輯斯迴歸預測模式，再放入Brigham and Women's Hospital （Boston, MA）2002/01/01~2004/09/30的資料到此七個模型中，作外部驗證。其外部驗證的AUC

結果範圍在0.82~0.90，研究的樣本數最小為2,804，最大為100,253，其研究的外部驗證AUC結果都較本研究來高許多。Matheny研究中可以看到邏輯斯迴歸有好的外部驗證結果，表示邏輯斯迴歸建造的模型在運用在不同的資料預測上有良好的預測效果，但本研究的三種外部驗證的AUC都很低，邏輯斯迴歸的外部驗證AUC為0.6351~0.7060，也許還有其更好的邏輯斯迴歸設定可以增加邏輯斯迴歸的外部驗證結果，未來研究中可以加強。

在Delen 2005的研究中，探討類神經網路、邏輯斯迴歸與決策樹，用於預測乳癌存活情形，其中有對類神經網路變項相對重要性作敏感度分析（sensitivity analysis），其較重要變項的前五名為，病理分級＞臨床分期＞放射治療類型＞腫瘤個數＞淋巴結個數，較不重要的變項為種族、組織學型態、腫瘤擴散程度[14]。

預測模型是臨床實證資料對於現代醫學中最具實質意義的生物統計分析之一，尤其是近年因應資料探勘技術在各領域的興起，也使得所謂類似資料探勘想法的實證醫學（evidence based medicine）在衛生醫藥護理的研究及實作需求有日益增加的趨勢。本研究利用SEER癌症登記資料進行子宮頸癌患者存活預測資料探勘的實作，可以作為臨床估計癌患者存活情形的實證參考，同時，本研究更進一步在預測模型用於未來子宮頸癌患者（不屬於訓練樣本群）預測存活的準確率作深入的研究，這是預測模型建立要外推運用於非模型訓練個案及真正應用於實務時，校正預測存活準確率的重要依據。

本研究使用資料探勘技術中的類神經網路、邏輯斯迴歸、及決策樹等分析，進行子宮頸癌患者存活預測模型的建立；並探討其外部驗證的準確率，作為預測模型外部通用性效能的參考。此外，為了釐清訓練樣本數目大小對預測模型建立及外推時準確率的影響，本研究也設計模型1~8，依照年度遞增的8個訓練樣本組合；並分別以其下年度為外部驗證的測試資料來研究探討訓練樣本數大小的影響，以釐清並證實Sargent 2001研究中所提樣本數大小會影響預測模型的效能。

在模型1~3的預測結果的AUC與準確率的表現，類神經網路會優於決策樹與邏輯斯迴歸；而在模型4~8，決策樹會優於類神經網路與邏輯斯迴歸。邏輯斯迴歸的模型在結果上的趨勢大致上都與類神經網路相似。內部驗證的AUC與準確率，都以模型1表現較好。外部驗證的AUC與準確率都以模型7表現較好。

建置好的模型，如果沒有做外部驗證的話，會錯估模型預測能力的AUC與準確率參數的高低。大體上，內、外部驗證準確率與AUC差值，都以決策樹的差值為最大，表示用決策樹建造的模型，很需要再進行外部驗證，以確定模型適用於非訓練樣本資料時的預測能力。

大體而言，外部驗證準確率的結果，以邏輯斯迴歸表現較好，但與其他兩者預測模式的結果相差不大；大體而言，外部驗證AUC的結果，以類神經網路表現較好。若想要得到較好的外部驗證結果，訓練樣本可以取過去的2~3年以上的資料。

研究限制

此研究的邏輯斯迴歸變項的納入，沒有考慮交互作用，可能會使邏輯斯迴

歸的預測能力受到影響。每個模型的測試組不一樣，造成模型間比較上的不公平。為了將求效率本研究的類神經網路訓練採用快速，可能會降低類神經網路模型的外部適用性。每個模型的外部驗證樣本都不同，所不同模型在比較上會不公平。本研究的三種演算法的內部設定是以得到較好的外部驗證結果為依據，但往往有好的外部驗證結果下，內部驗証結果會較差，所以再做內部驗證結果比較上可能會有偏差。

未來展望

未來研究中，外部驗證資料要為一樣的樣本，在做模型外部通用性比較將會更公平。在資料刪0除方面，不採用刪除所有含有遺漏值的個案的方法，看是否可以得到較好的模型預測能力。根據能得到較好外部驗證結果的文獻去作模型設定，以找出最佳的預測模式設定。

參考文獻

1. 世界衛生組織，http://www.who.int/whr/2002/annex/en/. 3-24, 2005.
2. 行政院衛生署全國衛生統計資訊網：臺灣地區死因統計資料，http://www.doh.gov.tw/statistic/index.htm. 3-24, 2005.
3. Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP: External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. J Clin Epidemiol 2003; 56: 721-9.
4. Anonymity, http://www.gct.ntou.edu.tw/Lab/aiwww/neural.html. 3-24, 2005.
5. Baxt WG: Complexity, chaos and human physiology: the justification for non-linear neural computational analysis. Cancer Lett, 1994; 77: 85-93.
6. Baxt WG: Application of artificial neural networks to clinical medicine. Lancet 1995; 346: 135-8.
7. Tu JV: Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996; 49: 1225-31.
8. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M: A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001; 34: 28-36.
9. Accessed online April 20, 2006, at: http://download.microsoft.com/download/6/b/3/6b3eb4a5-1ba1-4e37-a501-73b977f9a5c8/021606_SQLServer2005WEB-DT.ppt.
10. Accessed online April 20, 2006, at: http://ir.hit.edu.cn/download/qinbing_01.ppt.
11. Accessed online April 20, 2006, at: http://gim.unmc.edu/dxtests/ROC3.htm.
12. Snow PB, Kerr DJ, Brandt JM, Rodvold DM: Neural network and regression predictions of 5-year survival after colon carcinoma treatment. Cancer 2001; 91 (8 Suppl): 1673-8.
13. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E: Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005; 5: 3.
14. Delen D, Walker G, Kadam A: Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005; 34: 113-27.
15. Matheny ME, Ohno-Machado L, Resnic FS:

Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform 2005; 38: 367-75.

16. http://seer.cancer.gov/manuals/CD2.SEERDic.pdf
17. http://seer.cancer.gov/manuals/EOD10Dig.pub.pdf
18. 何子銘、盧瑜芬、許家瑋等：運用三種資料探勘方法預測子宮頸癌存活情形之比較。台灣家醫誌 2006; 16: 192-203。

Predicting Cervical Cancer Survivability: A Comparison of Three Data Mining Methods

Yu-Tieng Chang[1], Chi-Ming Chu[1], Wu-Chien Chien[1], Yu-Ching Chou[1], Tsang Yang[1]
Yu-Fen Lu[1], Chian-You Pai[2], Lu Pai[1] , Thomas Wetter[4]
Chien-An Sun[1] and Ching-Hui Loh[3]

The purpose of the study was to compare the performances of an artificial neural network(ANN), decision tree (C5), and logistic regression (LR) for predicting the 5-year survivability of cervical cancer and their external validation for generalization.

The data was collected from SEER (Surveillance, Epidemiology, and End Results) of the NCI (National Cancer Institute) in the United States during the years 1973~2000. There were 156,502 cases with 72 variables. After the data was cleaned, there were 2,022 cases and 18 variables remaining during years 1988~1996. The dataset was divided into 8 categories of training sets and test sets, according to the year the patients were diagnosed. The 8 training sets were applied to three algorithms: 1) ANN, 2) C5, and 3) LR to build 8 models. The parameters of performance of the models were accuracy and AUC (Area under the ROC curve) for predicting 5-year survivability of cervical cancer patients.

ANN had the best internal validation of the AUC and accuracy (AUC, 0.9392; accuracy, 0.9474) on model 1 and the best external validation of the AUC (0.6455) on model 7. ANN and C5 outperformed LR with respect to internal validation. ANN and LR both performed better than C5 in the external validation of the AUC.

All in all, algorithms of ANN and LR performed better for external generalization, and algorithms of ANN and C5 performed more accurately for classification.

（*Taiwan J Fam Med 2007; 17: 222-38*）

[1]School of Public Health, National Defense Medical Center, National Defense University; Departments of Pathology[2], Community Medicine[3], Tri-Service General Hospital, Taipei, Taiwan; [4]Department of Medical Informatics, University of Heidelberg, Heidelberg, Germany.

Received: August 7, 2006; Accepted: October 28, 2007.