# 血液透析關鍵因子分析及透析病患分群技術

呂慈純
朝陽科技大學副教授
tclu@cyut.edu.tw

曾俊雅
朝陽科技大學研究生
s9814604@cyut.edu.tw

## 摘要

腎臟位於人體內的腹腔後壁的脊椎兩側，是身體新陳代謝、排除廢物、毒素、調節血壓、和維持體內液體平衡的重要器官。一旦腎臟功能異常，無法正常運作，身體便會產生毒素，危害器官，甚至導致死亡。為了延長或挽救腎臟病患者的生命，常使用腎臟替代療法：腎臟移植(Kidney Transplantation)、血液透析(Hemodialysis, HD)及腹膜透析(Peritoneal Dialysis, PD)進行醫治。血液透析就是常稱的「洗腎」，利用人工腎(透析)設備將尿毒素、水份排出體外，減輕尿毒症狀。至於何時該進行血液透析，一般而言，當尿素氮大於 90，肌酐酸大於 9 時，肌酐酸清除率小於 0.17 ml/s 或肌酐酸大於 707.2 ml/L，醫生就會建議可以開始進行血液透析治療。但是，腎臟組織在一般正常人約僅發揮 30%的功能，就能維持日常的排毒功能，因此腎臟功能要在損壞高達 70%時，血中尿毒指數才會明顯升高。當血中尿毒指數開始升高時，腎臟已是相當脆弱，稍微不慎就會迅速惡化。換句話說，當上述的指數到達可以洗腎的門檻時，腎臟其實已經損壞了 1/3 以上。因此，是否存在其他關鍵檢驗項目，例如腎臟功能的蛋白比值；血液檢查的紅血球數量；尿液檢查的尿液白血球數值等，與腎臟功能有著潛在的關係，這些指標判斷腎臟功能是否異常的能力會較 BUN、Cr 或 CC 更好呢？因此，本論文提出一個血液透析關鍵因子分析技術，採用熵函數找出關鍵檢驗項目，並利用關鍵項目進行透析病患分群，用以判斷關鍵因子之準確性及分群效果。

**關鍵字：**血液透析、洗腎、資料探勘、分群、熵函數

## 1. 前言

腎臟位於人體內的腹腔後壁的脊椎兩側，是身體新陳代謝、排除廢物、毒素、調節血壓、和維持體內液體平衡的重要器官。全身的血液每小時經過腎臟二十次，若腎臟功能受損，身體廢物無法代謝，就會引起腰酸背痛、水腫、尿毒、高血壓、尿道發炎、疲倦、失眠、耳鳴、脫髮、視力模糊、反應遲鈍、情緒低落、恐懼感，甚至神經錯亂等現象。此外，腎臟還是紅血球生成素，若紅血球分泌不足，則病患會有貧血的現象；腎臟也是維持血中鈣磷平衡的最重要成份，若患者有腎衰竭現象，則可能會引發骨病變。

一旦腎臟功能異常，無法正常運作，身體便會產生毒素，危害器官，甚至導致死亡。為了延長或挽救腎臟病患者的生命，常使用腎臟替代療法：腎臟移植(Kidney Transplantation)、血液透析(Hemodialysis, HD)及腹膜透析(Peritoneal Dialysis, PD)進行醫治。腎臟移植是最有效的方法，但會因捐腎者人數稀少，和患者身體條件限制等因素而窒礙難行，取而代之是以人工裝置進行血液透析治療，延續患者生命。

雖然現在醫療技術已經有相當成熟的發展，但是隨著環境的變動，產生疾病的因素也不斷地在改變，任何潛在性的因素都有可能導致疾病的發生，因此許多的學者將資料探勘的技術使用在醫療上，例如分群技術分析、關聯規則分析或者是時間序列分析等，透過分析之目的不同，使用的探勘技術也不相同，選擇一個適合的探勘工具是取得有價值資訊的重要關鍵。但是資料探勘技術對一般的醫療院所而言是很困難的，尤其是在整理原始資料以及選擇探勘工具方面，而且也必須考量到領域

與目的的問題進而修改演算法，以解決問題產生的差異。因此本研究擬設計一套資料探勘系統，協助醫療人員發現與血液透析相關的潛在性因子。

## 2. 文獻探討

### 2.1 血液透析

血液透析就是常稱的「洗腎」，利用人工腎(透析)設備將尿毒素、水份排出體外，減輕尿毒症狀。使用的原理是擴散與超過濾作用，利用半透膜將血液與透析液隔開成二邊，其中一邊將血液不斷帶入人工腎臟中；另一邊則是以透析液不斷帶走擴散過來的尿毒素，並將乾淨的血液引回體內。不斷循環，達到淨化血液的目的。

至於何時該進行血液透析，則依急性或慢性而有不同。若為急性腎衰竭，不需等到尿毒症出現，就會建議病患開始洗腎；若慢性腎衰竭大部分會以藥物先調整，待尿毒症狀出現才開始洗腎。此外，醫師也會依據病患的腎衰竭原因、腎臟尺寸大小、貧血情形、腎功能退化速度及恢復狀況等因素進行評估。此外，也會搭配各項檢驗指標進行評估，目前較常使用的評估指標有：血尿素氮(BUN)、血中肌酐酸 (Creatinine, Cr) 、肌酐酸清除率(Creatinine Clearance)、尿液比重及滲透壓等[2, 3]。

血尿素氮(BUN)：尿素是蛋白質、氨基酸代謝產物，由腎臟排泄，檢測血液中尿素氮的濃度，可以用來評估腎臟排泄尿素的功能是否正常。正常的血尿素氮的範圍為 10～20mg%，如果超過 20mg%則稱之為有高氮質血症。但是因為血尿素氮容易因缺乏水份、吃大量蛋白質食物、上消化道出血、嚴重肝病、感染、使用類固醇藥物，及腎的血流量不足等影響，而暫時性上升。因此如果只有血尿素氮濃度升高，而血肌酐酸濃度正常，腎機能是正常的。所以血尿素氮雖然可做為判斷腎功能的指標，但不如“血中肌酐酸”及“肌酐酸清除率”來得準確。

- **血中肌酐酸(Creatinine, Cr)：**

血中的肌酐酸主要是來自於身體肌肉活動的代謝產物，每天的產量全部都經腎臟由尿液排泄。當腎功能產生問題時，則無法完全排出每日所產生肌酐酸，造成血中肌酐酸濃度上升。上升越高，腎功能越不好。由於肌酐酸是肌肉代謝的廢物，因此血中肌酐酸的濃度與每個人的肌肉總量或體重多少有關，卻與飲食或水份攝取無關。肌酐酸的濃度高低，較血中尿素氮更能準確的顯示腎功能的好壞，血中肌酐酸濃度正常時，並不一定代表腎功能一定正常，最好能夠再檢查所謂"肌酐酸清除率"較為準確。也由於腎臟有相當大的代償功能，一般人血肌酥酸濃度雖然只從 1.4 上升到 1.5mg%而已，事實上，整個腎功能可能已經喪失了 50%以上。

- **肌酐酸清除率 (Creatinine Clearance)：**

是目前在臨床上使用廣泛，較準確的腎機能評估方法，主要評估肌酐酸每分鐘清除幾 C.C。正常人的肌酐酸清除率約為每分鐘 80 到 120C.C，平均約為每分鐘 100C.C。如果算出來清除率只有每分鐘 50 到 70C.C，即表示腎機能有輕度損傷。如果只有每分鐘 30-50C.C，則代表腎機能中度損傷。如肌酐酸清除率小於每分鐘 30C.C，表示腎機能重度損傷，此時尿毒症的症狀會逐漸出現。到了清除率小於每分鐘 10C.C 以下時，則病患應準備開始洗腎治療。肌酐酸清除率的計算方法相當簡單。只要收集整天 24 小時的尿液，檢驗其尿中及血中肌酐酸濃度即可計算。

$$CC = \frac{\text{尿中肌肝酸濃度}(mg\%) \times 24\text{小時尿液總量}(c.c.)}{\text{血中肌肝酸濃度}(mg\%) \times 1440(\text{分鐘})} \quad \text{。(1)}$$

- **尿液比重及滲透壓：**

用以反映腎臟對尿液的濃縮能力。如果測定全天中各次尿液比重均無法達到 1.018 以上，或各次尿液比重差距不到 0.008 以上時，即表示濃縮功能已經受損。另外，如果收集 24 小時尿液，檢測其滲透壓與同時之血

液滲透壓比值，此值應大於 1.0；否則，表示腎濃縮能力失常。或在禁水十二小時後，測其尿及血滲透壓比值，正常比值應該大於 3 以上，否則也是腎濃縮能力受損。濃縮能力異常，經常出現在止痛劑腎病變的病人。

一般而言，當尿素氮大於 90，肌酐酸大於 9 時，肌酐酸清除率小於 0.17 ml/s 或肌酐酸大於 707.2 ml/L，醫生就會建議可以開始進行血液透析治療。但是，腎臟組織在一般正常人約僅發揮 30%的功能，就能維持日常的排毒功能，因此腎臟功能要在損壞高達 70%時，血中尿毒指數才會明顯升高。當血中尿毒指數開始升高時，腎臟已是相當脆弱，稍微不慎就會迅速惡化[1]。換句話說，當上述的指數到達可以洗腎的門檻時，腎臟其實已經損壞了 1/3 以上。因此，是否存在其他檢驗項目，例如腎臟功能的蛋白比值，如表 1；血液檢查的紅血球數量，如表 2；尿液檢查的尿液白血球數值，如表 3，等，與腎臟功能有著潛在的關係，這些指標判斷腎臟功能是否異常的能力會較 BUN、Cr 或 CC 更好？因此，本論文提出一個血液透析關鍵因子分析技術，採用熵函數找出關鍵檢驗項目，並利用關鍵項目進行透析病患分群，用以判斷關鍵因子之準確性及分群效果。

**表 1 腎臟功能檢驗項目**

| 腎臟檢驗項目 | | 參考值 | 單位 |
|---|---|---|---|
| BUN | 血尿素氮（尿毒） | 5-25(最好小於 20) | mg/dl |
| Creaitinine | 肌酐酸 | 0.3-1.4 | mg/dl |
| Uric acid (UA) | 尿酸（痛風） | 2.5-7.0 | mg/dl |
| A/G ratio | 蛋白比值 | 1.0-1.8 | |
| C.C.R./24hrs urine | 腎擴清率/24 小時尿 | M:71-135 F:78-116 | ml/min |
| Penin | 腎活素腎酵素 | 0.15-3.95 | pg/ml/hr |
| Creatinine Urine | 尿中肌酸甘 | 60-250 | mg/dl |

洪冠群學者應用多重最小支持度關聯規則探勘演算法進行洗腎病患住院預測分析[4]，其利用關聯規則分析出可能導致洗腎病患住院的因子，用以降低腎衰竭病患住院次數。

**表 2 血液檢驗項目**

| 血液檢驗項目 | | 參考值 | 單位 |
|---|---|---|---|
| Hb | 血紅素 | 男：14-18<br>女：12-16 | |
| RBC | 紅血球 | 男：450-600<br>女：400-550 | gm% |
| WBC | 白血球 | 5000-10000 | 萬/mm3 |
| Hct | 血容積 | 男：40-55<br>女：37-50 | mm3 |
| Plateles | 血小板 | 15--40.0 | % |
| MCV | 平均血球容積 | 83-100 | 萬/mm3 |
| MCH | 平均血色素蛋白 | 27-32.5 | u3 |
| MCHC | 平均血球血素濃度 | 32-36 | uug |
| Reticulocyte | 網狀球 | 0.5-2.0 | % |
| Malaria | 瘧疾血片檢查 | (-) | /mmhr |
| E.S.R. | 血沉澱率 | 男：1-15<br>女：1-20 | |
| D.C. | 白血球分類 | | % |
| Band | 帶狀球 | 0-2 | % |
| Neutrophils | 藥狀球 | 50-70 | % |
| Lymphocytes | 淋巴球 | 20-40 | % |
| Monocytes | 單核球 | 2--6 | % |
| Eosinophils | 伊紅球 | 1--4 | % |
| Basophils | 鹼性球 | 0--1 | % |
| Bleeging Times | 出血時間 | 0-3 | 分 |
| Coagulation Times | 凝血時間 | 2 月 6 日 | 分 |
| Blood type | 血型 | | |
| Rh Factor | 因子 | (+) | |
| B.P. | 血壓 | | |
| Height | 身高 | | mm/Hg |
| Weight | 體重 | | CM |

他們以病患每月常規的血液透析檢驗項目為主，包含尿素氮(BUN)、肌酐酸(Cr)、尿酸(UA)、鈉(Na)、鉀(K)、鈣(Ca)、磷(IP)和鹼性磷酸酶(ALP)等，搭配 667 項衍生變數(例如，白蛋白指數、單核球是否有感染、是否營養不足等)進行分析。根據實驗結果顯示，他們從 5793 筆有效記錄中取得九項規則，例如糖尿病患伴隨高血脂的病人是最主要的住院病患；透析不足是住院發生的高危險因子；洗腎病患年齡大於 65 歲，很容易因營養不足而住院；女性，年齡介於 40-49 歲之間，有單核球感

染，且最近一次的血紅素檢驗值 (Hb/Ht)過低，則住院的機率很高；當血球容積比(Ht)最近三個月有二次異常，平均血小板體積(MPV)二次異常，總蛋白 (TP)有一次異常時，則住院的可信度是 93 %；若洗腎病患近三個月總蛋白 (TP)、GOT、GPT指數有二次異常，且尿酸異常時，則住院危險度為 100%等。

**表 3 尿液檢驗項目**

| 尿液檢驗項目 | | 參考值 | 單位 |
|---|---|---|---|
| Color-Appearance | 外觀 | | |
| Reaction PH | 酸鹼度 | 5.5-8.5 | |
| Protein | 尿蛋白 | <(+) | mg/ml |
| Sugar | 尿糖 | (-) | g/dl |
| Bilirubin | 膽紅素 | (-) | |
| Urobilinogen | 尿膽素原 | <=1;4 | umol/l |
| BBC | 尿液紅血球 | 0-3 | /1H.F. |
| WBC | 尿液白血球 | 0-5 | /1H.F. |
| Pus Cell | 膿細胞 | 0-1 | /1H.F. |
| Epith Cell | 上皮細胞 | 男：0-3<br>女：0-15 | /1H.F. |
| Casts | 圓柱 | not found | /1H.F. |
| Ketones | 酮體 | (-) | mmol/l |
| Crystals | 結晶 | <=(+) | /1H.F. |
| Bacteria &Other | 細菌 黴菌和其他 | <(+) | /1H.F. |

黃世淵學者於 2009 年進行血液透析患者死亡風險評估與分析[5]，他們利用決策樹(Classification and Regression Tree)、Mann-Whittney U Test、卡方分配、皮爾遜積差相關分析(Pearson Correlation)、Nomogram 等統計方法，對 992 位血液透析病患進行分析。由他們的實驗結果得知，白蛋白、年齡是最重要的影響死亡因子。他們對病患進行分群，再針對分群結果進行分析得知年紀輕、營養好的病患當中，若同時有糖尿病，其死亡機率是無糖尿病患者的 5.45 倍；而年紀大、營養不好的病患，白蛋白及肌酐酸是二個與死亡有顯著差異之關鍵指數，兩者呈正相關關係，兩者相差越低死亡機率越高。由結論可知，白蛋白、年齡、是否有糖尿病、肌酐酸可以協助用來預測死亡危險因子。

針對洗腎患者住院問題，曹全偉學者也提出一個結合時間性摘要與資料探勘技術的序列樣式探勘系統 [6]，他們一樣採用多重最小支持度關聯規則探勘方法找出關聯規則，接著利用 C4.5 決策樹進行時間性摘要的探勘，進而判斷病患是否會住院。他們的實驗結果顯示 6284 筆資料中，共取得 26 條與是否住院相關的規則，例如若白蛋白小於 2.5 gm/dl 以下，則死亡危險性升高 16 倍，住院機率也高；若 Hbc 過低或不足，則有貧血現象，結合營養不足則易造成抗體不足，易受感染而住院等。

葉進儀等學者於2011年使用資料探勘預測血液透析病患住院記錄[19]，假使太多的病患都提早住進醫院接受治療，可能會導致醫療資源之浪費而且血液透析檢驗科的服務品質也會下降，因此他們利用決策樹 C4.5 以及多重最小支持度關聯式規則探勘技術進行分析。C4.5 是用來消去資料中的空值情形，再將 C4.5 處理過後無空值的資料利用最小支持度關聯式規則探勘找出血液透析病患住院的情形，根據病患在住院期間的記錄顯示中發現，有些病患發生慢性疾病的期間可能很常，甚至是根本不會發生，只是在檢驗時醫師依照判斷建議他是否住院，

林宇健學者利用病患的住院記錄結合了關聯式規則以及時間序列樣式演算法建立一套慢性疾病健康照護管理系統[7]，使用關聯式規則找出慢性疾病的相關併發症，並且使用序列樣式找出慢性病患應該注意的疾病，在研究中發現了許多規則，如腦動脈阻塞病患，可能會引發腦血栓症以及腦栓塞症，經醫師鑑定後發現對於避免二次中風的醫療建議方向吻合，而其它未被證實之規則如診斷欠明之心臟疾病，仍需要進一步確定是否有其它併發症等。作者以此做為慢性病患者家屬以及照護人員的參考依據，讓疾病能夠及時預防並進行控制。

上述學者多使用已知的血液檢驗項目進行規則探勘，然而是否存在尚未被發現的檢查項目，跟腎臟功能也有關呢？因此，本論文設計一套血液透析關鍵因子探勘系統，找出與腎臟功能相關的關鍵檢驗

因子，使用方法是資料探勘技術中的熵函數分析法，分析檢驗項目間的差異度，用以判斷檢驗項目與血液透析間的相關性。

## 2.2 熵函數

資訊亂度（Information Gain）是由 Quinlan 於 1979 年提出[15]，主要是做為 ID3（Interactive Dichotometer 3，ID3）決策樹歸納演算法建構決策樹的依據，用於判斷分割點的屬性。計算分類問題時，資訊亂度也可用於判斷特徵屬性與其他類別屬性的差異，常用於決策樹演算法上選擇分裂點(Split Point)。

假設分類問題包含有 $N$ 筆資料，$m$ 個特徵維度，以及 $k$ 個類別，針對單一個特徵的資訊增益測量方式必須計算兩個相關性的數值，這兩個相關性的數值稱為熵值（Entropy），而兩個熵值的差即稱作為資訊亂度。

$$Entropy(N)=\sum P_t \times log\left(\frac{1}{P_t}\right)=-\sum p_t \times log\left(p_t\right) \quad (2)$$

$$Entropy(D_j)=\sum_{v=1}^{|D_j|}\frac{D_{jv}}{N}\times Entropy\left(D_{jv}\right) \quad (3)$$

$$Gain(D_j)=Entropy\left(N\right)-Entropy\left(D_j\right) \quad (4)$$

方程式(1)是 $Entropy(N)$主要計算的是整個分類問題具有的總資訊含量，以此總資訊含量做為計算單一個特徵獲得資訊亂度的依據，其中 $P_t$ 代表的是第 $t$ 個類別在 $N$ 筆資料中出現的機率。

$Entropy(D_{jv})$主要計算的是第 $j$ 個特徵維度中，第 $v$ 種數值、類別與資料數量之間的資訊含量，將特徵中不同數值、類別與資料數量之間的資訊含量加總即為單一個特徵的總資訊含量。其中，$D_{jv}$ 代表的是第 $j$ 個特徵維度中包含有 $v$ 種不同的數值，第 $j$ 個維度共有$|D_j|$個數值。

$Gain(D_j)$代表的即為分類問題中第 $j$ 個特徵所獲得的資訊亂度，是以整個分類問題具有的總資訊含量與第 $j$ 個特徵的總資訊含量之間的差為計算方式。經由方程式(1)至(3)的計算後，就可以獲得分類問題中每一個特徵的資訊亂度，再藉由所設定的門檻值進行評估後，挑選資訊量亂度大於理想目標的特徵，組合成具有足夠分類辨識資訊的特徵集合[15]。我們利用熵函數找出關鍵檢查因子後，再以關鍵因子進行透析病患分群，用以判斷關鍵因子的準確性及分群效果。

## 2.3 分類演算法

目前已有非常多的分類技術被提出來，k-means 演算法是其中最具代表性且被廣泛地使用的一種方法，它亦被稱之為 Generalized Lloyd Algorithm，簡稱 GLA[17]。k-means 分群演算法將每筆資料記錄轉變成一個一個的資料節點(Data Point)，並且利用亂數產生初始的分群中心(Cluster Center)節點，以此來劃分各資料節點歸屬於哪一個中心節點。做為劃分資料節點歸屬的依據是利用資料節點與各個分群中心節點的距離(Distance)遠近來判斷資料節點的歸屬，例如：資料節點離某個分群中心比其它的分群中心還來的近，則此資料節點就會被劃分為此分群中心的資料節點；將每個資料節點的歸屬劃分完後，每一組不同群的資料節點做平均計算，算出此一分群最新的中心點，並以此新中心點為依據進入下一次迭代。以此類推，直至所有的中心點都不再變動為止，則結束分群並呈現最終的分群結果。k-means 簡單的執行步驟如下所示：

(1) 亂數產生初始的分群中心 $C_i$ = {1, 2, 3, …, $k$}。

(2) 計算每一個資料節點 $X$={$x_1$, $x_2$, …, $x_m$} 與每一個中心點 $C_i$ 的歐式距離 $d(X, C_i)$，並將距離最近的節點歸類至該群 $C_i$。距離公式如下所示：

$$d(X,C_i)=\sqrt{\sum_{j=1}^{m}(x_j-c_{ij})^2} \quad (5)$$

(3) 重新計算新的中心節點 $C_i$，假使所有新的中心節點移動的幅度達要求或是不會再移動了，則結束所有的分群作業，否則將繼續回到(1)和(2)步驟執行分群作業。

## 3. 研究方法

本研究將針對血液透析資料進行分析，利用熵函數技術找出的關鍵因子，並進行透析病患分群，將特性相同的病患分類至同一群，以觀察每群之間的同質性與相異程度，進而判斷關鍵因子的準確性。

本研究共分為四個階段，如下圖所示。

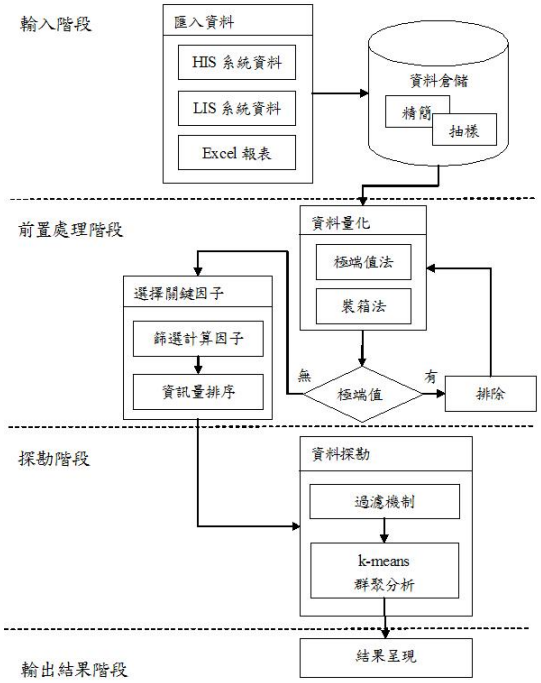


**圖 1 系統架構圖**

第一階段為輸入階段，主要目標在確認目標疾病，將來自不同來源與格式之資料統整至資料庫中，並做精簡與抽樣處理。此階段的處理將會對後期階段有很大的影響，因此得小心的處理。第二階段為前置處理階段，本階段共為二個子階段，其一為將對原始數據進行量化轉換，將所有數據轉換成可進行探勘研究工作之形式，例如中文字串轉換成數字代號，數值資料轉換成相似間距之資料格式；其二為計算每個檢驗項目與疾病之間的資訊量，並以此為關鍵因子。第三階段為探勘階段，進行病患記錄分群，將性質相似者與性質相異者進行劃分。第四階段為輸出階段：將整個探勘結果呈現出來，並且請專業人員解釋結果，從中找出任何可能是引發疾病的潛在因子。

### 3.1 輸入階段

檢驗資料來源眾多，例如有些是由醫院資訊系統(Hospital Information System, HIS)資料庫中篩選出來的資料、有些是由實驗室系統(Laboratory Information System, LIS)產生的，或是由已產生的 Excel 報表檔匯入等。不同來源資料格式或儲存方式可能不同，資料型態、表示也可能會不同，例如在 A 資料庫性別是用 1 和 2 表示，1 表示男生、2 表示女生，但是在 B 資料庫性別直接是用男和女來表示，當資料彙整時就會發生錯誤。為了確保所分析的資料是正確、完整且充足的，我們必須在資料匯入前先進行一些事前處理作業。一般而言，事前處理時所要解決的問題大致可以分成四類：

- **統一資料儲存方式：**為了方便進行探勘，需先將異質資料統一整理成相同格式。
- **無關的資料：**針對特定主題的資料探勘，僅須針對與該主題相關之資料作處理，至於無關的資料若不加以篩選，將會影響探勘效率，甚至探勘成果準確度。
- **不正確的資料：**錯誤的資料可能是來源發生錯誤或登錄資料庫時發生錯誤，對於錯誤的資料，應盡可能加以行修正或直接刪除。
- **格式不合：**資料庫設計時未考慮到日後的資料探勘應用。此外，不同主題的資料探勘或探勘方法、工具要求不同的資料格式，因此，為使探勘能夠順利的進行，有必要將資料作適當的轉換。
- **不完整的資料：**資料不齊全是常見的現象，可能的狀況包括記錄中的欄位資料遺失，或缺乏某期間的資料，或缺少某些分析時需要用到的屬性。

### 3.2 前置處理階段

接著進行資料正規化，其目的在將不同標準之下的記錄轉換到同一個標準，以便提高分析時的準確性。每項檢驗項目若

有檢驗標準數值，則以檢驗標準數值進行區隔，例如 TG 為三甘油脂(Triglycerides)，其檢驗標準數值若大於或等於 201 則超過標準，則正規化為 100；若 20～200 則為正常，則正規化為 50；若小於等於 19 則低於標準，則正規化為 0。若無檢驗標準數值，且若為連續數質型欄位則以裝箱法正規化方式進行轉換，其公式為：

$$v' = \left\lfloor \left( \frac{v - \min}{\max - \min} \right) \times Q \right\rfloor , \qquad (6)$$

其中 *v* 為原始資料，*min* 為該欄位最小值，*max* 為該欄位最大值，*v'*為正規化後之值。*Q* 為量化間距，使用者可依檢驗項目值域範圍，設定不同的量化間距。例如，WBC 最大值 10.7 與最小值 3.5 之間的差異為 7.2，我們設定間距 *Q* 為 4；RBC 最大值 5.4 與最小值 3.89 之間的差異為 1.51，我們設定間距 *Q* 為 3。表 4 為正規化後之資料。

**表 4 裝箱法正規化後之檢驗資料**

| | 性別 | 年齡 | WBC | RBC | HB | BUN | CRE | UA | GOT | GPT | TP | ALB | GLO | A/G | TG | 洗腎 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | 2 | 5 | 4 | 3 | 3 | 4 | 2 | 2 | 4 | 5 | 2 | 2 | 2 | 2 | 3 | 2 |
| 1 | 1 | 3 | 1 | 1 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 3 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 5 | 1 | 3 | 1 | 1 | 1 | 2 | 1 | 0 | 3 | 4 | 1 | 1 | 1 | 0 | 1 | 0 |
| 6 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 7 | 0 | 4 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 8 | 1 | 1 | 1 | 1 | 2 | 3 | 1 | 0 | 2 | 4 | 0 | 0 | 0 | 1 | 2 | 1 |
| 9 | 1 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 0 |
| 10 | 0 | 2 | 3 | 1 | 0 | 2 | 0 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | 2 | 0 |
| 11 | 1 | 0 | 1 | 2 | 2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 12 | 0 | 0 | 1 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 13 | 0 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 14 | 1 | 2 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| 15 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

上表為進行資訊量計算時所使用的正規化表，為了使系統進行分群時，可以有效的劃分不同病患記錄之間的差異，本研究同時利用極值正規化法進行量化，其公式為：

$$v'' = \frac{v - \min}{\max - \min} \times 100 , \qquad (7)$$

其中 *v* 為原始資料，*min* 為該欄位最小值，*max* 為該欄位最大值，*v"*為極值正規化後之值。例如，WBC 最大值 10.7 與最小值 3.5 之間的差異為 7.2，經由公式(7)計算後得到 $v'' = \left[ (1 - 3.5) / (10.7 - 3.5) \right] \times 100 = 38.89\%$ 。

在量化的過程中，對量化結果影響最大的是每個項目在整個資料集中的最大值與最小值，假使其中出現極端值，在量化時會導致異常的現象，舉例而言，CRE 區間值 *Q*=80，其指數大多在 0.37~2.99 之間，但某筆記錄卻為 6990，在量化時就會量化出兩極化的資料，0.37~2.99 會全落在區間值為 1，而 6990 這筆記錄則會落在區間值為 80。因此，為了消除資料集中的極端值，在量化的過程中設定一個機制，以防止極端值影響量化的正確性。當量化完成後，系統會先進行一次判斷，假使某個項目的某個區間內的記錄筆數低於整個資料庫的 0.05%，則此一區間內的所有記錄會被修正成佔多數的區間值之內容，再以此新的記錄重新進行量化，直到所有區間內之記錄個數都達到標準。

## 3.3 資訊量分析

本論文主要針對洗腎病患的相關檢查進行分析，故先針對“洗腎”欄位進行總資訊含量的計算。以表 4 為例，有洗腎的病患共有 6 筆 (洗腎=1)，發生機率為 $P_1 = 6/15$ ，資訊量為 $P_1 \times \log(1/P_1) = \frac{6}{15} \times \log(\frac{6}{15}) = 0.528771$ ；沒有洗腎的病患共有 9 筆，發生機率為 $P_0 = 9/15$ ，資訊量為 $P_0 \times \log(1/P_0) = \frac{9}{15} \times \log(\frac{9}{15}) = 0.442179$ 。二個資訊量相加的結果，即為總資訊含量 0.970951。

接著進行每個檢查項目與“洗腎”欄位之相關性運算，計算每個檢查項目的資料量。以下我們以“性別”欄位為例，由表 5 可知，性別為女(性別=0)的總筆數有 7 筆，性別為女(性別=0)且沒有洗腎(洗腎=0)的記錄有 4 筆，故性別為 0 且洗腎為 0 的機率為 $P_{D_{jv}}$ =4/7，資訊量為 0.46；而性別為女(性別=0)且有洗腎(洗腎=1)的記錄有 3 筆，機率為 $P_{D_{jv}}$ =3/7，資訊量為 0.52，二者相加得到性別為女的資訊量為 0.46+0.52=0.99。由於性別為女的資料只占

全部記錄的 7/16，故計算“性別”欄位的總資訊量時，需將性別為女的資訊量乘上機率 $7/16$，得到 0.99×(7/16) =0.459773，該資訊量與性別為男的資訊量 0.509031 相加得到“性別”欄位的總資訊量 0.459773+0.509031 = 0.968805。由公式(3) *Entropy*(*N*) - *Entropy*($D_j$)，可推得“性別”相對於“洗腎”的資訊亂度，故其亂度為 0.002145996 = 0.970951 - 0.968805。

以此類推，我們可以找出每個檢查項目與“洗腎”之間的關係，亂度越大表示資訊量越高。接著再將資訊亂度由大到小排序，找出前幾項做為關鍵因子，進行關聯規則探勘。各檢驗項目之資訊含量如表 6，若取前三項進行探勘，則亂度最高的前三項：年齡、WBC、BUN，將視為關鍵因子。

**表 5 “性別”相對於“洗腎”之資訊含量計算**

| 性別 $j$ | 洗腎 | Count($D_{jv}$) | $P_{D_{jv}}$ | $P_{D_{jv}} \times \log(1/P_{D_{jv}})$ | *Entropy*($D_{jv}$) | *Entropy*($D_j$) |
|---|---|---|---|---|---|---|
| 0 | 0 | 4 | 4/7 | 0.46 | 0.99 | 0.459773 |
| | 1 | 3 | 3/7 | 0.52 | | |
| 1 | 0 | 5 | 5/8 | 0.42 | 0.95 | 0.509031 |
| | 1 | 3 | 3/8 | 0.53 | | |
| sum | | | | | | 0.968805 |

**表 6 各檢驗項目之資訊含量**

| 欄位 | 性別 | 年齡 | WBC | RBC | HB | BUN | CRE | UA | GOT | GPT | TP | ALB | GLO | A/G | TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 資訊量 | 0.002 | 0.577 | 0.329 | 0.14 | 0.06 | 0.28 | 0.05 | 0.09 | 0.18 | 0.2 | 0.05 | 0.24 | 0.02 | 0.06 | 0.03 |

## 3.4 資料探勘

在進行資料探勘前，得先對數據進行過濾，病患會因身體狀況或醫師的建議進行身體檢查，每個記錄並非是完整的，可能會有遺失的記錄，而某些記錄雖然有遺失值，但是卻可能會有一些隱藏的資訊藏在其中。因此設定一個門檻值，若關鍵因子欄位空值的個數大於門檻值，則此筆病患記錄會被排除。過濾後之資料再利用分群演算法進行分類。

本研究將關鍵因子做為分群依據，進行 k-means 分群，令關鍵因子為 $d_1, d_2, \ldots, d_m$，共有 $m$ 個關鍵因子，假設病患記錄為 $X=\{x_1, x_2, \ldots, x_m\}$，$x_j$ 為病患 $X$ 第 $j$ 項關鍵因子的檢驗值，$1 \leqq j \leqq m$。將所有病患資料以下列演算法進行分群，群體個數為 $k$。

(1) 首先取得資料庫中的所有病患記錄，並且以隨機的方式產生出 $k$ 個初始中心點，$C_i=\{c_1, c_2, \ldots, c_m\}$ 為第 $i$ 群中心點記錄。如圖 2 (a)所示，實心圓共有 $N$=10 個，代表各個記錄所在的位置，三角形共有 $k$=3 個，代表群中心點 $C_i$ 所在的位置。

(2) 利用公式(5)之 $d(X, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - c_{ij})^2}$ 計算出每筆病患記錄與每個群中心點 $C_i$ 的歐式距離，將病患記錄 $X$ 歸類至距離最近的群中心。例如某個病患 $X$ 與各個群中心之距離為 $d_2 < d_8 < \ldots < d_6$，$d_2$ 的距離與其它群中心之距離相較之下最小，則病患 $X$ 會被歸類至 $C_2$，以此類推，此產生出各群的成員集合。

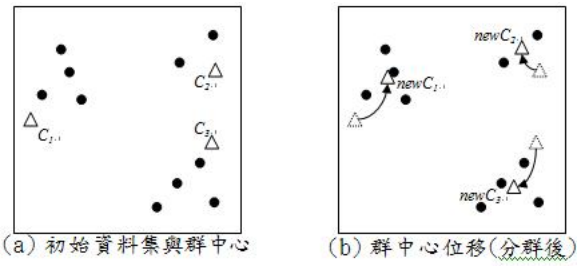


**圖 2 分群演算法示意圖**

(3) 根據每個群中心 $C_i$ 各自的成員集合 $C_i=\{X_{c_i1}, X_{c_i2}, \ldots, X_{c_iS}\}$，其中 $S$ 表示 $C_i$ 群內成員數有 $S$ 個，$X_{c_iu}$ 為 $C_i$ 群中第 $u$ 個病患記錄。將成員的關鍵因子各自加總後除以 $C_i$ 群的成員總數 $S$，即可得到新的群中心 $newC_i = \left\{\sum_{u=1}^{S} x_{u1}/S, \sum_{u=1}^{S} x_{u2}/S, \ldots, \sum_{u=1}^{S} x_{uj}/S\right\}$，並以此做為新的群中心點。如圖 2 (b)所示，經過步驟(2)與(3)後，$C_1$ 群其歸屬的 4 個記錄節點經平均計算後，得到新的群中心點 $newC_1$，並且以 $newC_1$ 取代 $C_1$。

(4) 重複步驟(2)和(3)，直到所有的群中心 $C_i$ 不再變動為止，即可完成 k-means 分群。

## 4. 實驗結果

本研究以某醫院檢驗科提供之檢驗資料進行分析，其中血液透析相關資料主要

來自“洗腎科”及一般門診各科室執行與腎臟相關之檢驗報告記錄。原始資料集建立完全後，發現此家醫院之檢驗相關項目多達 100 多種，但是並不是每個病患都會執行所有檢查，導致 100 多個項目欄位將近有 70%以上會出現空值的情形，因此得先對這些不完整的記錄進行篩選。本研究依據市面上確定與腎功能檢查相關的 2 個項目血中尿素氮以及血清肌酐酸為基準，判斷這 2 個指標項目只要其中一個出現空值，則排除這筆記錄。經過篩選過濾後，可以使用的檢驗記錄共 4662 筆記錄。

當檢驗資料建立完成後，接著要進行量化作業，量化作業旨在將連續性的數值或是數值差異性很大的值轉化成相似的有限區間值。各檢驗項目之間距依據醫療人員的建議進行區分，間距如下表所示。

**表 7 各檢驗項目之間距**

| ID | 項目 | 分群間距 |
|---|---|---|
| 1 | TG | 50 |
| 2 | AST(GOT) | 20 |
| 3 | Ch | 50 |
| 4 | ALT(GPT) | 20 |
| 5 | UA | 2 |
| 6 | K(Boold) | 2 |
| 7 | BUN | 5 |
| 8 | Amylase(B) | 50 |
| … | … | … |

## 4.1 尋找關鍵因子

檢驗資料的項目眾多，如果全部都列入探勘項目進行探勘，會探勘出毫無意義的結果，因此得過濾掉不相關的項目。本研究利用熵函數計算出每個項目與“洗腎”的資訊含量，並且使用排序功能，依照資訊量之大小排序，找出對其影響最大的前幾名項目，在探勘時取前幾名的項目當成關鍵因子使用。假設取前四名關鍵因子，則 TG=0.92653 、 AST(GOT)=0.86141 、 Ch=0.85901 和 ALT(GPT)=0.85824 將會成為探勘時的關鍵因子。

## 4.2 資料探勘處理

依前述介紹之分群演算法，將病患記錄進行分群，並且自動評估當次分群結果的優劣。在研究中設定分群個數為 4 群，分群演算法重覆執行 10 次並從中選出評估最好的分群結果呈現。圖 3 為分群結果，呈現出 4 群各自的代表特徵項目指標，例如被分類至第 1 群的 1991 筆檢驗記錄，其代表的平均特徵各自為 AST(GOT)=24.05 、 ALT(GPT)=24.18 、 TG=144.05 和 Ch=190.37；群與群之間的整體平均差異為 Differences = 24.8735，而群內密度的整體平均為 5.6，群整體評估為 overall = 9.6，是執行 10 次分群計算中最好的結果。

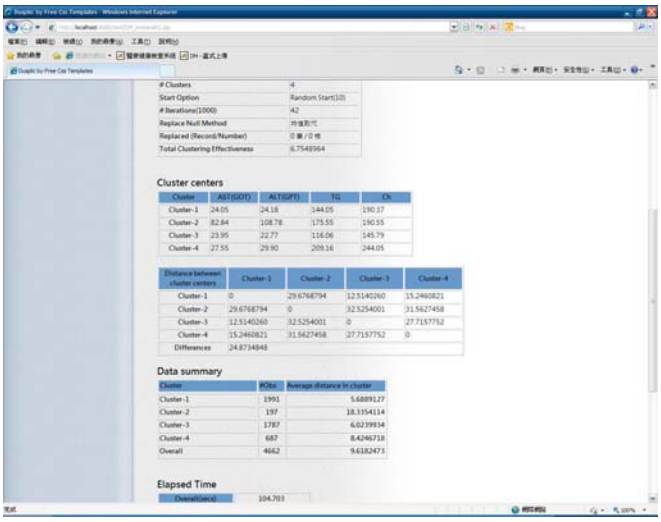


**圖 3 分群結果呈現**

## 4.3 結果呈現

本研究使用 4662 筆病患記錄以及 105 個檢驗項目，並計算出前四名與洗腎最相關的關鍵因子：麩氨基酸草醋酸轉氨基酵素 AST(GOT)、麩氨基酸焦葡萄轉氨基酵素 ALT(GPT)、三甘油脂 TG 和總體膽固醇 Ch，其中 AST(GOT)和 ALT(GPT)是肝功能檢驗的主要指標，但是在研究中發現，與洗腎的關係程度為第 2 及第 4 名。

將上述四個項目視為關鍵因子，分成 4 群利用 k-means 分群技術進行探勘，得到每群各自的代表特徵值：第 1 群特徵為 AST(GOT)=24.05 、 ALT(GPT)=24.18 、 TG=144.05、Ch=190.37，共有 1991 筆記錄且群內密度為 5.69；第 2 群特徵為 AST(GOT)=82.84 、 ALT(GPT)=108.78 、

TG=175.55、Ch=190.55，共有 197 筆記錄且群內密度為 18.34；第 3 群特徵為 AST(GOT)=23.95 、 ALT(GPT)=22.77 、 TG=116.06、Ch=145.79，共有 1787 筆記錄且群內密度為 6.02；第 4 群特徵為 AST(GOT)=27.55 、 ALT(GPT)=29.90 、 TG=209.16、Ch=244.05，共有 687 筆記錄且群內密度為 8.42。各群之間的平均差異為 24.87，群內密度之平均為 6.75。

將以上探勘結果與醫療人員討論後發現，在尋找關鍵因子階段，本研究找出 4 個項目，其中 2 個項目 TG 以及 Ch 是目前文獻上沒有提及到的檢驗項目，未來還有進一步研究的空間；分群結果各群的指標都落在應有的指數上。

## 5. 結論

為了降低疾病發生的可能性，各家醫療人員都試圖從病患的檢驗資料中發現更多的訊息，但是人為的觀察常常會導致某些資訊被忽略，或是想法被限制於教科書上的資訊。雖然已經有很多預測或探勘疾病的方法被提出來，但是大多數都是針對已知的檢驗項目進行研究，鮮少有為了尋找潛藏性關鍵因子的方法被提出，原因在於檢驗之項目眾多且不完整，很難利用系統找出項目之間的關聯性。

本研究利用探勘技術協助醫療人員尋找未知的潛在疾病發生因子，透過分群技術分析，協助醫療人員利用過往的檢驗記錄，從分群結果找出各種項目之間的潛藏關係，從中找出新的預測疾病關鍵因子，以幫助病患從新的指標中察覺疾病發生的可能性。

## 參考文獻


[1] 呂至剛，“腎臟功能為什麼會不好”，新光醫訊第 178 期，2006。

[2] 婦女健康醫療網 http://www.drkao.com/1st_site/health_wap/normal_main.htm

[3] "如何檢測腎功能"，綠十字健康網，2011/11/17 ， http://www.greencross.org.tw/kidney/symptom_sign/kid_func.html

[4] 洪冠群，「多重最小支持度關聯規則探勘演算法之醫療檢驗應用：以血液透析病患之住院預測為例」，國立東華大學資訊工程學系碩士在職專班論文，2004。

[5] 黃世淵，「血液透析患者死亡風險評估與分析」，台北醫學大學醫學資訊研究所碩士論文，2009。

[6] 曹全偉，「結合時間性摘要與資料探勘技術於血液透析病患資料之分析」，國立嘉義大學資訊管理學系碩士論文，嘉義縣，2008。

[7] 林宇建，「資料探勘技術應用於慢性疾病健康照護管理系統」，靜宜大學資訊管理學系碩士論文，2008。

[8] 行政院衛生署中央健康保險局 http://www.nhi.gov.tw/。

[9] 行政院衛生署 http://www.doh.gov.tw/cht2006/index_populace.aspx。

[10] 范仕遠，「聚類分析」，第三屆離島資訊技術與應用研討會，大葉大學電機工程研究所，2003。

[11] 楊正宏、柯兆軒、莊麗月和楊正三，「用於基因微陣列資料的兩階段式基因選擇」，Proceeding of International Medical Informatics Symposium in Taiwan，2007。

[12] 吳昆益，「植基於類神經網路之顧客交易特徵分析機制」，碩士論文，朝陽科技大學資訊管理系碩士論文，2008。

[13] 王偉驎、林文燦、賴政皓和陳慧敏，「應用資料探勘技術提升急診醫學檢傷分類之一致性－以台灣某醫學中心急診醫學部為例」，品質學報，第 15 期，第 4 卷，2008。

[14] 藍國誠、李昭輝、李語嫣、吳晉祥、黎煥中和曾新穆，「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」，國際醫學資訊研討會論文集，2009。

[15] J. R., Quinlan, “Induction of Decision Trees,” *Machine Learning*, No. 1, pp. 81-106, 1986.

[16] Z. C., Lai and Y. C., Liaw, “Improvement of the k-Means Clustering Filtering Algorithm,” *Pattern Recognition*, Vol. 41, pp. 3677-3681, 2008.

[17] Z. C., Lai, T. J., Huang, and Y. C., Liaw, “A Fast k-Means Clustering Algorithm Using Cluster Center Displacement,” *Pattern Recognition*, Vol. 42, pp. 2551-2556, 2009.

[18] W. T., Lin, S. T., Wang, T. C., Chiang, Y. X., Shi, W. Y., Chen, and H. M., Chen, “Abnormal Diagnosis of Emergency Department Triage Explored with Data Mining Technology: Emergency Department at a Medical Center in Taiwan Taken as an Example,” *Expert System with Applications*, Vol. 37, pp. 2733-2741, 2010.

[19] J. Y., Yeh, T. H., Wu, and C. W., Tsao, “Using Data Mining Techniques to Predict Hospitalization of Hemodialysis Patients,” *Decision Support Systems*, Vol. 50, pp. 439-448, 2011.

[20] C. D., Chang, C. C., Wang, and Bernard C. Jiang, “Using Data Mining Techniques for Multi-Diseases Prediction Modeling of Hypertension and Hyperlipidemia by Common Risk Factors,” *Expert Systems with Applications*, Vol. 38, pp. 5507-5513, 2011.

[21] N. A., Setiawan, P. A., Venkatachalam, and A. F. M., Hani, “Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System,” *Proceedings of the International Conference on Man-Machine Systems(ICoMMS)*, pp. IC3-1-IC3-5, 2009.

[22] H., Shoji, T., Shusaku, “Temporal Data Mining in Hospital Information Systems: Analysis of Clinical Courses of Chronic Hepatitis,” *IC-MED*, Vol. 1, No.1, Issue 1, pp. 11-19, 2007.

[23] E. M., Theodoraki, S., Katsaragakis, C., Koukouvinos, and C., Parpoula, “Innovative Data Mining Approaches for Outcome Prediction of Trauma Patients,” *J. Biomedical Science and Engineering*, Vol. 3, pp. 791-798, 2010.