

研究原著

以資料探勘技術預測健康檢查大腸息肉之風險因子

游雅雯^{1,*} 鄭博文¹ 林宏茂^{1,2} 梁玉芬³

摘要

目的：本研究透過妥善的健康檢查，以達早期發現早期治療。本研究建立大腸息肉風險因子與大腸鏡異常之發現預測模式，以提供醫師作為臨床輔助診斷，減少侵入性檢測，降低檢查成本。

方法：本研究收集西元 2009 年 1 月至西元 2009 年 12 月間，台灣中部個案醫院健康檢查中心做過大腸鏡篩檢之民眾資料，共計分析 809 筆。本研究以一般檢查、血液檢查、生化檢查—血脂肪、尿液檢查、癌胎抗原檢查、免疫法糞便檢查共 28 項變數，以大腸息肉與否為依變數，使用決策樹分類方法進行分析。

結果：本研究結果發現，以免疫法糞便檢查的預測績效最好，其訓練資料的 Az 值為 0.902，測試資料的 Az 值為 0.879。癌胎抗原檢查次之，其訓練資料的 Az 值為 0.897，測試資料的 Az 值為 0.843。

結論：本研究結果顯示，決策樹分類方法適用於醫學大腸息肉之健檢資料，可有效探勘其重要變數。本研究結果可提供醫院健康管理中心作為輔助決策。

» 關鍵字：大腸鏡、息肉、決策樹、風險因子

¹ 雲林科技大學工業工程與管理研究所

² 國立台灣大學醫學院附設醫院雲林分院 大腸直腸外科

³ 彰化基督教醫院 健康管理中心

* 通訊作者：游雅雯 聯絡地址：雲林縣斗六市大學路 3 段 123 號

e-mail : g9821802@untech.edu.tw

投稿日期：2012 年 06 月 01 日

接受日期：2012 年 08 月 09 日





前 言

由於生活環境的快速變遷，工作壓力的增加及飲食習慣改變，國人的健康狀況也慢慢亮起紅燈，使得癌症的人口數逐年上升。根據行政院衛生署所公佈的統計數字顯示，惡性腫瘤自西元 1982 年起，即列為國人死因之首位，其中大腸癌是常見的消化道惡性腫瘤 [1]。根據行政院衛生署所公佈的統計數字顯示，西元 2010 年大腸癌已經躍升國內癌症發生率第一，從西元 1982 年約有 2,855 個大腸直腸癌新發生個案，至西元 2009 年約有 12,488 個大腸直腸癌新發生個案；西元 2010 年有 4,680 人死於大腸直腸癌，其死亡人數逐年快速上升 [1]。

在大腸癌的風險因子研究中，大多數以飲食、生活型態、家族史、基因等做探討 [2-4]，鮮少有以預防角度的健檢項目來探討其可能的相關因子，因此本研究利用一般健檢資料進行初步分析，檢視大腸癌篩檢中的資訊。而在這些大量的資料中，可能隱藏著許多重要資訊，單憑一般的經驗與簡單的統計分析是很難挖掘出當中的資訊及各項關聯，因此若能使用廣泛應用於各領域的資料探勘（Data mining），可以有系

統、有效率的運用各項相關資料，將其資訊轉換為有用知識，產生決策 [5-6]。

目前醫學上應用資料探勘的研究，張語恬等人以類神經網路（artificial neural network）、決策樹（decision tree）以及邏輯迴歸（logistic regression）三種演算法造出模型，以 AUC（area under the ROC curve）、準確率（accuracy），作為演算法預測能力評估，並找出可以得到良好子宮頸癌五年存活預測結果的模型 [7]。蔡蕙如等人應用類神經網路與迴歸樹進行肝癌的分類模式 [8]。Ramirez 等人運用資料探勘於愛滋病患者特徵之研究，主要以決策樹來篩選變數，然後再以類神經網路來辨識愛滋病患者 [9]。陳銘樹等人提出新的分析概念，考慮因子變化量的決策樹模型，其預估的準確率最高，總正確率高達 82.51%，發現空腹血糖、血壓、三酸甘油脂、肝功能、白血球等五項檢驗數據為第二型糖尿病的統計上的重要危險因子 [10]。Dursun 等人使用 Logistic 回歸、決策樹、類神經網路三種方法，探勘乳癌病患的存活能力，分析變數與存活情形間的關係，進行預測病患存活能力，結果發現以決策樹表現最好、其次為類神經網路、

Logistic 迴歸 [11]。洪淑芬使用支援向量機、貝氏於潛在語意索引在生醫文件分類的應用上，結果發現以支援向量機表現最好 [12]。Belluacha 用簡單貝氏法、ANN、決策樹 C4.5，應用於乳癌存活率，其結果以決策樹最佳 [13]。由以上文獻可知，資料探勘於醫學上皆有不錯的分類績效，其中決策樹更是不錯的方法，故為本研究選用。

大腸癌一直是大家廣為討論的議題，而大腸癌早期並沒有什麼特殊症狀，必須要靠一些檢查才能夠發現，像是糞便潛血（Fecal occult blood test）及大腸鏡檢查（Colonoscopy）等 [14]。為了降低罹癌的機會，最佳的預防之道，就是定期接受大腸鏡的檢查，但一般民眾對此檢查，莫不面有難色，甚至退避三舍，因為許多人對於大腸鏡檢查前的清腸瀉藥、檢查中引起的不適感、大腸鏡可能造成的風險等等而為之卻步。因此本研究目的為使用中部某個案醫院之健檢資料，以 5 cross-validation 劃分方式，利用決策樹進行大腸息肉預測的分析，並探勘出對大腸息肉相關的影響因子，提供民眾在還未做大腸癌篩檢時，從一些基本體檢資料先進行初部檢視，以建議民眾是否需要再進一步作大腸鏡篩檢，免受大腸鏡檢查前的各種

不適，節省檢查成本。並提供醫院健康管理中心作為輔助，並作為預防醫學領域中大腸癌之醫療決策支援之防治工具，做好預防及健康保健工作。

方 法

一、研究對象

本研究收集西元 2009 年 1 月至西元 2009 年 12 月間，台灣中部個案醫院健康檢查中心，做過大腸癌篩檢之民眾資料，共計收集 946 筆；每一位民眾皆做過大腸鏡檢查，調查項目包括一般理學檢查（性別、年齡、身體質量指數、腰圍、收縮壓、舒張壓）、血液檢查（白血球、紅血球、血色素、血球容積比、紅血球體積、平均血紅素、平均血球血紅素濃度、血小板）、生化檢查—血脂肪（空腹血糖、糖化血色素、總膽固醇、高密度膽固醇—脂蛋白、低密度膽固醇—脂蛋白、三酸甘油脂）、尿液檢查（尿糖、膽紅素、尿酸鹼值、尿蛋白、尿液紅血球、尿液白血球）、癌胎抗原檢查（癌胎抗原指數）、免疫法糞便檢查（糞便潛血）。其大腸息肉資料皆由大腸直腸科主治醫師診斷，將大腸鏡檢查結果區分為大腸鏡檢查正常者和大腸鏡結果發現息肉者兩類。

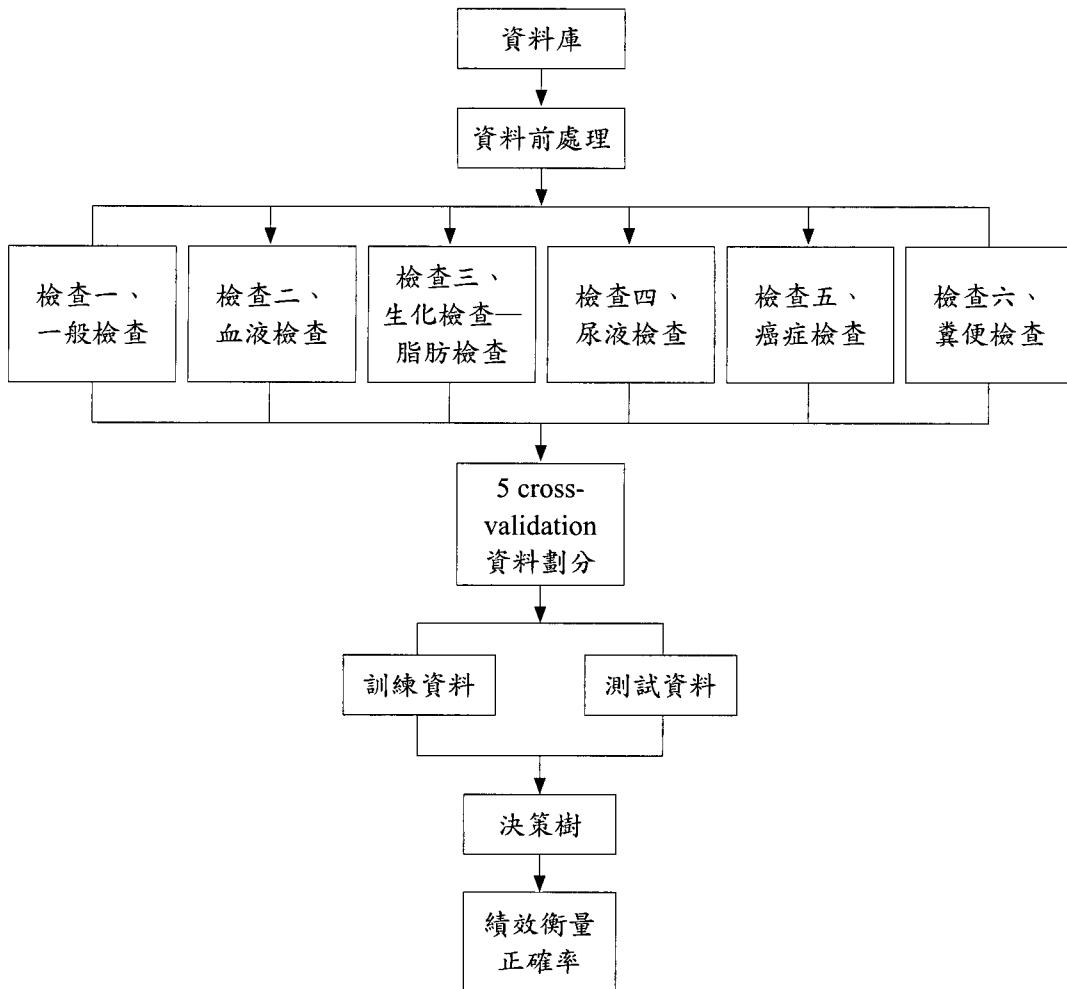




二、演算流程

本研究演算流程以資料前處理（Preprocessing）、資料劃分與分類（Classifier）為主，如圖一。首先將健康檢查資料進行資料前處理，將資料不全的予以剔除，並請主治醫師進行大腸鏡診斷歸類，接下來將資料以健康檢

查項目組合決定輸入變數，並且採用 5 cross-validation 劃分方式，再將訓練資料（Training data）建構出決策樹分類器，使用測試資料（Testing data）進行驗證，以正確率為績效評估值。本研究資料以 Weka3.6.6、SPSS18.0 進行分析。



圖一 研究流程

(一) 資料前處理

在進行資料探勘之前，為了讓資料更適合進行探勘工作，必須將不符合定義之資料予以剔除。曾憲雄等人提出首先整合資料變數，當資料變數具有重複性、不一致性時，將之刪除；其次資料如果不完整或有遺漏值（Missing value），處理方法分為兩種，第一種方法為採取人工填補法，亦即打電話、寄信詢問，若無法補齊則以平均值取代；第二種則為直接忽略法 [15]。本研究則採用直接忽略法，此方法適用於資料量大，是最直接也是最快速的處理方法，可避免不必要的資料，以達高品質探勘結果。本研究資料刪除 9 筆未完成健康檢查項目者、16 筆排除過去有大腸疾病者、112 筆為清腸未完全無法判斷者及其他紀錄未完全資料，研究資料共計分析為 809 筆。

接下來將前處理後的資料，進行資料劃分。本研究採用 Steinback 等人提出的 K 次交叉驗證（K cross-validation）劃分方式進行資料劃分 [16]。K 次交叉驗證，將資料分割成 K 個子樣本，以 K-1 個樣本用來訓練，一個樣本被保留作為測試模型的數據。這個方法的優勢在於，重複以隨機產生的子樣本進行訓練和測試，每次的結果

驗證一次，5 次交叉驗證為常用的交叉驗證，其為將資料視為一個族群，將資料集分成 5 份，輪流將其中 4 份當作訓練資料，1 份當作測試資料，5 次的結果的平均值作為對演算法好壞的估計。再依健康檢查項目組合，使用決策樹進行訓練與測試之分類。

(二) 健康檢查項目組合

本研究健康檢查項目如下：

1. 一般理學檢查

將性別、年齡、身體質量指數、腰圍、收縮壓、舒張壓，六項一般理學檢查變數，以 5 cross-validation 資料劃分，以決策樹於一般理學檢查預測績效評估。觀察一般理學檢查項目的預測準確度，探勘項目中關鍵性指標。

2. 血液檢查

將白血球、紅血球、血色素、血球容積比、紅血球體積、平均血紅素、平均血球血紅素濃度、血小板，八項血液檢查變數，以 5 cross-validation 資料劃分，以決策樹於血液檢查預測績效評估。觀察血液項目的預測準確度，探勘項目中關鍵性指標。

3. 生化檢查—血脂

將空腹血糖、糖化血色素、總膽固醇、高密度膽固醇—脂蛋白、低密度





膽固醇一脂蛋白、三酸甘油脂，六項血脂肪檢查變數，以 5 cross-validation 資料劃分，以決策樹於血脂肪檢查預測績效評估。觀察血脂肪項目的預測準確度，探勘項目中關鍵性指標。

4. 尿液檢查

將尿糖、膽紅素、尿酸鹼值、尿蛋白、尿液紅血球、尿液白血球，六項尿液檢查變數，以決策樹於尿液檢查預測績效評估。觀察尿液項目的預測準確度，探勘項目中關鍵性指標。

5. 癌胎抗原檢查

將癌胎抗原指數（Cartinoembryonic antigen, CEA），以 5 cross-validation 資料劃分，以決策樹於癌症檢查預測績效評估。觀察癌症項目的預測準確度，驗證癌胎抗原指數是否為大腸息肉檢查指標之一。

6. 免疫法糞便檢查

將糞便潛血，此項免疫法糞便檢查變數，以 5 cross-validation 資料劃分，以決策樹於糞便檢查預測績效評估。觀察糞便項目的預測準確度，驗證糞便潛血是否為大腸癌檢查指標之一。

三、分類與績效評估

決策樹（decision tree）是資料探勘中建立分類模式（classification

models）的方法之一，經常使用於醫療診斷與預測。在資料探勘中，決策樹是相當受到歡迎的分類和預測工具，決策樹主要以樹狀結構呈現，藉由內部節點（internal node）表示某屬性的測試、分支（brache）代表條件測試的結果、葉節點（leaf node）表示分類結果，由上而下構成容易瞭解且分類快速的樹狀結構，表達其所包含的知識結果，讓使用者可以容易了解。

而在分類中，同質性（homogeneous）高的類別分配和節點不純度（impurity）愈低是比較好的。因此有一些測量指標能夠測量分類的好壞，像是 Gini Index、Entropy、Gain Ratio 等，以下舉例 Entropy 算式 [15,16]：

$$\text{Entropy}(t) = - \sum_j p(j | t) \log p(j | t);$$

(NOTE : $p(j | t)$ is the relative frequency of class j at node t)

$$\text{GAIN}_{\text{split}} = \text{Entropy}(p) - \left(\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) \right);$$

(NOTE : Parent Node, p is split into k partitions; n_i is number of records in partition i)

「增益比值（Gain Ratio）」的計算公式來取代原有的分岔準則，但最根

本的內容還是透過亂度（Entropy）的概念作為決策樹的分岔準則。

$$GainRATIO_{split} = \frac{GAIN_{split}}{SplitINFO};$$

$$SplitINFO = -\sum_{i=1}^k \frac{n_i}{n} \log \frac{n_i}{n};$$

(Parent Node, p is split into k partitions
 n_i is the number of records in partition i)

藉由以上測量資訊量的數值來計算出個別資訊量，表達集合資料的複雜程度。資訊量也可以衡量資料複雜度、純度，資訊量越高代表其資料複雜度越高，使純度提升越多的變數就是有效變數。決策樹的解釋能力很高，能夠讓使用者清楚的了解其結果與關鍵屬性；因此對於運用在新資料或是不明性質的情況是非常有用的。一般常見的決策樹演算法有 C4.5、CART、CHAID[5, 6]。

本研究以 Andrew (2003) 的分類標準，完成模型建構後，必須評估其預測績效，一個較佳的預測模型之測試 Az 值愈高，表示該模型的預測績效愈好。操作特徵曲線（Receiver Operative Characteristic Curve, ROC；簡稱 Az 值）指標來評估模型績效，以計算面積之方式來評估效能，得到最客觀之評估結果；ROC 曲線藉由畫出假陽性（False Positive, FPF）與真陽性

（True Positive, TP）的函數圖形，代表兩個特定族群的相對關係範圍是從 0 到 1 之間。如果改變相對應的界線數值，就可以得到很多組 (X, Y) 數值，將這些座標點標示出來，我們可以在平面中畫出一條曲線，形成 ROC 曲線（Az 值）。特異度（Specificity）=

$$1 - \frac{FP}{TN + FP}$$
、敏感度（Sensitivity）=

$1 - \frac{TP}{FN + TP}$ 。因此透過「ROC 曲線下的面積」代表診斷工具所預測的機率有多大，預測的機率越大代表診斷工具相對就越好（0.9~1.0 優 Excellent；0.80~0.9 良 good；0.70~0.80 普通 fair；0.60~0.70 Poor；0.50~0.60 最差 Fail）[15,16]。

結 果

一、基本資料

本研究資料共計分析為 809 筆，描述性資料如表一所示，民眾的年齡從 20 歲至 82 歲，其中大腸鏡檢查正常者有 460 位（男性 48.7%、女性 51.3%）；大腸鏡結果發現息肉者有 349 位（男性 61.3%、女性 38.7%）。

健康檢查項目與大腸鏡檢查結果



中，大腸鏡結果發現息肉者平均年齡高於大腸鏡檢查正常者 ($51.5 > 48.0$, $p < .00$)，身體質量指數 ($24.6 > 23.9$, $p < .01$)、腰圍 ($84.6 > 81.4$, $p < .00$)、收縮壓 ($124.1 > 121.3$, $p < .02$)、舒張壓 ($80.2 > 78.7$, $p = .06$) 方面，皆以大腸鏡結果發現息肉者偏高。在血液項目中，白血球 ($p < .03$)、血球容積比 ($p < .02$)、紅血球體積 ($p < .02$)、平均血紅素 ($p < .01$)，結果顯示大腸鏡檢查結果具顯著影響。在生化檢查—血脂肪方面，空腹血糖 ($p < .04$)、糖化血色素 ($p < .02$)、總膽固醇 ($p < .05$) 與高密度膽固醇 ($p < .03$)、三酸甘油脂 ($p < .03$) 具顯著影響。在癌胎抗原指數 ($2.7 > 1.9$, $p < .04$) 方面則以大腸鏡結果發現息肉者偏高。在糞便潛血方面，其中大腸鏡檢查正常者（陽性 3.1%、陰性 59.2%）；大腸鏡結果發現息肉者（陽性 3.5%、陰性 34.2%），結果顯示具顯著影響，如表一。

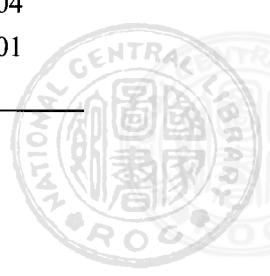
將大腸直腸科主治醫師以專家經驗方式，納入健康檢查變數 28 個因子，依照檢查項目區分作為獨立變數，以大腸息肉與否為依變數，以決策樹進行預測分析。健康檢查項目危險因子與預測結果，如表二。(1) 檢查一、

一般理學檢查，其訓練資料的 Az 值為 0.817，測試資料的 Az 值為 0.758，敏感度為 0.849，特異度為 0.789，探勘因子為性別、年齡、身體質量指數、收縮壓；(2) 檢查二、血液檢查，其訓練資料的 Az 值為 0.832，測試資料的 Az 值為 0.765，敏感度為 0.915，特異度為 0.829，探勘因子為紅血球體積、血球容積比、血小板數；(3) 檢查三、生化檢查—血脂肪，其訓練資料的 Az 值為 0.812，測試資料的 Az 值為 0.760，敏感度為 0.792，特異度為 0.768，探勘因子為空腹血糖、總膽固醇、高密度膽固醇—脂蛋白、三酸甘油脂；(4) 檢查四、尿液檢查，其訓練資料的 Az 值為 0.842，測試資料的 Az 值為 0.803，敏感度為 0.903，特異度為 0.742，探勘因子為尿糖、尿蛋白；(5) 檢查五、癌胎抗原檢查，其訓練資料的 Az 值為 0.897，測試資料的 Az 值為 0.843，敏感度為 0.995，特異度為 0.750，探勘因子為癌胎抗原指數；(6) 檢查六、免疫法糞便檢查，其訓練資料的 Az 值為 0.902，測試資料的 Az 值為 0.879，敏感度為 0.984，特異度為 0.817，探勘因子為糞便潛血。



表一 健康檢查大腸鏡結果

變數	(%)	大腸鏡判讀正常 (n = 460)	大腸鏡判讀異常 (n = 349)	p 值
性別	女	29.2	16.7	< 0.001
	男	27.7	26.4	
年齡		48.04 ± 12.42	51.50 ± 12.67	< 0.001
身體質量指數		23.92 ± 3.30	24.69 ± 3.53	0.01
腰圍		81.40 ± 9.45	84.67 ± 9.79	< 0.001
收縮壓		121.32 ± 14.84	124.17 ± 14.01	0.02
舒張壓		78.73 ± 9.03	80.27 ± 10.06	0.06
白血球		5.88 ± 1.70	6.21 ± 1.78	0.03
紅血球		4.70 ± 0.53	4.73 ± 0.48	0.28
血色素		14.21 ± 1.53	14.15 ± 1.46	0.41
血球容積比		41.06 ± 4.12	41.74 ± 4.03	0.02
紅血球體積		87.18 ± 7.80	88.87 ± 6.47	0.02
平均血紅素		30.18 ± 3.09	31.11 ± 2.55	0.01
平均血球血紅素濃度		34.57 ± 0.83	34.64 ± 0.69	0.14
血小板		247.11 ± 52.65	247.38 ± 56.6	0.95
空腹血糖		95.19 ± 24.68	100.15 ± 30.25	0.04
糖化血色素		5.51 ± 0.85	5.73 ± 1.07	0.02
總膽固醇		199.08 ± 38.27	201.50 ± 45.84	0.50
高密度膽固醇—		60.99 ± 16.13	57.40 ± 15.12	0.03
低密度膽固醇		118.34 ± 33.43	117.70 ± 34.96	0.84
三酸甘油脂		118.51 ± 209.99	117.79 ± 85.72	0.03
尿糖 (%)	陽性	2.4	2.9	0.19
	陰性	60.4	34.3	
膽紅素		0.71 ± 0.32	0.75 ± 0.25	0.25
尿酸鹼值		5.29 ± 0.54	5.33 ± 0.57	0.23
尿蛋白 (%)	陽性	2.3	2.5	0.27
	陰性	60.3	34.9	
尿液紅血球	(/ μL)	6.33 ± 15.09	9.7 ± 51.34	0.22
尿液白血球	(/ μL)	9.00 ± 36.63	8.93 ± 46.96	0.98
癌胎抗原指數		1.90 ± 1.39	2.78 ± 7.52	0.04
糞便潛血 (%)	陽性	3.1	3.5	0.01
	陰性	59.2	34.2	





表二 大腸鏡息肉預測準度整理表

	訓練資料	測試資料	Sensitivity	Specificity	重要因子
檢查一、一般檢查	0.817	0.758	0.849	0.789	性別、年齡、身體質量指數、收縮壓
檢查二、血液檢查	0.832	0.765	0.915	0.829	紅血球體積、血球容積比、血小板
檢查三、生化檢查—血脂	0.812	0.760	0.792	0.768	空腹血糖、總膽固醇、高密度膽固醇—脂蛋白、三酸甘油脂
檢查四、尿液檢查	0.842	0.803	0.903	0.742	尿糖、尿蛋白
檢查五、癌症檢查	0.897	0.843	0.995	0.75	癌胎抗原指數
檢查六、糞便檢查	0.902	0.879	0.984	0.817	糞便潛血

討 論

現今國民的生活水準逐漸提高，民眾對就醫觀念已經從以往的治療醫學變成預防醫學、健康醫學；「預防重於治療」，如果能預防得宜或能早期發現早期治療，皆可減輕龐大的醫療負擔；因此本研究以預防角度的健檢項目來探討其可能的相關因子，利用一般健檢資料進行初步分析，檢視大腸檢查中的資訊。本研究將其探勘結果與醫師進行討論，結果討論如下：

一、一般理學檢查

ACS (American Cancer Society)

等組織，建議成人於五十歲開始接受大腸癌篩檢 [18]。本研究發現之危險因子，身體質量指數、性別、年齡、收縮壓，與過去學者研究之危險因子相同 [2-4]。本研究大腸息肉檢查發現，男性較女性為多，且平均年齡也較女性來得高，因此建議男性大腸癌篩檢時間可以提早於女性；其中，身體質量指數過高易造成肥胖，而肥胖是心血管疾病的高危險因子；近來研究發現，肥胖不僅容易有腸息肉，甚至容易身陷罹患腸癌的高風險 [17]；根據研究發現，肥胖者的息肉發生率比一般人高，且腺瘤性息肉的再發率也較高 [18]。一般來說，腸癌大多數由腺瘤性息肉所發展，因此研究

顯示，腸道內有息肉的人，其罹患大腸癌的機率會是一般人的好幾倍 [19]。現今國人飲食逐漸西化且缺乏運動習慣，發現腸息肉的比例逐年升高，性別、年齡、身體質量指數、肥胖、壓力過大等皆可能是造成之因素 [2-4]，因此民眾應進行篩檢，並養成正確飲食及運動習慣。

二、血液檢查

過去研究指出大腸癌、息肉有疲勞、貧血等等徵兆 [1-3]，因此血液檢查也需要注意，可能造成貧血等等。平均血球容積指標代表每個紅血球平均體積，即紅血球體積平均值，平均血球容積值高表示紅血球體積過大，常見於缺乏維他命 B12 和葉酸之貧血、巨紅血球症；而平均血球容積值低即表示紅血球體積較小，小紅血球性貧血（缺鐵、慢性病、地中海型貧血），紅血球和血色素及血球容積比有密切的關聯，根據這些數值，大約可以判斷貧血的種類 [20]。而血小板具有黏著和凝集能力，負責止血的工作。當血小板數量減少時，容易引起出血傾向；相反的，血小板過多時，也是有止血作用降低的危險，所以血小板在 10 萬個以下，或 50 萬個以上時，就要在有血液內科的專科

醫院接受必要的精密檢查 [20]。一般來說，早期進行大腸鏡檢查，已被證實可以降低死亡率，但對於民眾做大腸鏡篩檢仍是個大問題；而血液檢驗為非侵入性檢驗，不僅可協助醫師利用血液樣本進行偵測，更能提供民眾簡單、便宜方式，提早發現。

三、生化檢查—血脂

大腸癌之形成，與肉食、高脂肪食物、低纖維食物之飲食習慣有密切關係。肥胖會引發其他疾病如糖尿病、高血壓、心血管疾病、癌症等 [17]。根據美國國家膽固醇教育計畫，如果腹圍肥胖、三酸甘油脂偏高、血中高密度脂蛋白膽固醇偏低、血壓偏高、空腹血糖偏高等五項指標中，具有三項或三項以上便符合代謝症候群，此時就需特別注意健康已亮紅燈 [21]。代謝症候群的危險因子有很多，如性別、年齡、種族、生活型態、肥胖及遺傳等 [21-23]，許多研究指出代謝症候群會增加心血管疾病及第二型糖尿病等慢性疾病的危險性，且會造成其死亡率上升，因此已成為重要的公共衛生議題 [22]。這群危險因子與台灣十大死因中，腦血管疾病、心臟病、糖尿病、高血壓性、慢性疾病等密切相關 [24]，其主要為現代人吃的好卻





動的少，引發高血糖、高血壓和三酸甘油脂高的比例也跟著增加，因此容易罹患代謝症候群。國內外已有研究發現，代謝症候群為大腸腺瘤風險增加因素，易引發大腸癌 [19,22,23]，因此及早預防疾病的發生是有其必要性。

四、尿液檢查

一般尿液檢查，尿蛋白陽性反應，並不代表病患一定有病理性尿蛋白。健康檢查為單一時間一次檢查，並非收集整天 24 小時所有的尿液檢測尿蛋白濃度，可能會因病患少喝水或流汗多，使得尿液相當濃縮，影響結果。因此如果尿蛋白為陽性反應，建議直接找專科醫師進一步做二十四小時尿液總蛋白量的檢測。若尿液蛋白質檢查呈陽性結果，則可能有罹患有慢性腎臟疾病，腎臟發炎、腎病症候群、妊娠毒血症等狀況 [20]。而尿糖是一般用來初步篩檢糖尿病的一種方法，若尿液中有葡萄糖時則應考慮是否有糖尿病，或是因腎臟疾病所引起；美國腸胃病協會的年會研究顯示，糖尿病患者的大腸癌癌前病變風險可能會增加 [24]；此方面的研究結果並不完全一致，因果關係尚須經由更多實驗證實。

五、癌胚抗原檢查

癌胚抗原指數是癌症檢查的參考指標，其會因疾病與當時的身體狀況等因素而所有變化；除了罹患大腸癌的病人指數會增加外，在胃癌、乳癌、胰臟癌、肺癌等等癌症都會增高；抽煙者數值也會有偏高現象；因此，指數升高或正常，不一定表示罹患癌症或體內沒有癌細胞。目前研究認為 CEA 不適宜作為初步篩檢之用 [25,26]，但於臨床上，對於民眾 CEA 指數高於標準值時，仍然會建議民眾進行大腸篩檢，進一步做檢查 [3,5]。整體來說，雖然癌胚抗原對於大腸癌的診斷特異性不高，不過當此值高時，已有癌轉移的現象居多；因此癌胚抗原指數在臨床應用上，最有幫忙的是在大腸癌追蹤，治療後指數下降，但經過一段時間後又高起來就要懷疑是否有復發，目前大都只用於手術前及手術後的一種評估，對於術後癌症復發及轉移的偵側較有價值 [27,28]。

六、免疫法糞便檢查

過去實驗已證實糞便潛血檢查可篩檢到早期癌症，降低死亡率 [2,3,4]。國民健康局鼓勵民眾踴躍作糞便潛血的篩檢，凡是超過 50 歲的民眾，每 2 年做 1 次糞便潛血檢查，希望民眾透過早



期篩檢，達到早期發現早期治療的目的[3]。潛血的反應，是偵測肉眼看不出的大便出血現象；陽性反應的原因可能是食物中含有過氧化酵素活性、胃腸道出血：如痔瘡或憩室或胃腸道息肉或腫瘤出血。陰性反應可能是沒有胃腸道出血，或出血未超過一定上限或腫瘤病變的出血為間歇性；因此糞便檢查呈現陽性，必須進一步接受大腸鏡檢查，因為糞便潛血檢查呈現陽性者有二至四成在大腸鏡檢查時，會發現腺瘤性息肉或惡性腫瘤[2]，因此糞便潛血可作為第一線初步篩檢的工具。

早期發現，早期治療是治療所有癌症的目標。在台灣，仍然常常可見延誤的病例，譬如直腸癌當做痔瘡，大腸癌當做腸胃炎，而未能做適當的診斷與治療。所以在五十歲以上的人，尤其是有症狀或屬於高危險群的人，定期做大腸鏡檢查是有必要的。本研究所建立之大腸息肉預測模型，無論是一般理學檢查、血液檢查等，均有七、八成以上的分類正確性，從風險因子與模型當中，更能清楚了解與大腸息肉的關係。本研究可提供民眾在還未做大腸篩檢時，從一些基本體檢資料先進行初步檢視，以建議民眾是否需要再進一步作大腸篩檢，免受大腸鏡檢查前的各種不適，節

省檢查成本。並提供醫院健康管理中心作為輔助，並作為預防醫學領域中大腸癌之醫療決策支援之防治工具，做好預防及健康保健工作。

七、結論

本研究之結果發現，決策樹分類方法適用於醫學大腸息肉之健檢資料，可有效探勘其重要變數。探勘影響因子於一般檢查、血液檢查、生化檢查—血脂、尿液檢查、癌胎抗原檢查、免疫法糞便檢查變數發現，疾病與疾病之間彼此有其關連性，其危險因子也有其共同性，只要身體指數發生異常，連帶而來的是許多疾病的產生，像是肥胖、代謝症候群等，因此建議民眾除了定期做健康檢查，也可以從一般身體指標管制自身健康，做好預防及健康保健工作。

八、建議與限制

為了有效提倡大腸癌篩檢，各相關機關應多方宣傳、積極鼓勵民眾進行篩檢，並提倡運動、保持標準體重、健康飲食等有益健康活動來降低大腸癌的發生，運動有助於加速大腸的蠕動，縮短糞便通過大腸的時間，減少大腸內膜接觸糞便內致癌物的機會。





本研究資料受限於樣本來源以及時間搜集限制，後續可再針對更多資料或相關因子、相關資料庫進行研究，例如：內分泌檢查變數、血清免疫檢查等變數，如此一來可以提供更完善的模式，得到更好的決策效果。本研究分析僅以大腸息肉為研究，後續可再針對息肉良惡性及其他研究議題來定義所選取的依變數，使研究更加完善。未來可以使用其他預測方法例如：類神經預測、灰預測、模糊理論或是其他結合演算預測方法，進行方法的比較；本研究對象為國內中部個案醫院，故外推性將受限。

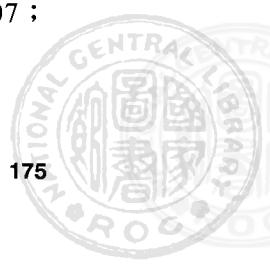
致 謝

本研究部份經費承國科會（編號：NSC 99-2221-E-224-033）補助，本文作者特此致謝。

參考文獻

- 行政院衛生署：86~100 年歷年死因統計。行政院衛生署，2011.08.15 摘自 http://www.doh.gov.tw/CHT2006/DM/DM2_2.aspx?now_fod_list_no=10326&class_no=440&level_no=3

- no=3。
- 劉易承、宋鴻樟、謝玲玲、唐瑞平、葉志清：大腸直腸癌之風險預測模式與風險指標。台灣衛誌 2008；27(1)：1-12。
 - 曾嘉慶、李嘉龍、吳啟華：大腸直腸腫瘤的篩檢與追蹤：文獻回顧與最新指引。內科學誌 2009；20：506-513。
 - 張簡俊榮：台灣大腸直腸癌的流行病學。中華癌醫會誌 2008；24(3)：143-147。
 - Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. 1st ed. Hoboken, NJ: Wiley-Interscience, 2003; 1-28.
 - Pang-Ning Tan, Michael Steinback, Vipin Kumar. Introduction to Data Mining. 1st ed. Boston: Addison Wesley, 2006; 42-105.
 - 張語恬、朱基銘、簡戊鑑等人：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。台灣家醫誌 2007；17(4)：222-238。
 - 蔡蕙如、柯明中、張偉斌、劉德明：應用類神經網路與分類迴歸樹於肝癌分類模式。北市醫學雜誌 2007；4(8)：658-667。



9. Ramirez JCG, Cook DJ, Peterson LL, Peterson DM. Temporal Pattern Discovery In Course-Of-Disease Data. IEEE Engineering In Medicine And Biology Magazine 2000; 19(4): 63-71.
10. 陳銘樹、王建智、王麗雁：應用決策樹演算法以探究高科技員工潛在的糖尿病之危險因子。健康管理學刊 2008；6(2)：135-146。
11. Dursun Delen, Glenn Walker, Amit Kadarm. Predicting breast cancer survivability: a comparison of three data mining methods. Computer and Information Science 2005; 34(2): 113-127.
12. 洪淑芬、葉吉原、董信煌：以潛在語意索引為特徵的生醫文件檢索系統。第十屆人工智慧與應用研討會，國立高雄大學，2005，高雄，台灣：國立高雄大學、中華民國人工智慧學會。
13. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Computer and Information Science 2006; 58(13): 10-110.
14. American Cancer Society. Colorectal Cancer facts and figures 2011-2013. American Cancer Society, 2012.04.28 from <http://www.cancer.org/Research/CancerFactsFigures/ColorectalCancerFactsFigures/index>
15. 曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯：資料探勘。初版。台北：旗標出版，2005；1-48。
16. Tan PN, Steinback M, Kumar V. Introduction to Data Mining. 1st ed. Boston: Addison Wesley, 2006; 6-36.
17. Giovannucci E. Metabolic syndrome, hyperinsulinemia, and colon cancer: a review. American Journal of Clinical Nutrition 2007; 86(3): 836-842.
18. Liu CS, Hsu HS, Li CI, et al. Central obesity and atherogenic dyslipidemia in metabolic syndrome are associated with increased risk for colorectal adenoma in a Chinese population. BMC Gastroenterology 2010; 10: 51.
19. Kim JH, Lim YJ, Kim YH, et al. Is metabolic syndrome a risk factor for colorectal adenoma? Cancer Epidemiol Biomarkers Prev 2007; 16(8): 1543-1546.
20. 青木誠孝、青木芳和：精準解讀健康檢查報告書。初版。台灣：瑞昇出版，2006；20-63。
21. 莊世杰、林秀美、林尚志：糖尿病合





- 併高脂血症之積極治療—新版美國國家膽固醇教育計畫建議之啟示。台灣醫界 2002 ; 45(6) : 21-22。
22. Wang YY, Lin SY, Lai WA, Liu PH, Sheu WH. Association between adenomas of rectosigmoid colon and metabolic syndrome features in a Chinese population. *J Gastroenterol Hepatol* 2005; 20(9): 1410-1415.
 23. Liou JM, Lin JT, Huang SP, et al. Screening for Colorectal Cancer in Average-Risk Chinese Population Using a Mixed Strategy with Sigmoidoscopy and Colonoscopy. *Dis Colon Rectum* 2007; 50(5): 630-640.
 24. Saydah SH, Platz EA, Rifai N, Pollak MN, Brancati FL, Helzlsouer KJ. Association of markers of insulin and glucose control with subsequent colorectal cancer risk. *Cancer* Epidemiol Biomarkers Prev 2003; 12(5): 412-418.
 25. Duffy MJ, van Dalen A, Haglund C, et al. Tumour markers in colorectal cancer: European Group on Tumour Markers(EGTM)guidelines for clinical use. *Eur J Cancer* 2007; 43(9): 1348-1360.
 26. Meeker WR Jr. The use and abuse of CEA test in clinical practice. *Cancer* 1978; 41(3): 854-862.
 27. 曾屏輝、林鴻儒、邱瀚模、李百卿、吳明賢、陳明豐：從實證醫學角度看自費健康檢查。內科學誌 2009；20：532-543。
 28. Chiu HM, Lin JT, Shun CT, et al. Association of metabolic syndrome with proximal and synchronous colorectal neoplasm. *Clin Gastroenterol Hepatol* 2007; 5(2): 221-229.



Original Articles

Applying Data Mining Technology to Predict Risk Factors for Colon Polyps on Physical Examination

Ya-Wen Yu^{1,*} Bor-Wen Cheng¹ Hong-Mau Lin^{1,2} Yu-Fen Liang³

Abstract

Objective: The aim of this study was to establish a predictive model for risk factors for colon polyps to help physicians reduce invasive testing and the costs of examinations.

Methods: Data were collected from a community hospital physical examination center located in Central Taiwan during the period of January 2009 to December 2009. We analyzed data from 809 patients who received colonoscopies. Risk factors associated with colon polyps were determined by using decision tree algorithms.

Results: The results showed that the best predictor was the presence of fecal occult blood. The receiver operative characteristic curve (Az value) of training data was 0.902, and the Az value of the test data was 0.879. The second best predictor was the Carcinoembryonic Antigen the Az value of training data was 0.897, and the Az value of the test data was 0.843.

Conclusions: Decision tree classification technology was an effective way to use physical examination data to make a decision index with regard to colon polyps. It was easy to determine and provided a highly accurate predictive model for the need for colonoscopy.

» **Key words:** Colonoscopy, Polyps, Decision tree, Risk Factors

¹ Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, No.123, Sec. 3, University Rd., Douliou City, Yunlin County 640, Taiwan, R.O.C.

² Department of Surgery, National Taiwan University Hospital Yunlin Branch, Taiwan, R.O.C.

³ Department of Health Exam Center, Changhua Christian Hospital, Taiwan, R.O.C.

* Correspondence author.

E-mail: g9821802@yuntech.edu.tw

Received: Jun 6, 2012

Accepted: Aug 9, 2012

