南華大學科技學院資訊管理學系

碩士論文

Department of Information Management
College of Science and Technology
Nanhua University

Master Thesis

以決策樹建立大腸癌之五年存活預測模式

Using Decision Tree to Build Five-year Survival Prediction Model for Colorectal Cancer

蘇家毅

Jia-Yi Su

指導教授：邱宏彬 博士

Advisor: Hung-Pin Chiu, Ph.D.

中華民國 109 年 7 月

July 2020

南 華 大 學

科技學院資訊管理學系

碩 士 學 位 論 文

以決策樹建立大腸癌之五年存活預測模式

Using Decision Tree to Build Five-year Survival Prediction Model for Colorectal Cancer

研究生：蔡 承翰

經考試合格特此證明

口試委員：林迪瀚

陳張宗榮

葉明憲

邱宏彬

指導教授：邱宏彬

系主任(所長)：陳

口試日期：中華民國 109 年 7 月 2 日

南華大學資訊管理學系碩士論文著作財產權同意書

立書人：蘇家毅 之碩士畢業論文

中文題目：

以決策樹建立大腸癌之五年存活預測模式

英文題目：

Using Decision Tree to Build Five-year Survival Prediction Model for Colorectal Cancer

指導教授：　邱宏彬　博士

學生與指導老師就本篇論文內容及資料其著作財產權歸屬如下：

■ 共同享有著作權

□ 共同享有著作權，學生願「拋棄」著作財產權

□ 學生獨自享有著作財產權

學　生：蘇家毅（請親自簽名）

指導老師：邱宏彬（請親自簽名）

中 華 民 國　109 年 7 月 21 月

誌　　謝

感謝在指導教授邱宏彬博士的指導之下，幫我開啟走向資料探勘的相關領域，一直到我完成本篇論文，算是對資料探勘領域有更進一步的研究與貢獻。期間感謝邱教授的指導與鼓勵及研究室每個夥伴的協助與提醒，讓我在論文寫作上得到一些建議、修正和指導，才使本篇論文得以完成。

另外，感謝南華大學資訊管理學系的每個成員，包括了系上各個領域的教授在課程上給予我相關的知識得以在論文寫作解決問題，而系上行政人員時常也在處理我們課務相關問題幫我們解決了不少疑難雜症，讓我們能順利的度過難關。

最後，感謝朋友與家人們在我研讀南華碩士學位到順利完成學業的期間，給予我各種支持與鼓勵，讓我有著堅持下去的動力以致我一步步完成向各種領域知識的挑戰。

蘇家毅　謹致於

南華大學資訊管理學系

2020 年 7 月

以決策樹建立大腸癌之五年存活預測模式

學生：蘇家毅　　指導教授：邱宏彬 博士

南 華 大 學　資訊管理學系碩士班

摘　　要

本研究目的在探究以數據分析與資料探勘在大腸癌存活預測模式的運用，應用於決策樹的演算法，由預測準確率以及對預測結果的解釋能力做為演算法的評估指標。本研究以嘉義大林慈濟醫院癌症登記資料庫與中醫門診記錄資料庫，自西元 2007 年至 2014 年期間就診的大腸癌患者納入研究。研究中再分為西醫病患與純西醫病患及中西醫病患三組與 12 個變項進行資料分析及每組區分為 3 種分別為 0.8/0.2、0.7/0.3、0.6/0.4 訓練集/測試集，用決策樹之演算法來比較預測五年存活準確度。結果顯示加入中醫輔助療法之病患存活率下降較純西醫治療存活率下降速度較為緩慢，顯示使用中醫輔助治療確實可延緩死亡。

關鍵字：大腸癌、存活預測、中醫用藥、決策樹

Using Decision Tree to Build Five-year Survival Prediction Model for Colorectal Cancer

Student：SU, JIA-YI Advisor： Hung-Pin Chiu, Ph.D.

Department of Information Management
The Graduated Program
Nan-Hua University

ABSTRACT

The purpose of this research is to explore the application of data analysis and data exploration in the prediction model of colorectal cancer survival. The algorithm applied to the decision tree uses the prediction accuracy and the ability to interpret the prediction results as the evaluation index of the algorithm. This research uses the cancer registration database of Chiayi Dalin Tzu Chi Hospital and the database of Chinese medicine outpatient records. Patients with colorectal cancer who were treated from 2007 to 2014 were included in the study. The study was further divided into three groups of Western medicine patients, pure Western medicine patients, and Chinese and Western medicine patients with 12 variables for data analysis, and each group was divided into 3 types, respectively 0.8/0.2, 0.7/0.3, 0.6/0.4 training set /Test set, Use decision tree algorithm to compare the accuracy of predicting five-year survival.

The results showed that the survival rate of patients with Chinese medicine adjuvant therapy decreased more slowly than that of pure Western

medicine treatment. It shows that the use of Chinese medicine auxiliary treatment can indeed delay death.

Keywords: Colorectal cancer, Survival prediction, Chinese herbal medicine, Decision tree

目錄

著作財產權同意書 i

誌謝 ii

摘要 iii

ABSTRACT iv

目錄 vi

圖目錄 viii

表目錄 x

第一章、緒論 1

第一節、研究背景 1

第二節、研究動機與目的 2

第二章、文獻探討 3

第一節、大腸癌的危險因子 3

第二節、大腸癌的治療方式 4

第三節、大腸癌結合中醫輔助治療 5

第三章、研究設計 6

第一節、研究流程 6

第二節、資料集的簡述 7

第三節、資料集的前置處理流程 12

第四節、研究架構 17

第五節、研究方法 18

第四章、研究結果 23

第一節、大腸癌病患之存活分布情形 23

第二節、探討相關因子對大腸癌存活預測的影響 26

第五章、結論 33

第一節、結論 33

參考文獻 34

圖目錄

圖 1 研究流程圖 6

圖 2 資料集的前置處理流程圖 12

圖 3 篩選中西醫組、純西醫組的病患流程圖 16

圖 4 不同治療存活預測 17

圖 5 研究樣本篩選流程 19

圖 6 存活分析預測流程示意圖 22

圖 7 大腸癌分組存活曲線圖 25

圖 8 所有病患組的訓練集/測試集 0.6/0.4 組 27

圖 9 所有病患組的訓練集/測試集 0.7/0.3 組 27

圖 10 所有病患組的訓練集/測試集 0.8/0.2 組 27

圖 11 所有病患組的訓練集/測試集 0.6/0.4 組之決策樹示意圖 28

圖 12 所有病患組的訓練集/測試集 0.8/0.2 組之決策樹示意圖 28

圖 13 純西醫治療組的訓練集/測試集 0.6/0.4 組 29

圖 14 純西醫治療組的訓練集/測試集 0.7/0.3 組 29

圖 15 純西醫治療組的訓練集/測試集 0.8/0.2 組 30

圖 16 純西醫治療組的訓練集/測試集 0.8/0.2 組之決策樹示意圖 30

圖 17 中西醫合治組的訓練集/測試集 0.6/0.4 組 31

圖 18 中西醫合治組的訓練集/測試集 0.7/0.3 組..... 31

圖 19 中西醫合治組的訓練集/測試集 0.8/0.2 組..... 31

圖 20 中西醫合治組的訓練集/測試集 0.8/0.2 組之決策樹示意圖 32

表目錄

表 1 癌登資料欄位定義表 8

表 2 中醫門診紀錄欄位定義表 11

表 3 ICD-O-3 西醫原發部位編碼表 13

表 4 ICD-9-CM 中醫診斷代碼表 15

表 5 大腸癌所有病患組存活分布資料表 24

表 6 大腸癌純西醫治療組存活分布資料表 24

表 7 大腸癌中西醫合治組存活分布資料表 24

第一章、緒論

第一節、研究背景

根據我國行政院衛生署統計，罹患大腸癌發生人數曾創下連續超過10年盤踞10大癌症之首位癌症成為台灣十大死因榜上已經多年[25]，甚至大腸癌是世界上癌症死亡原因排名的前茅，也是威脅現代人生命的健康殺手。

隨者生活水平型態的轉變，速食產業的興起，使得國民飲食習慣有極大改變，導致國民攝取過量高油高鈉且低纖的食物。由於飲食習慣的轉變，不僅使得國民罹患心血管等三高相關疾病的升高，甚至罹患大腸癌的發生率也逐漸增加。

本研究希望透過決策樹存活分析的分析方式[23]，探討罹患大腸癌的病患，在治療方式為：純西醫治療因子或是結合中醫的中西醫治療因子，對於病患的存活率是否有影響，並且在台灣健保癌症登記資料庫內的病患數據資料也逐漸增加，有了可靠且具有大量真實性之現有資料，該如何分析與預測相關因素可能對於癌症所造成的影響，是現今醫學研究一個重大環節。

第二節、研究動機與目的

近代以來國民患者就醫主要以西醫為主，但傳統的中醫治療對國人健康有著莫大的幫助，所以使用中醫治療的患者還是佔有一定的比例。而健保資料庫具有相當多優勢，但在相關研究上仍然有限制。例如資料的正確性、資料遺漏問題。

此外，建立癌症登記為癌症防治的基本要件，所以對於癌症患者的資料記錄都相當嚴格，才能提高資料可信度。且癌症患者資料產生過程複雜需耗費長時間的收集及分析，且在紀錄資料後再進行核對，所以在患者資料的可信度以及嚴謹度高於健保資料庫[4]。

最後，在大林慈濟醫院有完善與龐大的癌症登記資料庫，且在中醫治療也有完善的中醫門診紀錄資料庫。因此，本研究將該院的兩大資料庫針對就診方式的不同，探討治療相關因子對於存活之相關性。希望能透過此分析，找出最有效治療大腸癌或延長患者存活時間之方法。

第二章、文獻探討

第一節 、大腸癌的危險因子

罹患大腸癌發生的主因，目前尚未有確切的答案，但大多數研究指出認為與飲食及作息有高度相關性，以下列出四項危險因子：

一、年齡：據研究大腸癌的比較容易發生在中老年人口五十歲以上，然而現在國人生活習慣轉變，暴飲暴食，加上不固定且不足的運動，導致這幾年在壯年三四十幾歲的患者也越來越多，甚至出現更年輕病患。

二、飲食：研究發現肉類的過多攝取是影響形成大腸癌的因子之一，且攝取紅肉高於白肉還多的人，罹患機會將更增加。因此形成大腸癌與長期攝取高脂肪食物有相關，應適量攝取高脂肪食物。

三、遺傳因素：研究分析發現具有大腸癌家族病史的血親，罹患大腸癌機率比一般人高。統計分析上，若家族中有一位一等親患有大腸癌，則其本人罹患機率為一般人的二到四倍;如果有二位一等親患有大腸癌，則危險性提高至三到六倍[7]。

四、作息：因長時間的不固定輪班工作與身體休息，所以身體機能無法固定按時運作，導致腸道引發大腸息肉，甚至惡化成腸癌。

第二節、大腸癌的治療方式

大腸癌的治療方式主要有三種：手術治療、化學治療、放射治療。一般醫師會根據臨床檢查與癌症分期的結果來建議病人選擇適當的治療方式，大略介紹於下：

一、手術治療：手術切除長久以來都是多數癌症治療的主要方式，因為在癌症初期只有完全切除才有治癒的機會，同理對大腸癌也是。然而根據癌細胞位置的不同，所執行的手術方式也有些不一樣，除了將腸道癌細胞切除外，通常也會將附近的組織移除掉[14]。

二、化學治療：採取藥物治療癌症的方法，是一種全身性的治療。化學治療的原理是藉抑制或破壞癌細胞的生長等方式來完成治療癌症的目地[28]。

三、放射治療：放射線治療即是運用較高能量的放射線照射腫瘤，約是診斷用 的 X 光之能量的數百或數千倍，產生生物效應破壞癌細胞，防止癌細胞的生長與分裂[17]。

第三節、大腸癌結合中醫輔助治療

現今研究中，顯示癌症病患併用中醫輔助療法的情形相當普遍，高達七至九成，在癌症病患有生理或心理不適時有相當高的比例會尋求中醫或其他替代醫療的協助。然而，多數對中醫藥本身以及其使用時機較少有完整性的概念來協助治療。癌症診斷後，經由專科醫師評估後，接受西醫的手術、放療、化療或標靶治療後，因癌症或癌症治療所引起的症狀包括腸胃不適、眩暈、掉髮，以及療程中常出現口乾、口腔潰瘍、腹瀉、便秘、疲倦、惡病質、體重減輕、食慾不振、精神不濟以及癌性疼痛等不適的症狀，以上這些症狀大都會使患者的生活品質下降、生理心理問題增加，甚至可能無法使西醫治療完善，而使其預後惡化。目前研究證據指出配合使用中醫輔助療法對於癌症相關症狀來說是有統計上的意義，期待透過中醫的體質調理，提升患者的生活品質，進一步協助病患順利完成整個西醫療程而改善患者預後[12][13]。

第三章、研究設計

第一節、研究流程

見圖 1，本研究以大腸癌相關資訊及中西醫治療研究現況等進行文獻探討，確立研究架構進行研究設計，並透過人體試驗委員會合法取得大林慈濟醫院癌症登記資料庫與中醫門診紀錄資料庫，進行資料庫前置處理後，透過病患納入與排除條件將資料匯入資料預測分析軟體 RapidMiner，進行存活分析預測，最後呈現報告。

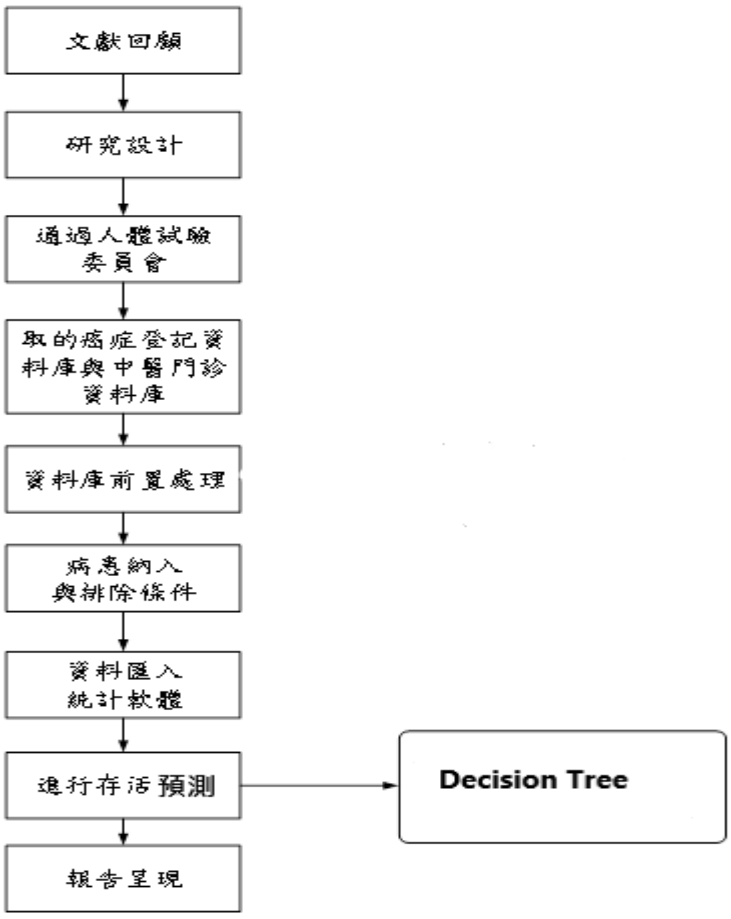

圖 1 研究流程圖

第二節、資料集的簡述

本研究資料庫建置的資料集是由大林慈濟醫院中醫部醫師所提供，分別為西醫癌症登記資料、中醫門診紀錄資料，分述如下：

壹、西醫癌症登記資料

衛生署為規劃癌症防治工作，在民國 68 年以行政命令方式針對 50 床以上醫院建立癌症登記系統，要求申報新發癌症個案的流行病學和診斷治療摘要資料。癌症登記自民國 85 年 7 月由衛生署委託『癌症登記中心』進行癌症資料收集，在癌症防治法於民國 92 年 5 月 21 日公布，該法第十一條規定「為建立癌症防治相關資料庫，癌症防治醫療機構應向中央主管機關所委託之學術研究機構，提報新發生之癌症個案與期別等相關診斷及治療資料。」自此確立了癌症登記的法源依據。在本研究中，針對癌症登記資料集所使用的資料欄位：ID、SEX、戶籍地代碼、診斷年齡、個案分類、最初診斷日、原發部位、測性、臨床 T、臨床 N、臨床 M、臨床期別組合、病理 T、病理 N、病理 M、病理期別組合、首次手術日、放療開始日、化學治療開始日、荷爾蒙治療開始日、最後聯絡日期、存活狀態、死亡原因[16]。本研究試圖從上述所提及的資料欄位，訂定研究變項如表 1，進行存活預測分析之探討。

表 1 癌登資料欄位定義表

欄位名稱	欄位定義
ID	記錄個案的身分證統一編號，用來辨識個案。
SEX	可作為各癌症部位性別比例及預後之比較。
戶籍地代碼	為個案流行病學之地域資料，並可作為癌症群聚或環境因素研究分析。
診斷年齡	有助於個案的確認，且對於統計分析癌症相關資料時，年齡常是一個重要的因素。
個案分類	在做治療和存活分析時，本欄位將個案分為可分析和不可分析個案兩類。進而提供申報醫院進行研究個案之選擇。
最初診斷日	可計算癌症最初診斷日期至完成分期或開始治療的時間間隔。
原發部位	依 ICD-O-3 腫瘤部位碼進行原發部位編碼。
側性	確認癌症起源於成對器官或身體的某一側。
臨床 T	指原發腫瘤大小或侵犯程度，腫瘤分期判斷以臨床主責醫師為主。
臨床 N	指是否有區域淋巴結的轉移和轉移的範圍，腫瘤

	分期判斷以臨床主責醫師為主。
臨床 M	指是否有遠端轉移，腫瘤分期判斷以臨床主責醫師為主。
臨床期別組合	基於臨床 T、N 和 M 來決定疾病於解剖部位上的侵犯程度，腫瘤分期判斷以臨床主責醫師為主。
病理 T	指原發腫瘤大小或侵犯程度，腫瘤分期判斷以病理醫師為主
病理 N	指是否有區域淋巴結的轉移和轉移的範圍，腫瘤分期判斷以病理醫師為主
病理 M	指是否有遠端轉移，腫瘤分期判斷以病理醫師為主
病理期別組合	基於臨床 T、N 和 M 來決定疾病於解剖部位上的侵犯程度，腫瘤分期判斷以病理醫師為主。
首次手術日	記錄在任何醫療機構，最早針對癌症進行手術的日期。
放療開始日	記錄在申報醫院的首次療程中，進行放射治療的開始日期。
化學治療開始日	記錄個案在申報醫院化學治療開始的日期。

荷爾蒙治療開始日	記錄個案在申報醫院荷爾蒙治療開始的日期。
最後聯絡日	記錄個案的最後聯絡日期或是死亡日期，作為個案追蹤和治療結果研究之用。
存活狀態	記錄個案「最後聯絡或死亡日期」的存活狀態。
死亡原因	作為癌症存活率統計分析時的死因分類，以區分非癌症死亡個案。

貳、中醫門診紀錄資料

本資料庫是大林慈濟醫院建立於中醫部的門診紀錄資料庫，在此資料庫紀錄的內容有：病患就診於該院中醫部的診療資料。資料內容包含：就醫日期、病歷號、個人基本資料、病患自我述說病況、醫師對於病患述說病況的回饋、中醫用藥、醫師對於病患的治療方式以及診斷代碼。針對病患資料萃取，本研究所使用到的資料為：就醫日期、病歷號及診斷代碼，透過本研究會使用到的資料內容，訂定研究變項如表 2，透過該資料庫與上述該院所提供之癌症登記資料庫進行癌症病患的資料樣本篩選，試圖透過資料分析，得知是否能發現結合中醫治療的病患對存活時間的影響。

表 2 中醫門診紀錄欄位定義表

欄位名稱	欄位定義
就醫日期	紀錄該病患就診的日期。
病歷號	記錄個案的身分證統一編號，用來辨識個案。
診斷代碼	ICD-9-CM 國際疾病分類代碼。

第三節、資料集的前置處理流程

本研究利用上述提及之資料集，進行有關存活預測議題的研究，但在資料集內容的部分，並非全員都為本研究所探討的對象，所以必須先將資料集做資料的前置處理，只萃取出本研究所探討之議題的資訊，在資料處理的前置流程有四個如圖 2，分別為：篩選中西醫組、純西醫組的病患、剃除 5 年內存活時間不確定之病患、判斷五年存活，各流程分述如下：

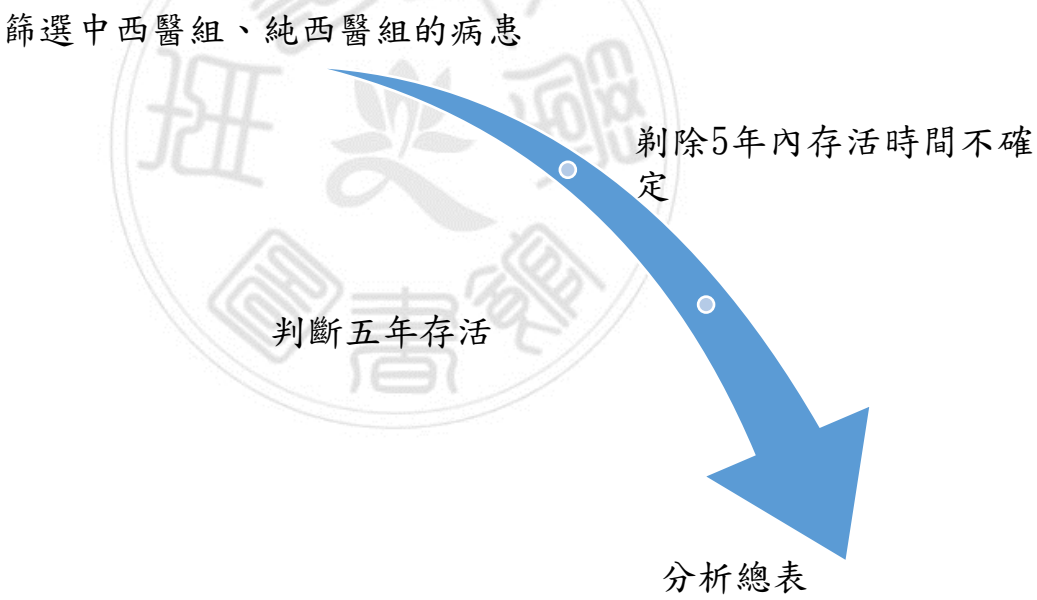

圖 2 資料集的前置處理流程圖

壹、篩選中西醫組、純西醫組的病患

本研究想探討癌症就診方式，如果只接受西醫治療或者是加上中醫

的輔助治療，對於病患的存活時間是否有任何影響。所以將上述所提及的西醫癌症登記資料及中醫門診紀錄資料，進行資料前置處理，將純西醫及中西醫的病患個別篩選出來，再依照與醫師討論的用藥暴露天數的規則，將中西醫未達用藥暴露天數的病患，歸納於純西醫組，最後資料篩選完成後，會得到純西醫組及中西醫組兩組病患，資料篩選流程如下所述：

一、篩選資料集內大腸癌病患及中西醫大腸癌病患

透過大林慈濟醫院提供的癌症登記資料以及中醫門診紀錄兩個資料集，進行研究樣本的萃取。透過ICD-O-3（國際疾病分類腫瘤學）的西醫原發部位編碼如表3以及ICD-9-CM（國際疾病分類標準）的中醫診斷代碼如表4，進行癌症篩選。本研究透過兩張代碼表，將上述兩個資料集內紀錄的大腸癌病患篩選出來，篩選完成後，分別儲存在對應資料表內，資料表為：西醫癌症登記大腸癌病患資料表及中醫門診紀錄大腸癌病患資料，再將有進行西醫及中醫治療的病患篩選出來，得到治療方式為中西醫的大腸癌病患。流程如圖3。

表 3 ICD-O-3 西醫原發部位編碼表

代碼	原發部位英文名稱	原發部位英文名稱
C180	Malignant neoplasm of cecum	盲腸惡性腫瘤

C181	Malignant neoplasm of appendix	闌尾惡性腫瘤
C182	Malignant neoplasm of ascending colon	升結腸惡性腫瘤
C183	Malignant neoplasm of hepatic flexure colon	右曲結腸惡性腫瘤
C184	Malignant neoplasm of transverse colon	橫結腸惡性腫瘤
C185	Malignant neoplasm of splenic flexure	結腸脾（彎）曲部惡性腫瘤
C186	Malignant neoplasm of descending colon	降結腸惡性腫瘤
C187	Malignant neoplasm of sigmoid colon	乙狀結腸惡性腫瘤
C188	Malignant neoplasm of overlapping sites of colon	大腸重疊部位之惡性腫瘤
C189	Malignant neoplasm of colon, unspecified	結腸惡性腫瘤
C199	Malignant neoplasm of rectosigmoid junction	直腸乙狀結腸連接處惡性腫瘤
C209	Malignant neoplasm of rectum	直腸惡性腫瘤
C210	Malignant neoplasm of anus, unspecified	肛門惡性腫瘤
C211	Malignant neoplasm of anal canal	肛管惡性腫瘤
C212	Malignant neoplasm of cloacogenic zone	泄殖腔帶惡性腫瘤
C218	Malignant neoplasm of overlapping sites of rectum, anus and anal canal	直腸、肛門及肛（門）管重疊部位之惡性腫瘤

表 4 ICD-9-CM 中醫診斷代碼表

代碼	診斷代碼名稱英文	診斷代碼名稱中文
153.0	Malignant neoplasm of hepatic flexure colon	右曲結腸惡性腫瘤
153.1	Malignant neoplasm of transverse colon	橫結腸惡性腫瘤
153.2	Malignant neoplasm of descending colon	降結腸惡性腫瘤
153.3	Malignant neoplasm of sigmoid colon	乙狀結腸惡性腫瘤
153.4	Malignant neoplasm of cecum	盲腸惡性腫瘤
153.5	Malignant neoplasm of appendix	闌尾惡性腫瘤
153.6	Malignant neoplasm of ascending colon	升結腸惡性腫瘤
153.7	Malignant neoplasm of splenic flexure	左曲結腸惡性腫瘤
153.8	Malignant neoplasm of other specified sites of large intestine	大腸其他特定部位之惡性腫瘤
153.9	Malignant neoplasm of colon, unspecified	結腸惡性腫瘤
154.0	Malignant neoplasm of rectosigmoid junction	直腸乙狀結腸連接部惡性腫瘤
154.1	Malignant neoplasm of rectum	直腸惡性腫瘤
154.2	Malignant neoplasm of anal canal	肛管惡性腫瘤

154.3	Malignant neoplasm of anus, unspecified	肛門惡性腫瘤
154.8	Malignant neoplasm of rectum, rectosigmoid junction, and anus, other	直腸，直腸乙狀結腸連接處及肛門之惡性腫瘤

二、篩選純西醫大腸癌病患資料

透過上述篩選後的西醫癌症登記大腸癌病患資料表及中醫門診紀錄大腸癌病患資料表，將西醫癌症登記大腸癌病患資料表內有中西醫大腸癌病患資料表的病患排除，即可萃取出在治療大腸癌期間只有使用純西醫治療的病患。流程如圖3。

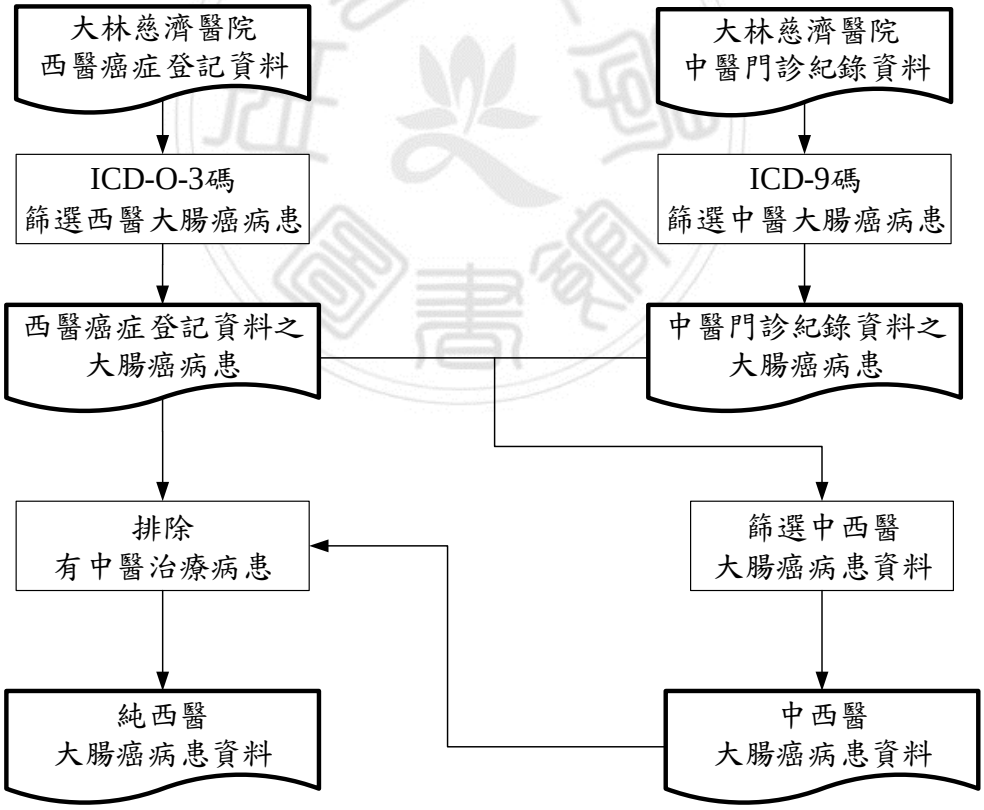

圖 3 篩選中西醫組、純西醫組的病患流程圖

第四節、研究架構

本研究架構論述如下：

研究架構一：以存活預測探討大腸癌病患在不同的就診方式下的影響存活時間的相關因子如圖 4，包括人口學特質（年齡、地域、性別）、疾病特質（期別、共病症）及治療方式（手術治療、化學治療、放射治療）對存活預測是否有影響?

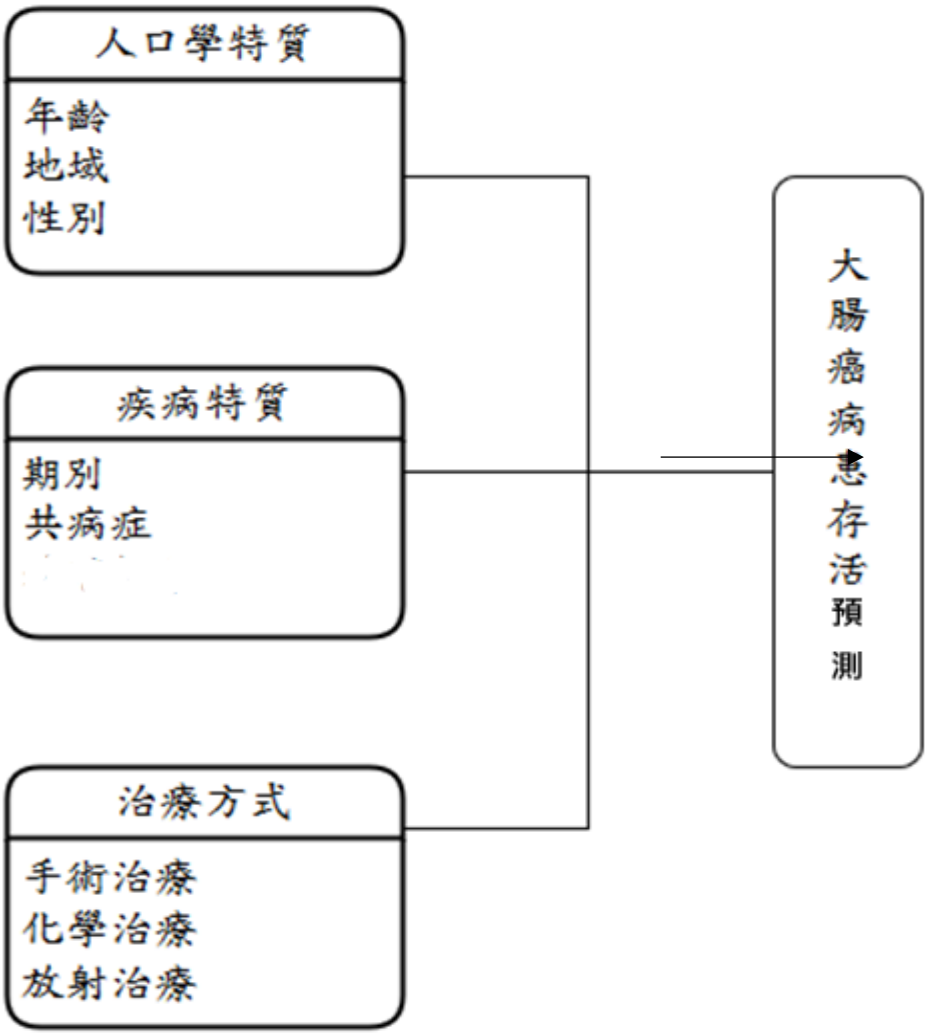

圖 4 不同治療存活預測

第五節、研究方法

壹、研究樣本

本研究樣本來源為嘉義大林慈濟醫院癌症登記資料庫與中醫門診記錄資料庫，自西元2007年至2014年期間就診的大腸癌患者，其納入研究與排除準則如下：

一、納入準則

1.癌症登記資料庫ICD－O－3代碼為C180至C218之大腸癌病患。

2. 中醫門診資料庫ICD－9－CM代碼為153至154.8之大腸癌病患。

3. 追蹤時間≥3個月

二、排除準則

1. 存活狀態空值：空值無法取得存活與否，所以需要排除。

2. 死亡原因代碼為7798：7798代表該病患非死於癌症。

3. 重複資料：由於病患可能罹患一個或多個癌症，可能會有多筆資料，因此以資料內容最完整的為優先納入。

4. 期別不明及零期患者：期別不明無法判斷該病人期別，而零期屬於原位癌，與醫生討論過後只保留侵襲癌的病患。

三、樣本篩選流程

本研究樣本來源流程如圖5。自運用癌症登記資料庫進行西醫與中醫用藥之存活預測以大腸癌為例，此論文研究總個案數為535人再剃除掉五年內存活時間不確定之個案。

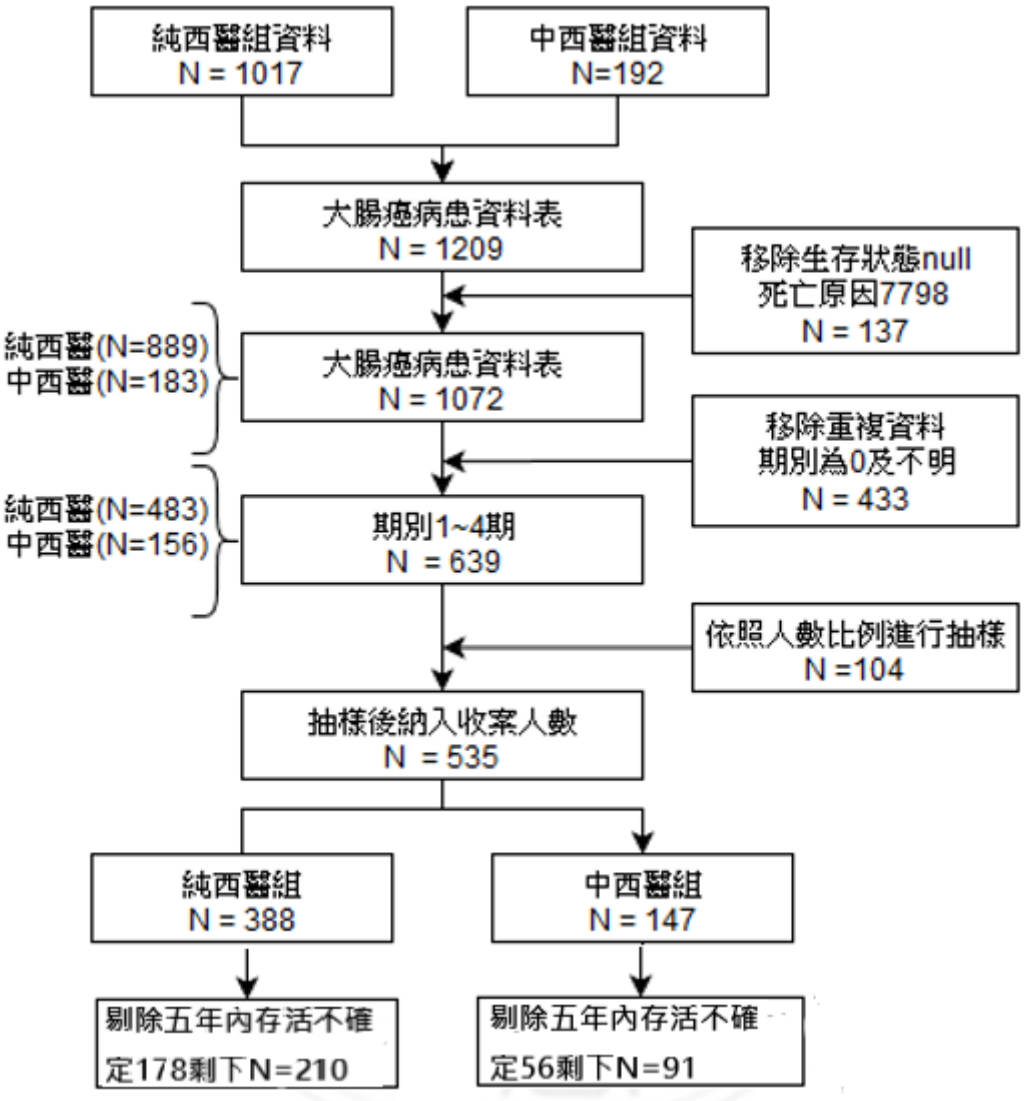

圖 5 研究樣本篩選流程

貳、資料預測分析模型

本研究以資料預測分析軟體RapidMiner進行預測分析與繪製存活預測決策樹模型，建立存活分析預測流程示意圖6，預測模型詳述如下：

1. 決策樹模型（Decision Tree）

建立存活預測流程圖在首先(1)建立一個母資料夾以提供研究資料集儲存的地方，再(2)建立子資料夾一個來儲存資料並且在儲存資料前設定資料格式與個別欄位屬性，(3)建立子資料夾一個來儲存存活分析預測流程。

(4)在建立存活預測流程時所需的工具將來自此區塊選取，(12)在選取建模所需工具後有些特定參數設定時需要在此區塊設定。

流程一開始我們需要告訴這個流程它使用的資料存放在哪裡，所以我們需從(4)區塊搜尋(5)Retrieve將此工具拉自中間設計區域中並在(12)區塊設定資料來源方可輸出資料。

有了資料我們需要進一步選取真正建模所需的資料變數，所以需從(4)區塊搜尋(6)Select Attributes將此工具拉自中間設計區域中與(5)Retrieve輸出接點連接並在(12)區塊設定選取包含了五年存活、就診方式、年齡分組、性別、期別、地域、放療、化療、

手術、共病1、共病2、共病6全部中的12個變數。

根據上述確認了所需的12個變數，下一步則是根據本研究重點設定目標變數給予它特殊身分，所以需要在(4)區塊搜尋(7)Set Role將此工具拉自中間設計區域中與(6)Select Attributes輸出exa接點連接並在(12)區塊選取五年存活變數作為目標變數。

再者，建立模型時我們需要一部分資料當作「訓練集」用來分析，另一部分當作「測試集」用來驗證預測，讓我們結果可以更客觀，所以我們需在(4)區塊搜尋(8)Split Data將此工具拉自中間設計區域中與(7)Set Role輸出exa接點連接並在(12)區塊設定訓練集/測試集切割比例，在此為了找出最後最佳準確性模型，我們將訓練集/測試集比例分成0.6/0.4、0.7/0.3、0.8/0.2三組進行比較。

接下來將正式建立決策樹模型，計算出各個分析變數對於預測目標變數的影響力，所以我們需在(4)區塊搜尋(9)Decision Tree將此工具拉自中間設計區域中與(8)Split Data第一個par輸出接點連接代表使用「訓練集」來訓練模型。

有了對「訓練集」訓練模型接下來也對「測試集」進行預測，所以需在(4)區塊搜尋(10)Apply Model將此工具拉自中間設計區域中並將mod接點與(9) Decision Tree的mod接點連接，將unl接點

與(8)Split Data的第二個par接點連接，將右側mod接點連接最後右側第一res接點。

最後當我們對測試資料作出預測後，如要評斷模型預測能力好不好，就需要一些評斷指標，所以需在(4)區塊搜尋(11)Performance將此工具拉自中間設計區域中並將lab接點連接(10)Apply Model的lab接點，將par接點連接最後右側第二res接點，將exa接點連接最後右側第三res接點。

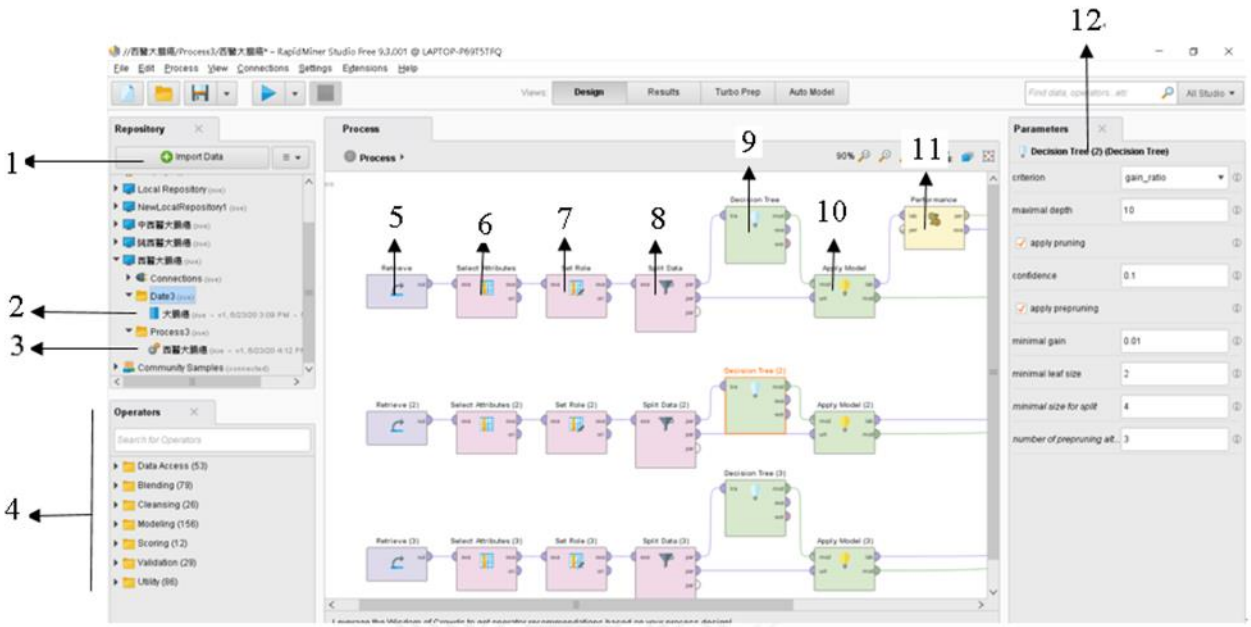

圖 6 存活分析預測流程示意圖

第四章、研究結果

本章將依研究主題敘述其研究結果:第一節為研究大腸癌病患之存活特性分布情形、第二節為探討相關因子對大腸癌存活預測的影響。

第一節、大腸癌病患之存活分布情形

研究結果發現，從表5、表6、表7三張不同分組的存活分布資料表得知病患在不同就診方式與期別不同時存活時間分布之差異，結果發現罹患大腸癌越後期其存活時間越低，而在存活超過五年的病患裡更是幾乎集中在第1、2、3期別。

而從表7大腸癌中西醫合治組存活分布資料表與圖7就診方式不同的面相來看在納入中醫輔助時其整體存活時間偏低，因為納入中醫輔助的病患其一開始就是比較晚期發現的3、4期，所以在病情比較嚴重狀況下選擇納入中醫輔助幫助晚期病患提升治療效果，結果發現在病患病情比較嚴重狀況下，仍然有一定程度的幫助延緩死亡效果。

所有病患組		各期別人數			
存活時間	人數	1 期	2 期	3 期	4 期
0~1	87	1	13	26	47
1~2	50	4	3	13	30
2~3	33	2	7	13	11
3~4	17	1	3	7	6
4~5	9	2	1	4	2
5~	105	27	31	42	5

表5大腸癌所有病患組存活分布資料表

純西醫治療組		各期別人數			
存活時間	人數	1 期	2 期	3 期	4 期
0~1	27	1	4	10	12
1~2	44	4	2	12	26
2~3	28	1	4	12	11
3~4	16	1	3	6	6
4~5	7	1	0	4	2
5~	88	20	27	36	5

表6大腸癌純西醫治療組存活分布資料表

中西醫合治組		各期別人數			
存活時間	人數	1 期	2 期	3 期	4 期
0~1	60	0	9	16	35
1~2	6	0	1	1	4
2~3	5	1	3	1	0
3~4	1	0	0	1	0
4~5	2	1	1	0	0
5~	17	7	4	6	0

表7大腸癌中西醫合治組存活分布資料表

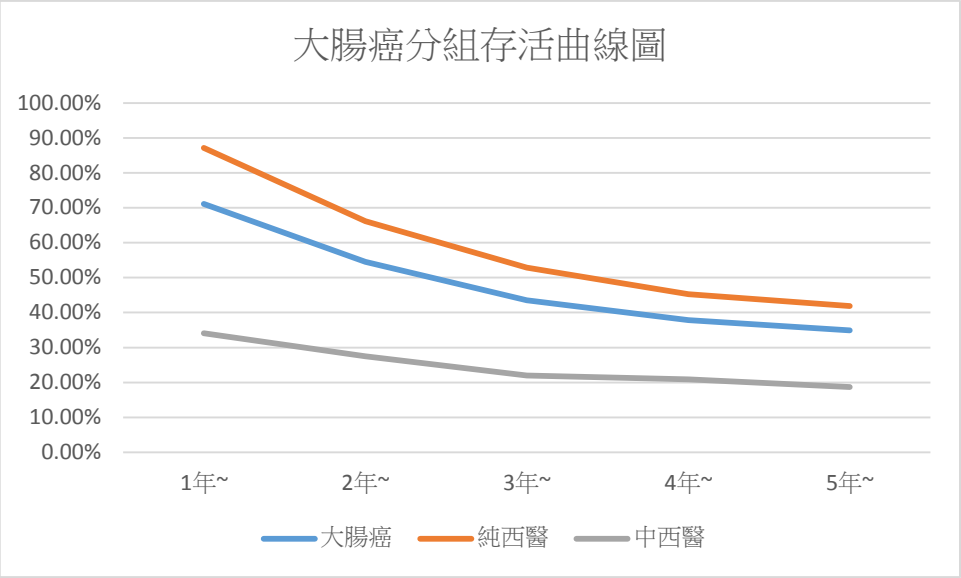

圖7大腸癌分組存活曲線圖

第二節、探討相關因子對大腸癌存活預測的影響

本研究利用就診方式的不同區分為所有病患組、純西醫治療組、中西醫合治組三個組別進行研究，並對每組別的五年存活、共病1、共病2、共病6、化療、地域、就診方式、年齡分組、性別、手術、放療、期別等12個變項對五年存活預測情形檢驗其個別變項的影響力，再將每組根據訓練集/測試集比例分成0.6/0.4、0.7/0.3、0.8/0.2三組進行比較，進而找出最具有準確性(accuracy)的結果分述如下：

一、所有病患組：在訓練集/測試集比例為0.6/0.4、0.7/0.3、0.8/0.2三組時其準確性(accuracy)如圖8所有病患組的訓練集/測試集0.6/0.4組、圖9所有病患組的訓練集/測試集0.7/0.3組、圖10所有病患組的訓練集/測試集0.8/0.2組，在比較之下0.6/0.4、0.8/0.2二組其準確性(accuracy)為最高，故選擇此二組作為最後決策樹之討論。在所有病患組之決策樹，如圖11所有病患組的訓練集/測試集0.6/0.4組之決策樹示意圖、圖12所有病患組的訓練集/測試集0.8/0.2組之決策樹示意圖，研究發現0.6/0.4組與0.8/0.2組手術變項均是最具有影響所有病患組五年存活之變項，其原因可能為開刀手術是最快速抑制病根的方法。

accuracy: 75.00%

	true 0	true 1	class precision
pred. 0	67	19	77.91%
pred. 1	11	23	67.65%
class recall	85.90%	54.76%	

圖8所有病患組的訓練集/測試集0.6/0.4組

accuracy: 72.22%

	true 0	true 1	class precision
pred. 0	50	16	75.76%
pred. 1	9	15	62.50%
class recall	84.75%	48.39%	

圖9所有病患組的訓練集/測試集0.7/0.3組

accuracy: 75.00%

	true 0	true 1	class precision
pred. 0	35	11	76.09%
pred. 1	4	10	71.43%
class recall	89.74%	47.62%	

圖10所有病患組的訓練集/測試集0.8/0.2組

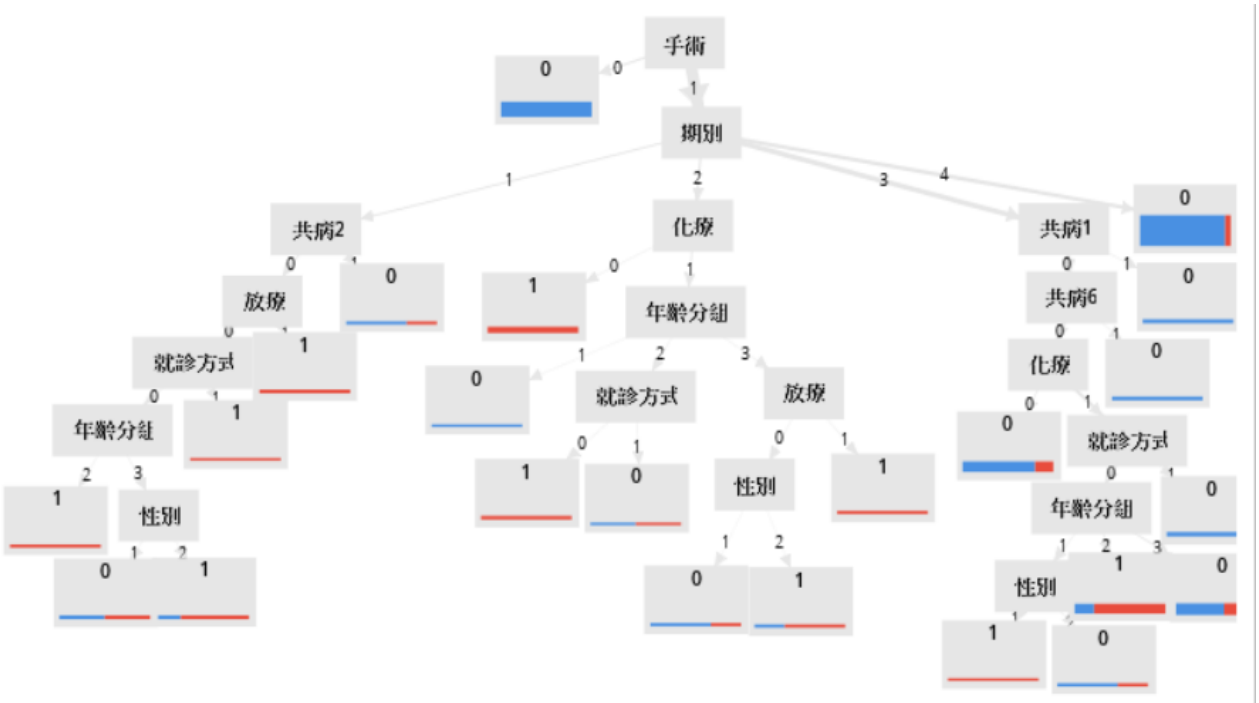

圖11所有病患組的訓練集/測試集0.6/0.4組之決策樹示意圖

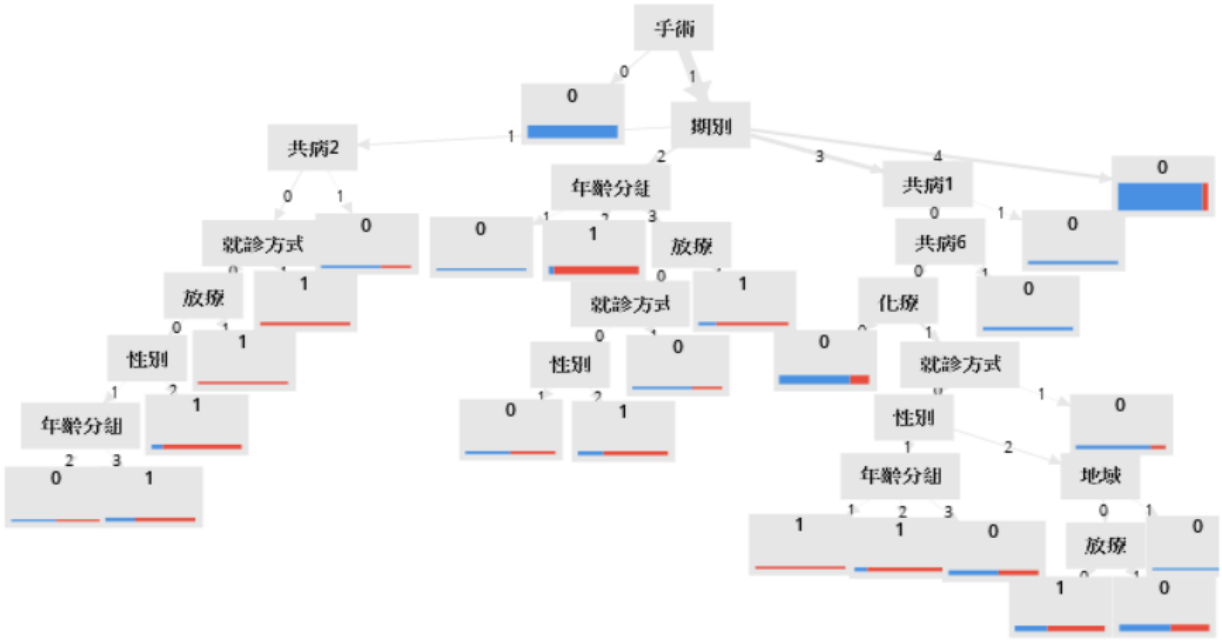

圖12所有病患組的訓練集/測試集0.8/0.2組之決策樹示意圖

二、純西醫治療組：在訓練集/測試集比例為0.6/0.4、0.7/0.3、0.8/0.2三組時其準確性(accuracy)如圖13純西醫治療組的訓練集/測試集0.6/0.4組、圖14純西醫治療組的訓練集/測試集0.7/0.3組、圖15純西醫治療組的訓練集/測試集0.8/0.2組，在比較之下0.8/0.2組其準確性(accuracy)為最高，故選擇此0.8/0.2組作為最後決策樹之討論。在純西醫治療組之決策樹，如圖16純西醫治療組的訓練集/測試集0.8/0.2組之決策樹示意圖，研究發現0.8/0.2組手術變項是最具有影響純西醫治療組五年存活之變項，其原因可能為開刀手術是最快速抑制病根的方法。

accuracy: 69.05%

	true 0	true 1	class precision
pred. 0	32	9	78.05%
pred. 1	17	26	60.47%
class recall	65.31%	74.29%	

圖13純西醫治療組的訓練集/測試集0.6/0.4組

accuracy: 65.08%

	true 0	true 1	class precision
pred. 0	24	9	72.73%
pred. 1	13	17	56.67%
class recall	64.86%	65.38%	

圖14純西醫治療組的訓練集/測試集0.7/0.3組

accuracy: 71.43%

	true 0	true 1	class precision
pred. 0	18	6	75.00%
pred. 1	6	12	66.67%
class recall	75.00%	66.67%	

圖15純西醫治療組的訓練集/測試集0.8/0.2組

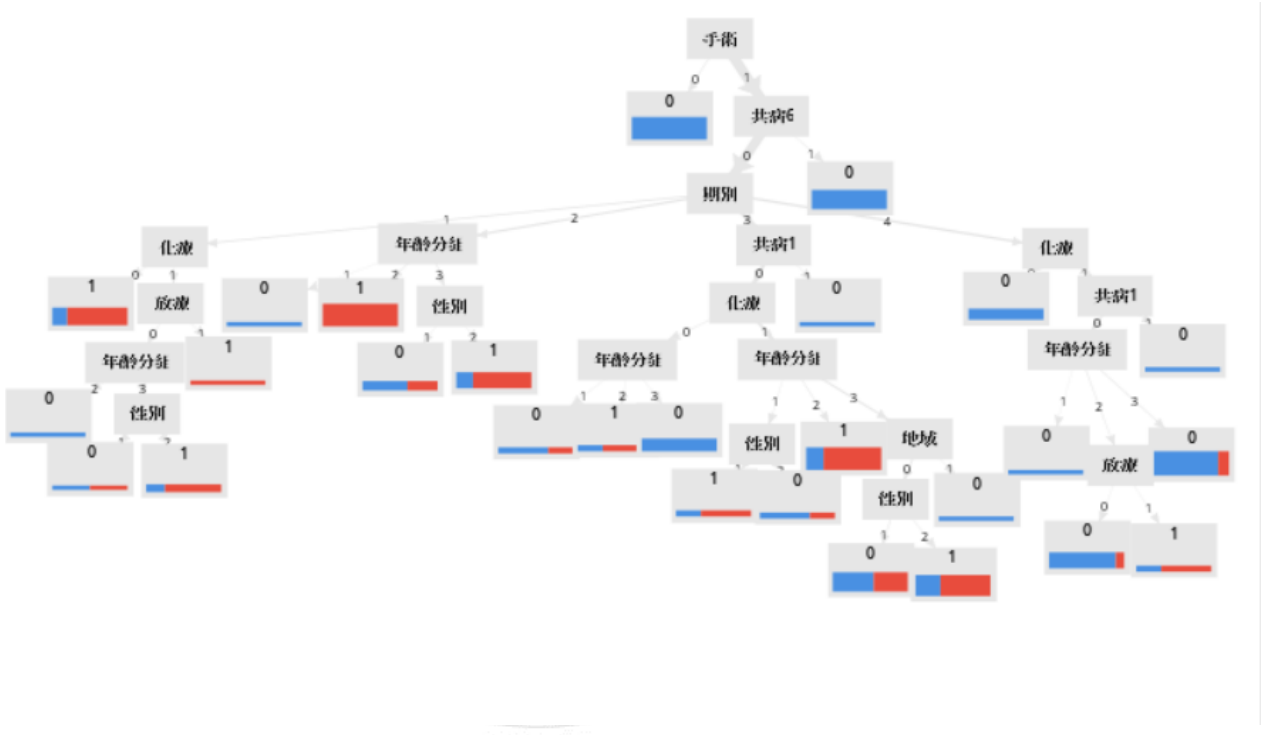

圖16純西醫治療組的訓練集/測試集0.8/0.2組之決策樹示意圖

三、中西醫合治組：在訓練集/測試集比例為0.6/0.4、0.7/0.3、0.8/0.2三組時其準確性(accuracy)如圖17中西醫合治組的訓練集/測試集0.6/0.4組、圖18中西醫合治組的訓練集/測試集0.7/0.3組、圖19中西醫合治組的訓練集/測試集0.8/0.2組，在比較之下0.8/0.2組其準確性(accuracy)為最高，故選擇0.8/0.2此組作為最後決策樹之討論。

在中西醫合治組之決策樹，如圖20中西醫合治組的訓練集/測試集0.8/0.2組之決策樹示意圖，研究發現0.8/0.2組期別變項是最具有影響中西醫合治組五年存活之變項，其原因可能為病患癌症期別若是一開始檢驗結果就是較晚期其存活率可能本就應該比較低，但仍可幫助一些病患延長存活時間。

accuracy: 78.38%

	true 0	true 1	class precision
pred. 0	28	6	82.35%
pred. 1	2	1	33.33%
class recall	93.33%	14.29%	

圖17中西醫合治組的訓練集/測試集0.6/0.4組

accuracy: 81.48%

	true 0	true 1	class precision
pred. 0	22	5	81.48%
pred. 1	0	0	0.00%
class recall	100.00%	0.00%	

圖18中西醫合治組的訓練集/測試集0.7/0.3組

accuracy: 83.33%

	true 0	true 1	class precision
pred. 0	15	3	83.33%
pred. 1	0	0	0.00%
class recall	100.00%	0.00%	

圖19中西醫合治組的訓練集/測試集0.8/0.2組

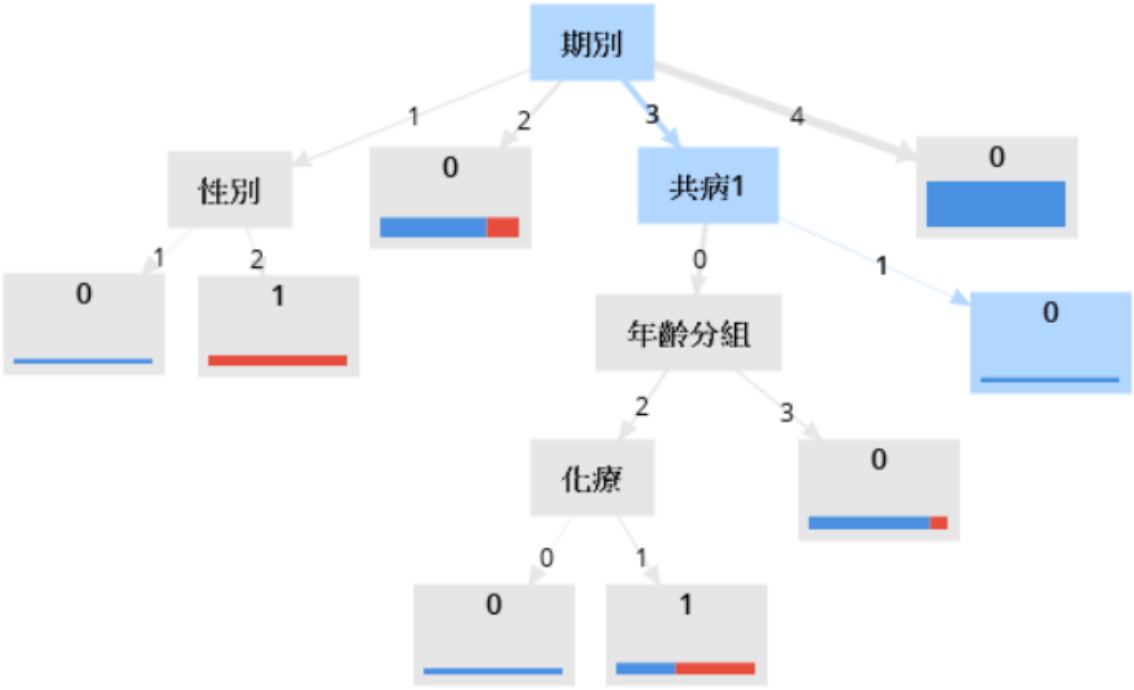

圖20中西醫合治組的訓練集/測試集0.8/0.2組之決策樹示意圖

第五章、結論

第一節、結論

長期以來台灣因生活型態改變,導致國人罹患大腸癌人數不斷增加，進而利用完善台灣健保資料庫找出其延長大腸癌病患存活時間的最主要變因，而本研究主要的個案來源，由西元 2007 年至 2014 年期間就診於大林慈濟醫院的大腸癌患者，探討治療追蹤中大腸癌病患的五年存活預測之最具影響力因子。

本研究透過 RapidMiner 分析軟體的(DecisionTree)分析五年存活預測之相關因子治療情形為何，依就診方式總共分別大腸癌病患、純西醫大腸癌病患、中西醫大腸癌病患三組而納入分析之因子分別為年齡、年齡分層、期別、地域、性別、手術、放療、化療，最後分析結果發現影響大腸癌病患組存活最重要的因子是手術因子而影響大腸癌純西醫組存活最重要因子為手術因子最後影響大腸癌中西醫組存活最重要為期別。

參 考 文 獻

一、中文部分

1. 行政院衛生福利部國民健康署，癌症登記報告 106 年，2017。
2. 呂浚都，應用資料探勘於淋巴癌病人存活預測之模式，中正大學資訊管理學系碩士班，碩士論文，民國 99 年。
3. 翁淑娟、葉明功、王舜睦、王俊翔，健保資料庫在藥學領域上之應用，藥學雜誌電子報 112 冊，2012。
4. 邱淑媞，衛生福利部國民健康署 2016 年癌症登記報告，2016。
5. 黃椒筠，資料探勘技術於結直腸癌患者分類模型之建構，國防醫學院公共衛生學研究所，碩士論文，2008。
6. 張語恬、朱基銘、簡戊鑑、周雨青、楊燦、盧瑜芬、白健佑、白璐、Thomas Wetter、孫建安、羅慶徽，運用三種資料探勘方法預測子宮頸癌存活情形之比較，台灣家庭醫學雜誌，17 卷 4 期，P222-238，2007。
7. 王煥昇，大腸癌會遺傳嗎，癌症新探 27 期，2008。
8. 陳建志，結腸直腸癌，和信治癌中心醫院 2011 年度報告，2011。
9. 吳雅琪，癌症試驗之存活資料分析，當代醫藥法規，Vol.14，2011。
10. 朱育增、吳肖琪，回顧與探討次級資料適用之共病測量方法，台灣衛誌，29 卷 1 期，P8-21，2010。
11. 吳肖琪、簡麗年、吳義勇，探討術前合併症指標與醫療利用及手術結果之關聯性－以全股(髖)關節置換健保申報資料為例，台灣衛誌，23 卷 2 期，P121-129，2004。
12. 徐瑜璟，西醫為主，中醫為輔，讓治療更有力，衛生福利部台南醫院中醫科，2014。
13. 陳俊良，癌症的中醫輔助治療，腫瘤護理雜誌，第十卷第二期，2010。

14. 姚遠賢、謝東呈、蘇淑芬，現今大腸癌之治療趨勢，弘光學報，65 期，2011。
15. 陳玫容，應用癌症資料庫之線上存活分析系統，成功大學工程科學系，碩士在職專班學位論文，2012。
16. 衛生福利部國民健康署，台灣癌症登記場長表摘錄手冊民國 100 年版，2013。
17. 曾玉華主任，請問醫師，什麼是放射線治療，嘉基院訊，第 146 期，2011。
18. 林星帆，醫護投稿實務一本通以 SCI 期刊為實作範例，2016。

二、西文部分

1. Deyo RA, Cherkin DC, Ciol MA., “Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases.”J Clin Epidemiol, 45:613-9, 1992.
2. Lee, Yuan-Wen, et al. “Adjunctive traditional Chinese medicine therapy improves survival in patients with advanced breast cancer: A population-based Study.” Cancer, 120.9: 1338-1344, 2014.
3. Tseng, Chin-Hsiao., “Use of insulin and mortality From breast cancer among Taiwanese women with diabetes.” Journal of diabetes research, 2015.
4. Hsu, I-Lin, et al., “Angiogenetic biomarkers in non-small cell lung cancer with malignant pleural effusion: correlations with patient survival and pleural effusion control.” Lung Cancer, 65.3: 371-376, 2009.
5. D. Kleinbaum and M. Klein., “Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health),” Springer, 2011.

三、網站部分

6. 行政院衛生福利部國民健康署，103 年度死因統計。取自於：http://www.mohw.gov.tw/cht/DOS/Statistic.aspx?f_list_no=312&fod_list_no=5488

7. 台灣癌症登記中心，取自於：http://tcr.cph.ntu.edu.tw/main.php?Page=A2。

8. 大腸癌照護網，取自於：http://www.crctw.org/Cognition_inner_P4.aspx。

9. 和信治癌中心醫院，取自於：http://www.kfsyscc.org/cancer/cancer-treatment/med/chemotherapy/。