科技部補助

大專學生研究計畫研究成果報告

計畫名稱：	運用 Hadoop 框架探討十大死因慢性病與人口地文間之關聯分析

執行計畫學生： 廖冠豪

學生計畫編號： MOST 104-2815-C-040-058-E

研 究 期 間 ： 104 年 07 月 01 日至 105 年 02 月 28 日止，計 8 個月

指 導 教 授 ： 曾明性

處理方式： 本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

執 行 單 位： 中山醫學大學醫學資訊學系

中華民國 105 年 03 月 04 日

一、 摘要

近年來因為人口老化，導致慢性病患人數增加，而在行政院衛生署國民健康局102年公布的資料中[1]也指出慢性疾病在十大死因中就佔了7項，因而耗費大量的健保支出，所以在未來人口老化的社會中，十大慢性疾病與人文、地文、經濟等變項間的關聯分析是個相當重要的研究課題。

由於健保資料庫資料龐大須耗費不少資料處理等待時間，因此本研究嘗試使用Hadoop技術框架及SQL程式來加速整體資料處理的效率。首先撰寫SQL程式及MapReduce程式進行資料擷取，繼之應用R軟體進行擷取資料集的關聯分析，最後再透過Ｒ軟體所提供的套件畫出十大慢性疾病間的網路關聯圖進行資料可視化分析。

本研究從100萬人之門診就醫紀錄最完整的健保資料庫中，探討十大慢性疾病間和性別、年齡、投保地、投保金額之關聯性。研究結果發現，國人罹患慢性疾病的機率與年齡、性別、投保金額、投保地息息相關，依年齡來看，慢性疾病普遍好發於18~59歲、60~79歲這兩個年齡層；依投保金額做探討，可發現慢性疾病患者主要分佈於極端值，最高或最低這兩個投保類別；依投保地做探討，則可發現慢性疾病患者集中於都市生活圈。此外，本研究也特別針對十大慢性病多重屬性關聯檔進行關聯規則分析，得到幾條令人感興趣的隱藏規則。

關鍵字：健保資料庫、慢性疾病、關聯分析、資料可視化

二、 研究計畫之背景及目的

隨著國人飲食和生活習慣的改變，各種文明病油然而生，且在高齡化社會中，又以慢性病最為重要，根據內政部統計處在 2005 年針對 50 歲以上「臺閩地區老人狀況調查」指出「50~64 歲國民患有慢性病或重大疾病者占 38.46%，65 歲以上老人患有慢性病或重大疾病者占 65.20%」，根據以上資料顯示，慢性病已成為國民個人與醫療機構應該嚴肅面對的問題。

人口老化導致慢性病患人數增加，同時也增加了治療高血壓、糖尿病及高血脂等藥品費用。97 年全民健康保險藥費支出達 1250 億，較 2006 年成長 6.9% ，其中心血管用藥達 263 億，占了 24%，成長率為 7.8%，亦為費用成長之主因。因人口老化迅速、壽命延長，衍生重大傷病及慢性病人數增加，導致藥費由 1998 年 772 億元，逐年快速成長到 2006 年 1141 億元，其中以門診慢性病（成長貢獻度 50.9%）高居藥費支出的第一名。「全民健康保險研究資料庫」擁有大量的就醫記錄，是從事挖掘醫學知識的最佳資料來源，但是過去的傳統分析方法常會耗費許多等待時間，因此本研究透過四大步驟：資料擷取、資料庫建置、資料分析、資料可視化。針對高血壓、高血脂症、糖尿病（高血糖）、心臟血管及腦血管疾病、慢

性肝炎、腎臟病、氣喘、骨質疏鬆症、痛風及關節炎，十大慢性疾病進行疾病關聯分析，探討在十大慢性病和年齡、性別、投保地、投保金額之間的關係，以過去未曾使用過的研究方法，找出醫學上未曾發現的新知識並佐證既有的醫學知識，提供個人疾病的預警、醫師臨床診斷和健保付費制度與合理支付制度之參考。

三、 文獻回顧與探討

(一)健保資料庫[2]

健保局從 2000 年起，委託財團法人國家衛生研究院建置全民健康保險研究資料庫，其中內容包含了完整的國人就醫的記錄。總共分為包括費用檔、醫令檔、基本資料檔。

由於資料龐大且類別複雜，所以國家衛生研究院針對不同研究需求，分為下列五種資料檔：

（1）基本資料檔：包含醫事機構病床主檔(BED)、醫事機構診療科別明細檔(DETA)、醫事機構基本資料檔(HOSB)、專科醫師證書主檔(DOC)、醫事人員基本資料檔(PER)、重大傷病證明明細檔(HV)，以及門診、住院費用總表等九類資料檔。

（2）系統抽樣檔：以每次門診或住院為抽樣對象的門診或住院資料檔

（3）抽樣歸人檔：以被保險人為抽樣對象的被保險人就醫記錄

（4）特定主題分檔:以不同疾病、病患身分及醫療層級區分之主題就醫記錄分檔

（5)教學用資料檔:含 85 年至 90 年共 6 年間每年 1000 人的就醫資料，免費提供教師教學使用

考慮到個人資料的保護，各個資料檔內之敏感資料（身分證字號、醫師代號及醫療機構代號欄位）均已經過加密的處理，研究者需利用國家衛生研究院提供之譯碼簿進行資料處理，譯碼簿提供健保資料庫中各種資料檔之欄位資（序號、英文欄位、中文欄位、資料型態、資料長度及資料貣始位置等）。

(二)Hadoop 技術框架[3]

Hadoop 是 Apache 軟件基金會旗下的一個開源分佈式計算平台。以 Hadoop 分佈式文件系統（HDFS，HadoopDistributedFilesystem）和 MapReduce（GoogleMapReduce 的開源實現）為核心的 Hadoop 為用戶提供了系統底層細節透明的分佈式基礎架構。HDFS 的高容錯性、高伸縮性等優點允許用戶將 Hadoop 在低廉的硬件上，形成分佈式系統；MapReduce 分佈式編程模型允許用戶在不了解分佈式系統底層細節的情況下開發並行應用程序。所以用戶可以利用 Hadoop 輕鬆地組織計算機資源，從而搭建自己的分佈式計算平台，並且可以充分利用集群的計算和存儲能力，完

成海量數據的處理。

HDFS 和 MapReduce 是 Hadoop 的兩大核心。而整個 Hadoop 的體系結構主要是通過 HDFS 來實現對分散式存儲的底層支持的，並且它會通過 MapReduce 來實現對分佈式並行任務處理的程序支持。

因本研究需要，特別說明 MapReduce 的體系結構，MapReduce 是一種並行編程模式，這種模式使得軟件開發者可以輕鬆地編寫出分佈式並行程序。在 Hadoop 的體系結構中，MapReduce 是一個簡單易用的軟件框架，基於它可以將任務分發到由上千台商用機器組成的集群上，並以一種高容錯的方式並行處理大量的數據集，實現 Hadoop 的並行任務處理功能。MapReduce 框架是由一個單獨運行在主節點上的 JobTracker 和運行在每個集群從節點上的 TaskTracker 共同組成的。主節點負責調度構成一個作業的所有任務，這些任務分佈在不同的從節點上。主節點監控它們的執行情況，並且重新執行之前失敗的任務；從節點僅負責由主節點指派的任務。當一個 Job 被提交時，JobTracker 接收到提交作業和配置信息之後，就會將配置信息等分發給從節點，同時調度任務並監控 TaskTracker 的執行。

MapReduce 編程模型的原理是：利用一個輸入的 key/value 對集合來產生一個輸出的 key/value 對集合。MapReduce 庫的用戶用兩個函數來表達這個計算：Map 和 Reduce。用戶自定義的 map 函數接收一個輸入的 key/value 對，然後產生一個中間 key/value 對的集合。MapReduce 把所有具有相同 key 值的 value 集合在一貨，然後傳遞給 reduce 函數。用戶自定義的 reduce 函數接收 key 和相關的 value 集合。reduce 函數合併這些 value 值，形成一個較小的 value 集合。一般來說，每次 reduce 函數調用只產生 0 或 1 個輸出的 value 值。通常我們通過一個迭代器把中間的 value 值提供給 reduce 函數，這樣就可以處理無法全部放入內存中的大量的 value 值集合了。

MapReduce 的過程簡而言之就是將大數據集分解為成百上千個小數據集，每個（或若干個）數據集分別由集群中的一個節點（一般就是一台普通的計算機）進行處理並生成中間結果，然後這些中間結果又由大量的節點合併，形成最終結果。

由以上資料可得知 MapReduce 計算模型非常適合在大量計算機組成的大規模集群上並行運行。每一個 map 任務和每一個 reduce 任務均可以同時運行於一個單獨的計算節點上，可想而知，其運算效率是非常高。

(三)關聯規則

資料探勘(Data Mining)可從大量的資料中找出未知、正確且有用的隱藏知識，是知識發現的重要技術。Chen et al. (1996) 認為資料探勘可從資料庫中萃取出重要、事先 未知、潛在有用的資訊；Berry 與 Linoff (1997) 認為資料探勘可針對大量的資料，應用自動或半自動的方式進行分析，以找出有意義的關係或法則。Fayyad(1996)和 Fayyad et al.(1996) 將知識發

現流程分為資料選擇、資料前置處理、資料轉換、資料探勘、解釋與評估等步驟，如圖 1 所示。Han 與 Kamber(2000)將資料探勘方法分為四大類，分別為關聯分析、分類與預測、群集分析、推估與偏差分析。

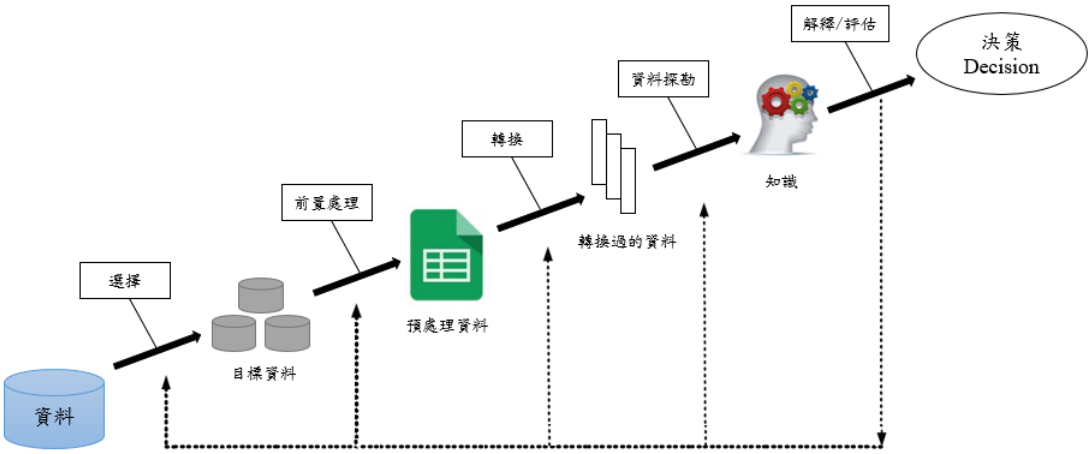


圖1　知識發現流程 (Fayyad et al., 1996)

其中關聯分析(Association Analysis)又稱為購物籃分析(Basket Analysis)，因擅於分析購物籃中購物品項間的關係而得名，可從眾多交易中(多個購物籃)分析出哪些商品同時被顧客購買的頻率較高，分析結果可協助業者訂定相關商品促銷策略，可應用於市場行銷及客戶社群分析等許多應用上(Khan et al. 2008)。

關聯分析由 Agrawal 等學者於 1993 年提出(Agrawal et al. 1993; Agrawal & Srikant 1994)，主要觀念建立在條件機率上，並且運用支持度(Support)與信賴度(Confidence)篩選 關聯規則，以關聯規則 X→Y 而言，其支持度 s = 包含項目 X 的交易數量，信賴度 c = (同時 包含項目 X、Y 的交易數量) / 包含 X 的交易數量。舉例而言，若在某賣場的交易中，60 筆 有購買香菸，40 筆有購買啤酒，同時購買香菸及啤酒的交易有 20 筆，則對於買香菸就會 買啤酒的關聯規則(以 X→Y 表示)來講，其支持度 s = 60，信賴度 c = 20 / 60 = 0.33。關聯 分析的目的是找出具備足夠支持度與信賴度的關聯規則，所以一個關聯規則要成立必須 同時滿足預先設定的最小支持度與最小信賴度。

(四) 慢性病關聯研究

吳瑞堯等人[4]利用 MSSQLServer 進行健保資料庫中僅僅 4 萬人 3 年門診紀錄分析，經資料淨化與篩選後，進行統計分析，以便了解上述慢性疾病在不同生活圈、年齡層及性別上之盛行率，並利用 MSSQLServerBI 之關聯規則探勘找出不同生活圈、性別、年齡之慢性疾病間的關聯性差異。獲得下列研究結果:

(1)高血壓性疾病，男、女性盛行率均居 6 種慢性疾病之冠，65 歲以上族群盛行率更高達 37.61%。而且不管哪種生活圈、性別、年齡，腎臟病變、腦血管疾病、心臟疾病、糖尿病、肝臟病變與高血壓性疾病間都有關聯，突顯高血壓性疾病照護之重要性。

(2)這6種慢性疾病在一般生活圈男女性之盛行率均高於都會生活圈，可能是一般生活圈民眾對慢性疾病防治較不積極，或醫療資源較為不足，無法早期醫療、早期控制，顯示一般生活圈民眾健康意識宣導及醫療資源均應加強。

(3)肝臟病變在 40~64 歲年齡層之盛行率高達 10.88%，高於 65 歲以上的 8.03%，提醒民眾肝臟保護應從年輕做貣。

(4)男性之慢性疾病盛行率普遍較女性高，但一般生活圈的女性之高血壓性疾病及心臟疾病盛行率卻高於男性，一般生活圈的女性應特別注意這兩項疾病。離島生活圈女性在高血壓性疾病、心臟疾病、糖尿病之盛行率也高於男性，值得注意。

(5)一般生活圈 65 歲以上高血壓性疾病及心臟疾病盛行率明顯高於都會生活圈。離島生活圈 65 歲以上高血壓性疾病及腦血管疾病盛行率更明顯高於其他兩個生活圈。顯示政府應加強一般生活圈及離島生活圈老人此兩項疾病之照護。

(6)都會生活圈中信賴度最高的關聯規則「腎臟病變、腦血管疾病→高血壓性疾病」在一般生活圈中不存在。顯示不同生活圈間有疾病關聯的差異。

(7)在男性中信賴度達到 0.78 的關聯規則「腎臟病變、腦血管疾病→高血壓性疾病」並未存在女性中，可能是女性腎臟病變及腦血管疾病盛行率都遠較男性低。另外，女性心臟疾病與高血壓性疾病、糖尿病間的關係均較男性明顯。顯示性別間有疾病關聯的差異。

(8)年齡層愈高，慢性疾病間的關聯也愈明顯，顯示慢性疾病相互引發之關係密切，因此慢性病早期防治對醫療資源的有效運用相當重要。

(9)在 40~64 歲年齡層中患有腎臟病變及肝臟病變者很可能同時會有高血壓性疾病，這樣的知識尚未見諸於文獻中，值得醫界注意。

(10)關聯分析顯示高血壓性疾病、腎臟病變及腦血管疾病間有高度關聯；高血壓性疾病、腦血管疾病及心臟疾病間亦有高度關聯；而女性的高血壓性疾病、腦血管疾病及糖尿病有高度關聯。提醒民眾及醫生應注意其關聯性，一發現有其中一種疾病，應該確實檢查是否有相關疾病，或積極進行相關疾病防治。

本研究採用 100 萬人於 2011 年的 20 個門診紀錄檔進行分析。

四、　研究方法及步驟

研究架構:本研究分為資料擷取、資料庫建置、資料分析以及資料可視化四部分進行，如圖2所示。

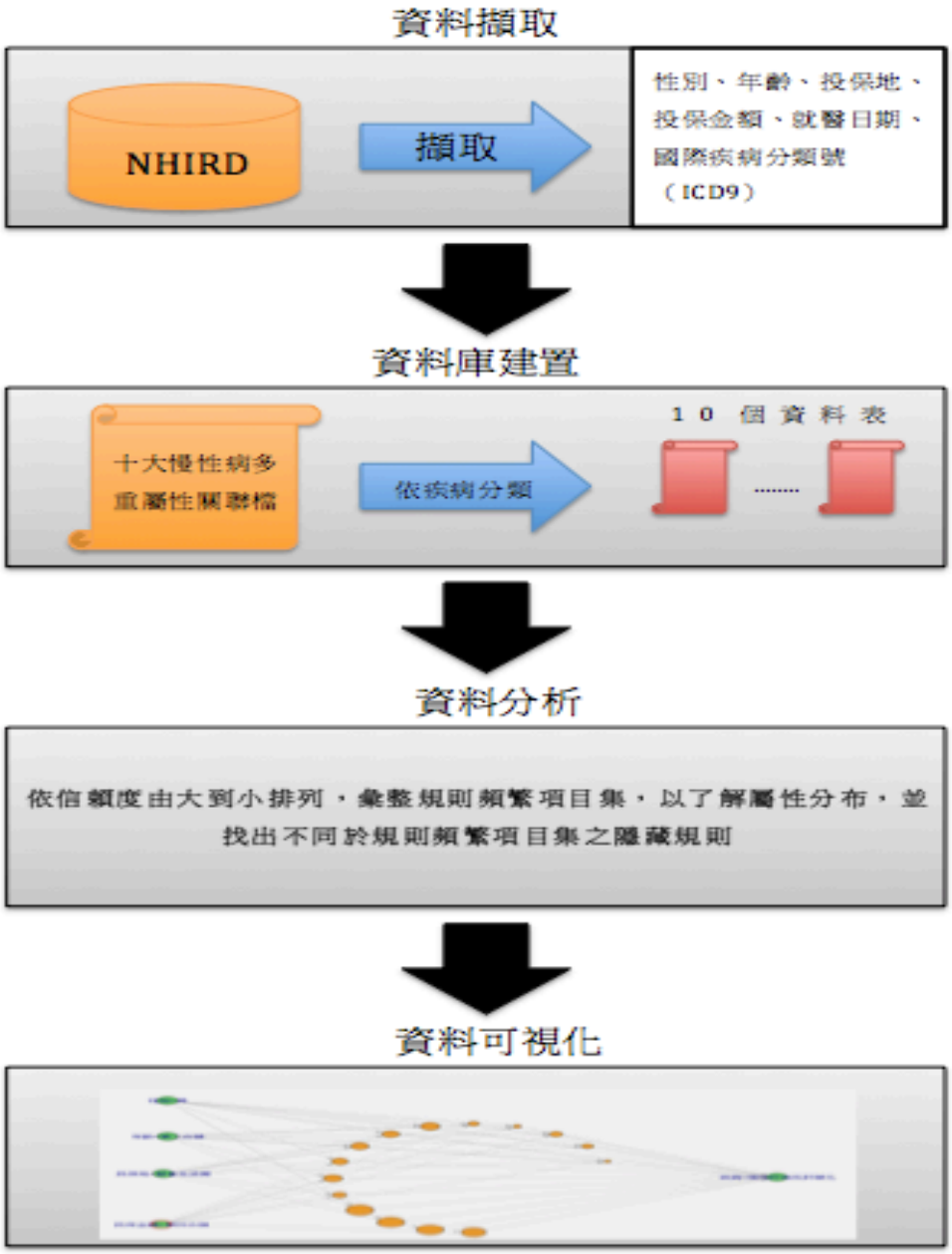


圖2、研究架構

(1)資料擷取

本研究首先為了瞭解使用Hadoop技術框架進行整體資料處理的加速效率，並探討節點數量對運算效能的影響，初步使用5萬人門診資料及四個計算節點進行效能測試。結果由圖3可以發現，當節點數越多時，運算速度亦隨之增快。

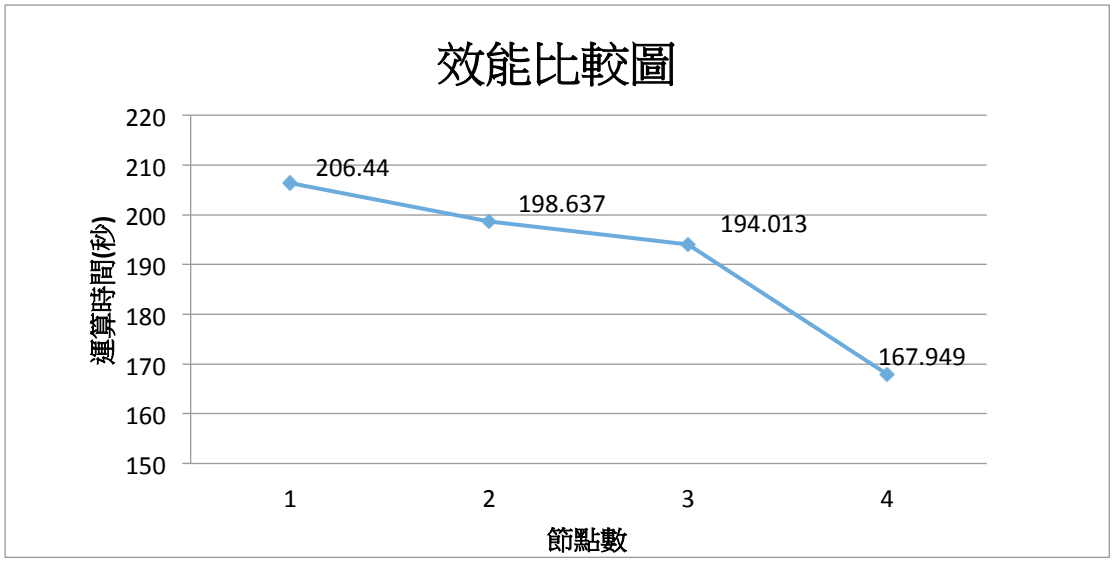


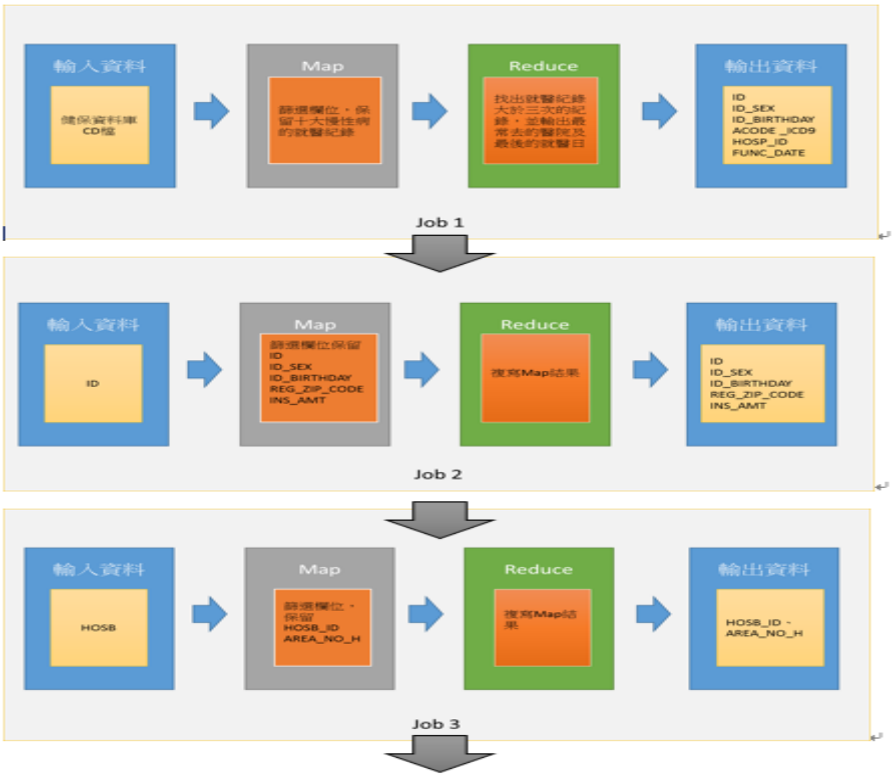
圖3、效能比較圖

為取得研究用資料，本研究依據國家衛生研究院提供之譯碼簿逐列擷取承保資料檔中的身分證字號、出生日期、性別及投保單位所在地區代碼等欄位。本研究利用100萬人於2011年的20個門診紀錄檔，依據性別、年齡、投保地、就醫日期及國際疾病分類號(ICD-9)分割及擷取，其程式流程如圖4所示。由於本研究僅對慢性疾病進行研究，非屬慢性疾病之記錄不納入資料庫。因門診處方及治療明細檔中所載疾病為國際疾病分類代碼ICD-9-CM，因此本研究依據表1之疾病名稱與國際疾病分類代碼(ICD-9-CM 2001)對照表將十大死因慢性疾病篩選出來。

表 1 疾病名稱與國際疾病分類代碼(ICD-9-CM 2001)

疾病名稱	國際疾病分類代碼	疾病名稱	國際疾病分類代碼
糖尿病	250	氣喘	493
骨質疏鬆症	73300,73301,73302, 73303,73309	慢性肝病及肝硬化	571
心臟性疾病	390-392,393-398, 410-414,420-429	腎炎腎徵候群及腎性病變	580-589
高血壓性疾病	401-405	痛風性關節病	2740

腦血管疾病	430-438	高血脂症	2722,2724



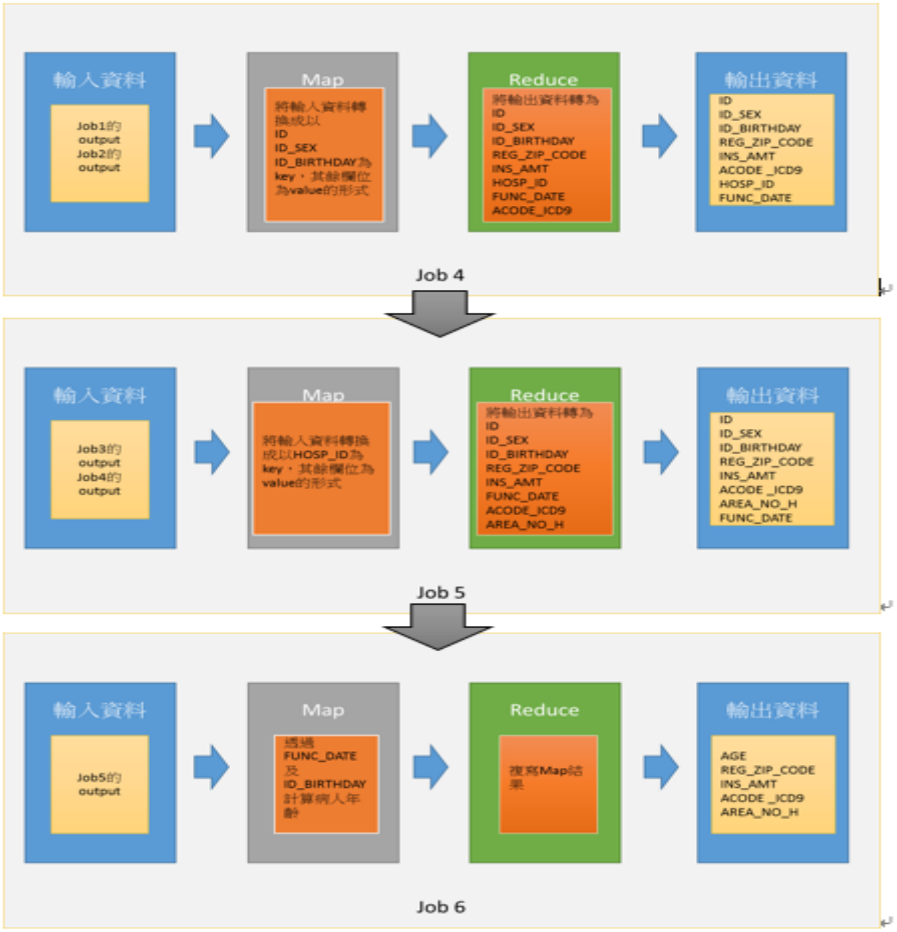


圖 4、MapReduce 過程

(2)資料庫建置

在進行資料分析探勘之前，必須先建立分析用的資料庫，本研究所建立的分析用資料庫為十大慢性病多重屬性關聯檔，其依疾病區分為 10 個資料表。十大慢性病多重屬性關聯檔包含疾病、性別、年齡、投保地、投保金額共計 5 個欄位，由於此資料表僅針對十大慢性病，故其他疾病之相關統計數據均不列入此資料表中。其中，疾病的欄位為高血壓性疾病、高血脂症、糖尿病（高血糖）、心臟性疾病、腦血管疾病、慢性肝病及肝炎、腎炎腎病症候群及腎病、氣喘、骨質疏鬆症、痛風及關節炎共 10 種慢性疾病；性別則分為男性與女性；年齡則分成 0~17 歲（第一類別）、18~59 歲（第二類別）、60~79 歲（第三類別）、80 歲以上（第四類別）四種類別；投保地如表 2 分為都會生活圈、一般生活圈、離島生活圈三種；投保金額利用四分位數法分成第一至第四類別。

表 2 投保地分類

生活圈	縣市

都會生活圈	台北，桃園，新竹，台中，台南，高雄
一般生活圈	基隆，宜蘭，苗栗，南投，彰化，雲林，嘉義，屏東，花蓮，台東
離島生活圈	澎湖，金門，連江

(3)資料分析

為了了解十大慢性病和年齡、性別、投保地、投保金額之間的關係，本研究利用 R 軟體[5]所提供的關聯規則探勘演算法對資料擷取後的目標資料集進行關聯分析。本研究運用 R 軟體所提供之關聯分析套件對十大慢性病多重屬性關聯檔所劃分的 10 張資料表進行關聯分析，擷取規則條件為:最小支持度設為 0.3、最小信賴度設為 0.9、且增益值大於等於 1.0 以上，並依信賴度由大到小排列，彙整規則頻繁項目集，了解屬性分布，找出不同於規則頻繁項目集之隱藏規則。

(4)資料可視化

為了清楚的表示資料關聯分析後的結果，本研究利用R軟體對關聯規則分析結果進行可視化之圖形呈現。

五、　結果

本研究採用之十大慢性病多重屬性關聯檔，依疾病分為 10 個資料表。利用 R 軟體進行關聯規則分析後，針對各資料表進行規則頻繁項目集整理得到下列的結果：

十大慢性病多重屬性關聯檔__心臟性疾病、氣喘、高血壓疾病、腎炎腎病症候群及腎病、糖尿病這 5 個資料表之規則頻繁項目集呈現以下屬性分佈，如表 3~表 7：依性別而言，男性或女性均容易罹患上述六種慢性疾病；依年齡來看，上述 5 種慢性疾病均好發於 18~59 歲、60~79 歲這兩個年齡層；依投保金額做探討，可發現患者主要均分佈於極端值，最高或最低這兩個投保類別；依投保地做探討，則可發現患者均集中於都市生活圈。

此外，本研究利用 R 軟體針對十大慢性病多重屬性關聯檔__骨質疏鬆症、痛風性關節炎、腦血管疾病、慢性肝病與肝硬化、高血脂症 5 個資料表進行關聯規則分析時，發現較特別之規則屬性分佈：以骨質疏鬆症為例，

此疾病之患者主要為 60~79 歲的女性，如表 8，根據台大醫院護理部護理長李嘉玲於 2013 年 2 月 63 期健康電子報[6]所提出的資訊，骨質疏鬆症好發於停經後 15~20 年的女性，因為在停經後體內雌激素急遽減少，破骨細胞活性增強而吸收骨小樑，使骨小樑變細、斷折、數目減少、不連續、減低骨強度。以台灣婦女平均停經年齡 51.2 歲[7]做推算，恰好落在本研究的年齡第三分類（60~79 歲）區間；以痛風性關節炎、慢性肝病與肝硬化為例，這兩種疾病的患者主要為 18~59 歲且投保金額屬於最高投保類別的男性，如表 9、表 10，壢新醫院過敏免疫風濕科葉松峰醫師曾在文中提到，痛風是典型的文明病，好發於經濟能力較佳的族群，這些人日常生活多半優渥，勞動少，美食文化也較風行，其中又以中年的男性為主，高峰為 50 歲。而慢性肝病的相關研究指出[8]，男性患有肝病的機率比女生多 2~3 倍，其中最大的影響因子為男性的賀爾蒙，在男生的肝細胞表面上，有許多男性荷爾蒙受體，容易與病毒結合，促進病毒的繁殖複製，讓肝細胞發炎壞死，因此，男生變成肝硬化的機率便增加許多；以腦血管疾病為例，此疾病之患者主要介於 60~79 歲之年齡層，如表 11，根據 101 年衛生福利部的死因統計分析顯示，腦血管疾病排名 65 歲以上人口的死因第三名；以高血脂症為例，此疾病的患者主要集中於投保金額屬於最高投保類別的族群，根據相關研究，多與經濟條件較佳者的飲食習慣有關，如表 12。

綜合上述，可知本研究探勘的十大死因慢性病關聯樣式，大部分與現有醫學知識可相互佐證。

表 3 心臟性疾病關聯規則

規則	支持度	信賴度	增益
{生活圈=都會生活圈} => {疾病=心臟性疾病}	0.716	1	1
{性別=男性} => {疾病=心臟性疾病}	0.506	1	1
{性別=女性} => {疾病=心臟性疾病}	0.494	1	1
{年齡=第三分類} => {疾病=心臟性疾病}	0.477	1	1
{投保金額=第四分類} => {疾病=心臟性疾病}	0.452	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=心臟性疾病}	0.370	1	1
{性別=女性,生活圈=都會生活圈} => {疾病=心臟性疾病}	0.346	1	1
{年齡=第三分類,生活圈=都會生活圈} => {疾病=心臟性疾病}	0.333	1	1
{年齡=第二分類} => {疾病=心臟性疾病}	0.327	1	1
{投保金額=第一分類} => {疾病=心臟性疾病}	0.320	1	1

表 4 氣喘關聯規則

規則	支持度	信賴度	增益
{生活圈=都會生活圈} => {疾病=氣喘}	0.708	1	1
{性別=女性} => {疾病=氣喘}	0.534	1	1
{性別=男性} => {疾病=氣喘}	0.466	1	1
{年齡=第二分類} => {疾病=氣喘}	0.457	1	1
{投保金額=第四分類} => {疾病=氣喘}	0.442	1	1
{性別=女性,生活圈=都會生活圈} => {疾病=氣喘}	0.379	1	1
{年齡=第三分類} => {疾病=氣喘}	0.350	1	1
{投保金額=第一分類} => {疾病=氣喘}	0.342	1	1
{年齡=第二分類,生活圈=都會生活圈} => {疾病=氣喘}	0.341	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=氣喘}	0.329	1	1

表 5 高血壓關聯規則

規則	支持度	信賴度	增益
{生活圈=都會生活圈} => {疾病=高血壓性疾病}	0.706	1	1
{性別=女性} => {疾病=高血壓性疾病}	0.505	1	1
{性別=男性} => {疾病=高血壓性疾病}	0.495	1	1
{投保金額=第四分類} => {疾病=高血壓性疾病}	0.489	1	1
{年齡=第三分類} => {疾病=高血壓性疾病}	0.478	1	1
{年齡=第二分類} => {疾病=高血壓性疾病}	0.389	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=高血壓性疾病}	0.356	1	1
{性別=女性,生活圈=都會生活圈} => {疾病=高血壓性疾病}	0.350	1	1
{年齡=第三分類,生活圈=都會生活圈} => {疾病=高血壓性疾病}	0.328	1	1
{投保金額=第一分類} => {疾病=高血壓性疾病}	0.309	1	1
{生活圈=都會生活圈,投保金額=第四分類} => {疾病=高血壓性疾病}	0.300	1	1

表 6 腎炎腎病症候群及腎病關聯規則

規則	支持度	信賴	增益

		度	
{生活圈=都會生活圈} => {疾病=腎炎腎徵候群及腎性病變}	0.712	1	1
{性別=男性} => {疾病=腎炎腎徵候群及腎性病變}	0.566	1	1
{年齡=第三分類} => {疾病=腎炎腎徵候群及腎性病變}	0.487	1	1
{性別=女性} => {疾病=腎炎腎徵候群及腎性病變}	0.434	1	1
{投保金額=第四分類} => {疾病=腎炎腎徵候群及腎性病變}	0.428	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=腎炎腎徵候群及腎性病變}	0.409	1	1
{投保金額=第一分類} => {疾病=腎炎腎徵候群及腎性病變}	0.340	1	1
{年齡=第三分類,生活圈=都會生活圈} => {疾病=腎炎腎徵候群及腎性病變}	0.336	1	1
{年齡=第二分類} => {疾病=腎炎腎徵候群及腎性病變}	0.331	1	1
{性別=女性,生活圈=都會生活圈} => {疾病=腎炎腎徵候群及腎性病變}	0.303	1	1

表 7 糖尿病關聯規則

規則	支持度	信賴度	增益
{生活圈=都會生活圈} => {疾病=糖尿病}	0.700	1	1
{性別=男性} => {疾病=糖尿病}	0.507	1	1
{年齡=第三分類} => {疾病=糖尿病}	0.499	1	1
{性別=女性} => {疾病=糖尿病}	0.493	1	1
{投保金額=第四分類} => {疾病=糖尿病}	0.479	1	1
{年齡=第二分類} => {疾病=糖尿病}	0.396	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=糖尿病}	0.362	1	1
{年齡=第三分類,生活圈=都會生活圈} => {疾病=糖尿病}	0.341	1	1
{性別=女性,生活圈=都會生活圈} => {疾病=糖尿病}	0.338	1	1
{投保金額=第一分類} => {疾病=糖尿病}	0.321	1	1

表 8 骨質疏鬆症關聯規則

規則	支持度	信賴度	增益
{性別=女性} => {疾病=骨質疏鬆症}	0.821	1	1
{生活圈=都會生活圈} => {疾病=骨質疏鬆症}	0.662	1	1
{年齡=第三分類} => {疾病=骨質疏鬆症}	0.564	1	1
{性別=女性,生活圈=都會生活圈} => {疾病=骨質疏鬆症}	0.545	1	1
{性別=女性,年齡=第三分類} => {疾病=骨質疏鬆症}	0.478	1	1
{投保金額=第四分類} => {疾病=骨質疏鬆症}	0.407	1	1
{投保金額=第一分類} => {疾病=骨質疏鬆症}	0.389	1	1
{年齡=第三分類,生活圈=都會生活圈} => {疾病=骨質疏鬆症}	0.367	1	1
{性別=女性,投保金額=第一分類} => {疾病=骨質疏鬆症}	0.345	1	1
{性別=女性,投保金額=第四分類} => {疾病=骨質疏鬆症}	0.332	1	1
{生活圈=一般生活圈} => {疾病=骨質疏鬆症}	0.328	1	1
{性別=女性,年齡=第三分類,生活圈=都會生活圈} => {疾病=骨質疏鬆症}	0.313	1	1
{生活圈=都會生活圈,投保金額=第一分類} => {疾病=骨質疏鬆症}	0.305	1	1

表 9 痛風性關節炎關聯規則

規則	支持度	信賴度	增益
{性別=男性} => {疾病=痛風性關節炎}	0.833	1	1
{生活圈=都會生活圈} => {疾病=痛風性關節炎}	0.694	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=痛風性關節炎}	0.589	1	1
{投保金額=第四分類} => {疾病=痛風性關節炎}	0.562	1	1
{年齡=第二分類} => {疾病=痛風性關節炎}	0.554	1	1
{性別=男性,年齡=第二分類} => {疾病=痛風性關節炎}	0.504	1	1
{性別=男性,投保金額=第四分類} => {疾病=痛風性關節炎}	0.483	1	1

{年齡=第二分類,生活圈=都會生活圈} => {疾病=痛風性關節炎}	0.400	1	1
{年齡=第二分類,投保金額=第四分類} => {疾病=痛風性關節炎}	0.382	1	1
{性別=男性,年齡=第二分類,生活圈=都會生活圈} => {疾病=痛風性關節炎}	0.367	1	1
{生活圈=都會生活圈,投保金額=第四分類} => {疾病=痛風性關節炎}	0.362	1	1
{性別=男性,年齡=第二分類,投保金額=第四分類} => {疾病=痛風性關節炎}	0.351	1	1
{年齡=第三分類} => {疾病=痛風性關節炎}	0.347	1	1
{性別=男性,生活圈=都會生活圈,投保金額=第四分類} => {疾病=痛風性關節炎}	0.322	1	1

表 10 慢性肝病與肝硬化關聯規則

規則	支持度	信賴度	增益
{性別=男性} => {疾病=慢性肝病及肝硬化}	0.734	1	1
{年齡=第二分類} => {疾病=慢性肝病及肝硬化}	0.730	1	1
{生活圈=都會生活圈} => {疾病=慢性肝病及肝硬化}	0.723	1	1
{投保金額=第四分類} => {疾病=慢性肝病及肝硬化}	0.674	1	1
{性別=男性,年齡=第二分類} => {疾病=慢性肝病及肝硬化}	0.577	1	1
{年齡=第二分類,投保金額=第四分類} => {疾病=慢性肝病及肝硬化}	0.551	1	1
{年齡=第二分類,生活圈=都會生活圈} => {疾病=慢性肝病及肝硬化}	0.539	1	1
{性別=男性,生活圈=都會生活圈} => {疾病=慢性肝病及肝硬化}	0.524	1	1
{性別=男性,投保金額=第四分類} => {疾病=慢性肝病及肝硬化}	0.509	1	1
{生活圈=都會生活圈,投保金額=第四分類} => {疾病=慢性肝病及肝硬化}	0.461	1	1
{性別=男性,年齡=第二分類,投保金額=第四分類} => {疾病=慢性肝病及肝硬化}	0.434	1	1
{性別=男性,年齡=第二分類,生活圈=都會生活	0.419	1	1

圈}=>{疾病=慢性肝病及肝硬化}			
{年齡=第二分類,生活圈=都會生活圈,投保金額=第四分類}=>{疾病=慢性肝病及肝硬化}	0.397	1	1
{性別=男性,生活圈=都會生活圈,投保金額=第四分類}=>{疾病=慢性肝病及肝硬化}	0.345	1	1
{性別=男性,年齡=第二分類,生活圈=都會生活圈,投保金額=第四分類}=>{疾病=慢性肝病及肝硬化}	0.307	1	1

表 11 腦血管疾病關聯規則

規則	支持度	信賴度	增益
{生活圈=都會生活圈}=>{疾病=腦血管疾病}	0.701	1	1
{性別=男性}=>{疾病=腦血管疾病}	0.566	1	1
{年齡=第三分類}=>{疾病=腦血管疾病}	0.529	1	1
{性別=女性}=>{疾病=腦血管疾病}	0.434	1	1
{性別=男性,生活圈=都會生活圈}=>{疾病=腦血管疾病}	0.403	1	1
{投保金額=第一分類}=>{疾病=腦血管疾病}	0.383	1	1
{投保金額=第四分類}=>{疾病=腦血管疾病}	0.370	1	1
{年齡=第三分類,生活圈=都會生活圈}=>{疾病=腦血管疾病}	0.360	1	1
{生活圈=都會生活圈,投保金額=第一分類}=>{疾病=腦血管疾病}	0.301	1	1

表 12 高血脂症關聯規則

規則	支持度	信賴度	增益
{生活圈=都會生活圈}=>{疾病=高血脂症}	0.744	1	1
{投保金額=第四分類}=>{疾病=高血脂症}	0.510	1	1
{性別=女性}=>{疾病=高血脂症}	0.509	1	1
{性別=男性}=>{疾病=高血脂症}	0.491	1	1
{年齡=第二分類}=>{疾病=高血脂症}	0.481	1	1
{年齡=第三分類}=>{疾病=高血脂症}	0.450	1	1
{性別=女性,生活圈=都會生活圈}=>{疾病=高血脂症}	0.376	1	1
{性別=男性,生活圈=都會生活圈}=>{疾病=高血脂症}	0.368	1	1

{年齡=第二分類,生活圈=都會生活圈}=>{疾病=高血脂症}	0.365	1	1
{生活圈=都會生活圈,投保金額=第四分類}=>{疾病=高血脂症}	0.344	1	1

為了更清楚展示各屬性與疾病死亡率之關聯，本研究利用 R 軟體，將 10 大慢性疾病之關聯規則，進行資料可視化，如圖 5～圖 14 所示。

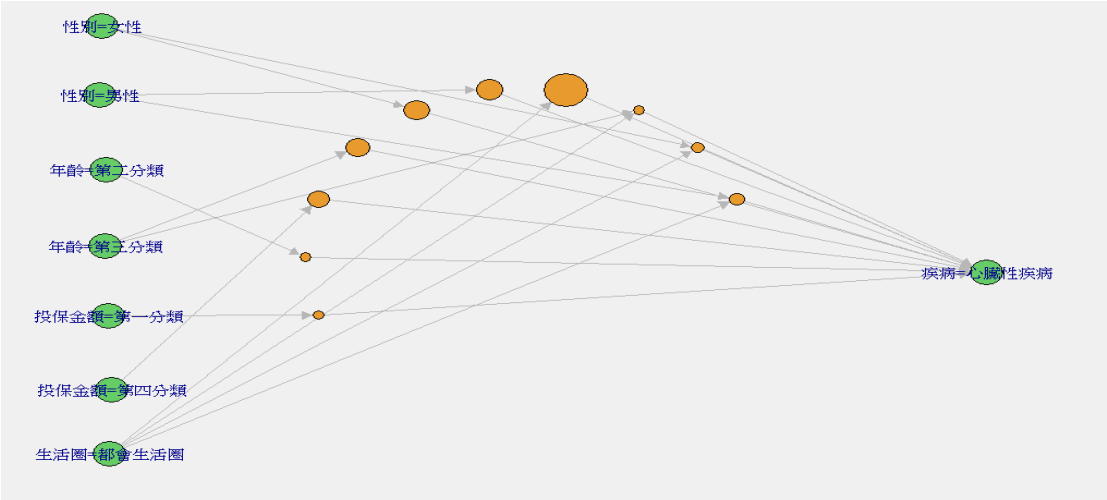


圖 5、心臟性疾病關聯規則圖

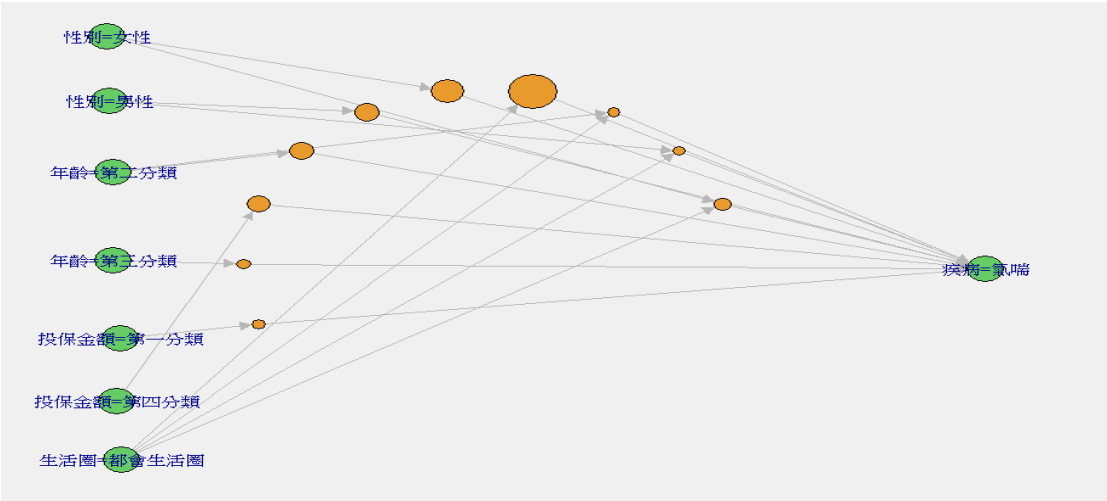


圖 6、氣喘關聯規則圖

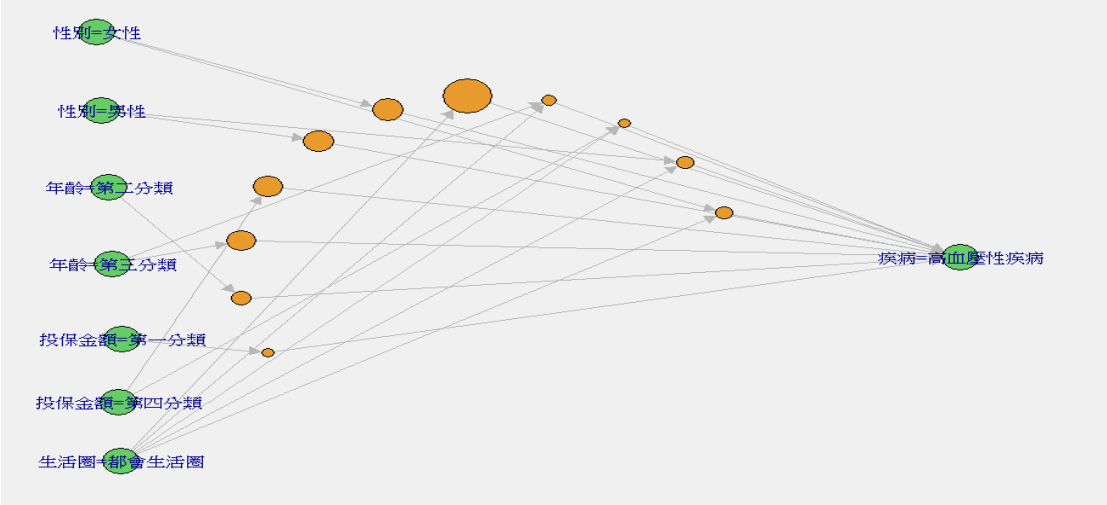


圖 7、高血壓性疾病關聯規則圖

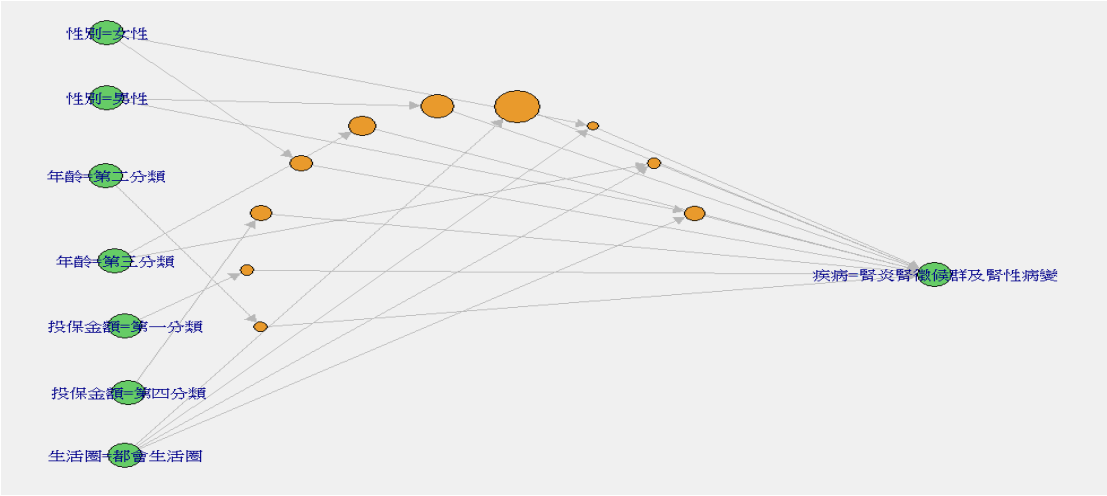


圖 8、腎炎腎徵候群及腎性病變關聯規則圖

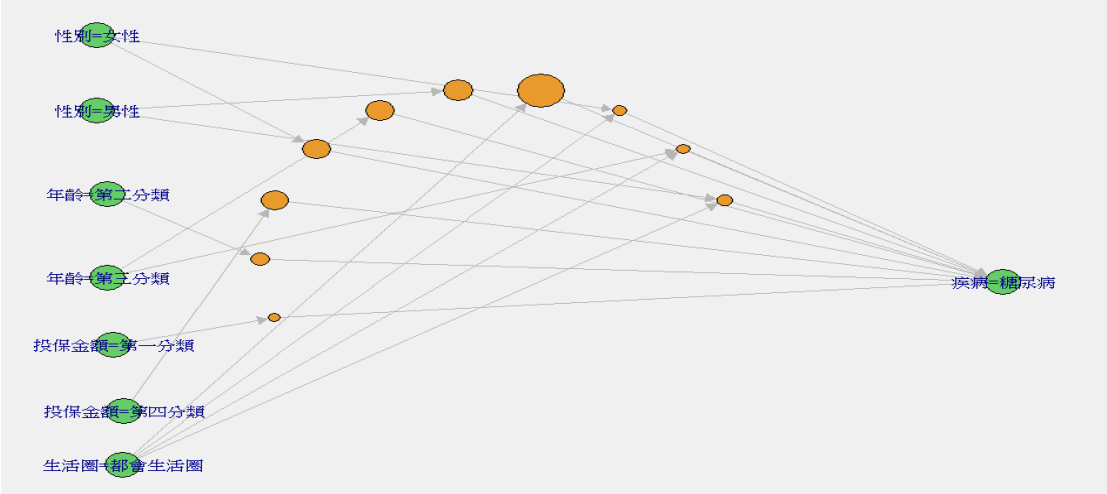


圖 9、糖尿病關聯規則圖

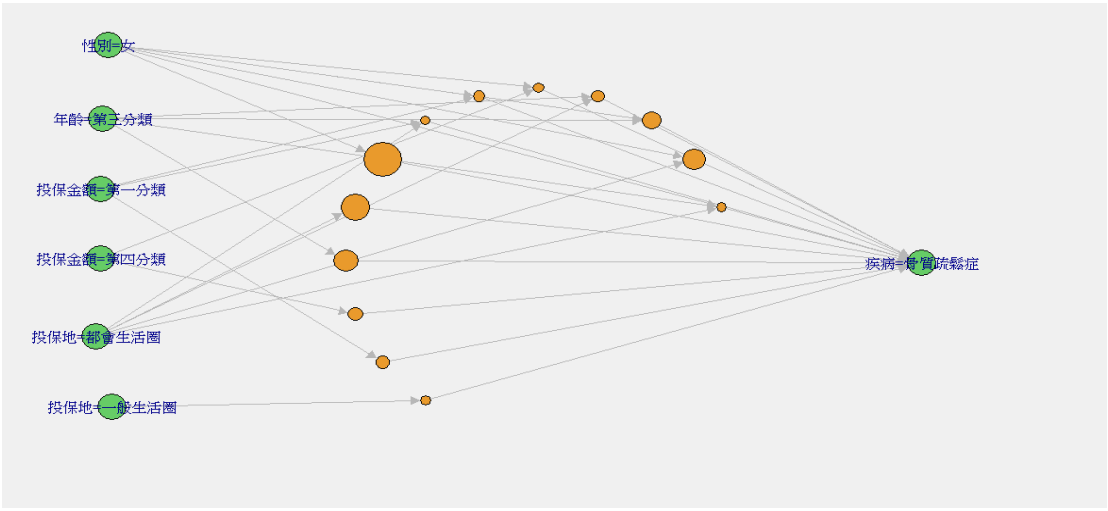


圖 10、骨質疏鬆症關聯規則圖

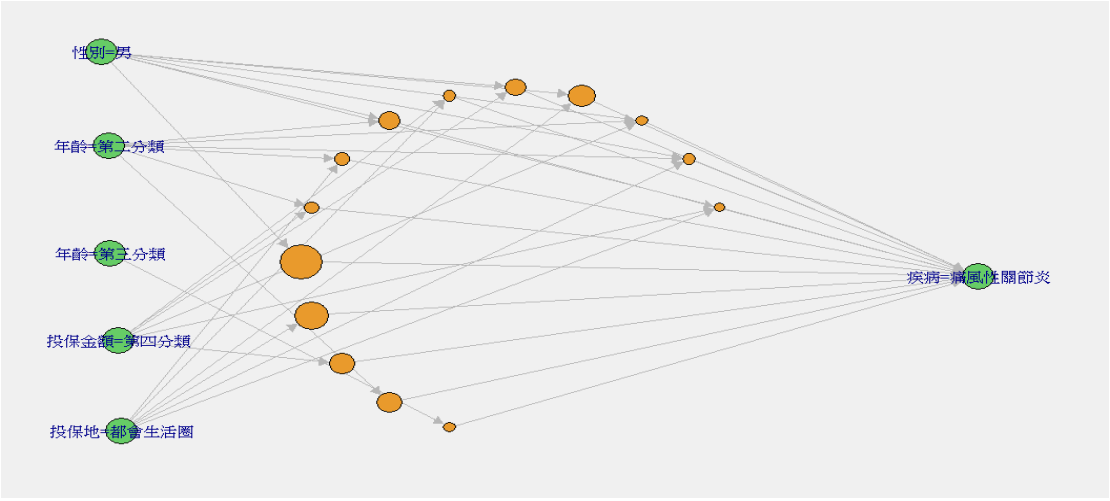


圖 11、痛風性關節炎關聯規則圖

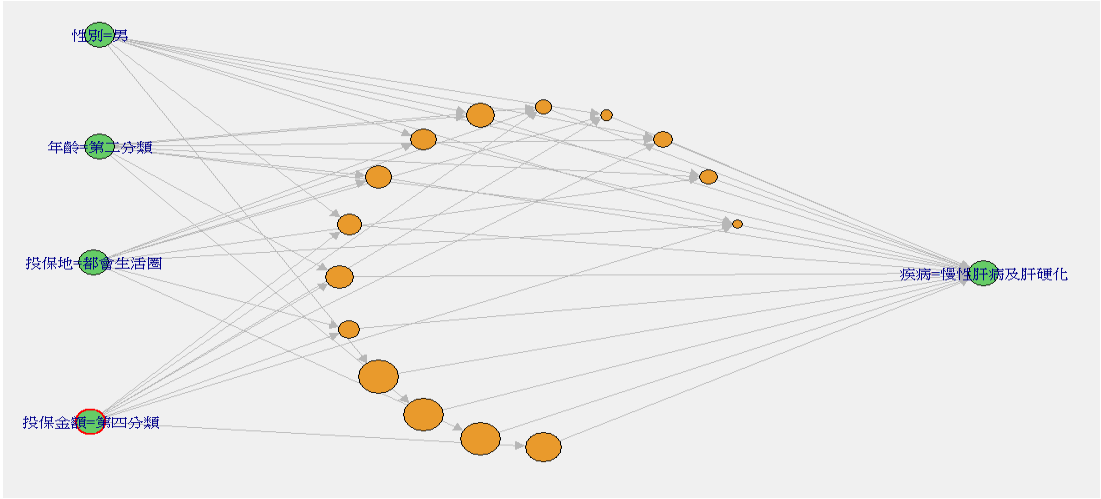


圖 12、慢性肝病及肝硬化關聯規則圖

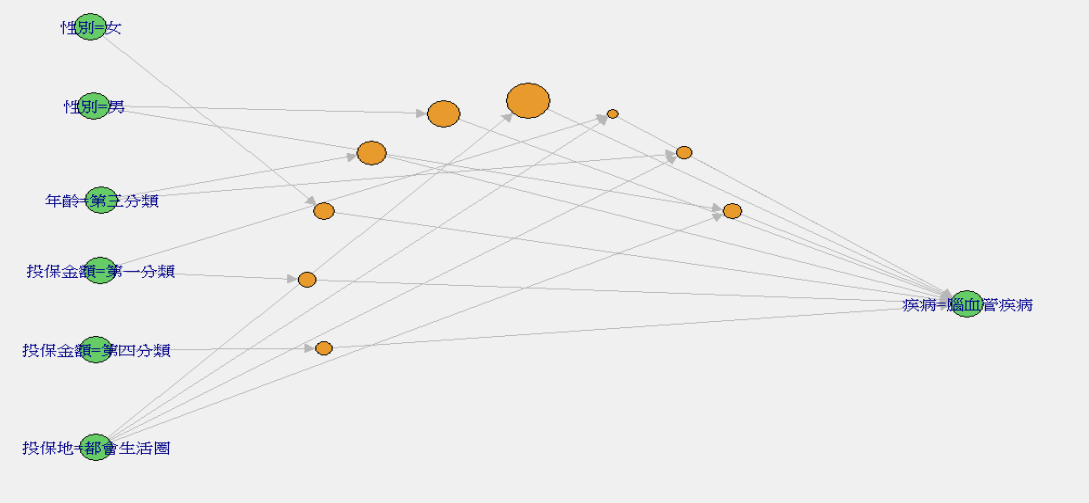


圖 13、腦血管疾病關聯規則圖

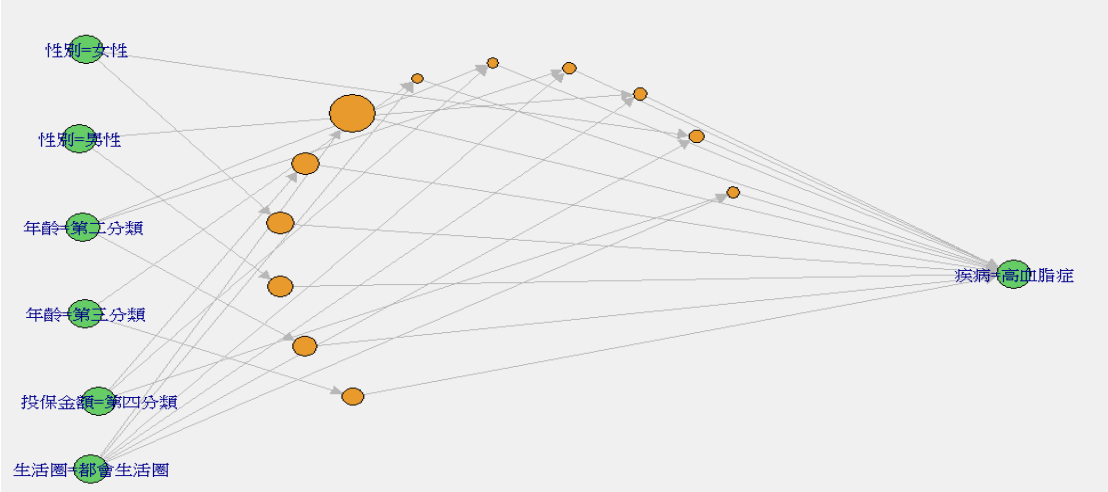


圖 14、高血脂症關聯規則圖

六、　結論

本研究運用 Hadoop 技術框架及 SQL 程式從 100 萬人的健保資料庫中進行資料擷取，並利用 R 軟體對十大慢性病多重屬性關聯檔所劃分的 10 張資料表進行關聯分析及資料可視化後，進行規則頻繁項目集整理，得到下列的結果：

(一) 十大慢性病多重屬性關聯檔__心臟性疾病、氣喘、高血壓疾病、腎炎腎病症候群及腎病、糖尿病，這 5 個資料表之規則頻繁項目集呈現以下樣式分佈：

(1) 依性別而言，男性或女性均容易罹患上述 5 種慢性疾病。

(2) 依年齡來看，上述 5 種慢性疾病均好發於 18~59 歲、60~79 歲這兩個年齡層。

(3) 依投保金額做探討，可發現患者主要均分佈於極端值，最高或最

低這兩個投保類別。

(4) 依投保地做探討，則可發現患者均集中於都市生活圈。

(二)十大慢性病多重屬性關聯檔__骨質疏鬆症、痛風性關節炎、腦血管疾病、慢性肝病與肝硬化、高血脂症，這 5 個資料表進行分析時，發現較特別之規則樣式分佈：

(1) 以骨質疏鬆症為例，此疾病之患者主要為 60~79 歲的女性。

(2) 以痛風性關節炎、慢性肝病與肝硬化為例，此兩類疾病之患者主要為 18~59 歲且投保金額屬於最高投保類別的男性。

(3) 以腦血管疾病為例，此疾病之患者主要介於 60~79 歲之年齡層。

(4) 以高血脂症為例，此疾病之患者主要集中在投保金額屬於最高類別之族群。

本研究結果顯示國人罹患慢性疾病的機率與年齡、性別、投保金額、投保地息息相關，且某些疾病好發於特定的年齡層、性別或是投保金額最高的族群。停經後的女性，可透過諮詢醫生索取建議，降低罹患骨質疏鬆症的危險；男性及生活條件較優渥的族群則應盡量改變生活和飲食習慣，舉凡酗酒、熬夜等應盡量避免。上述研究結果可供政府單位在訂定公衛政策或推動疾病防治工作時作為依據及參考，落實真正的醫療資源平等。

七、 參考文獻

[1] 行政院衛生署國民健康局，102 年國人死因統計結果（http://health99.hpa.gov.tw/Article/ArticleDetail.aspx?TopIcNo=846&DS=1-life）

[2] 全民健康保險研究資料庫 (http://nhird.nhri.org.tw/date_01.htm)

[3] 陸嘉恆，《Hadoop 實戰》

[4] 吳瑞堯、周駿賢，『運用資料探勘技術於六大死因慢性疾病之研究』，資訊管理學報，第十八卷，第一期

[5] RDataMining (http://www.rdatamining.com)

[6] 李嘉玲，健康電子報，**2013**年**2**月**63**期（http://epaper.ntuh.gov.tw/health/201302/health_2.html）

[7] 姚智偉，台北縣政府衛生局電子報（http://www.health.ntpc.gov.tw/web66/_file/1459/upload/ehealth/9905/pages/index-01-1.html）

[8] 許金川，台北市立大健保藥師藥局部落格（http://chuckdon.pixnet.net/blog/post/30086580-%E7%94%B7%E7%94%9F%E5%A4%9A%E8%82%9D%E7%97%85%E8%8D%B7%E7%88%BE%E8%92%99%E5%AE%B3%E7%9A%84）