

運用關聯規則分析技術探討某醫院患者特質

曾若嫻¹ 林家宏² 吳惠鳳³ 陳俞成⁴

¹ 嘉南藥理科技大學醫務管理系碩士在職專班 angelar52007@yahoo.com.tw

² 嘉南藥理科技大學醫務管理系碩士在職專班 s582580@hotmail.com

³ 嘉南藥理科技大學醫務管理系碩士在職專班 k55991@yahoo.com.tw

⁴ 嘉南藥理科技大學醫務管理系(所) ycchen@mail.chna.edu.tw

摘要

健康檢查與健康管理在目前醫學下逐漸受到重視，以科學方法與資訊科技來分析與客戶關係的特性更是一種現代趨勢。建立分析模式了解個案醫院就診的顧客屬性，以作為醫院擴大服務對象的參考。本研究將所收集的病患資料，先進行資料過濾，再採用「關聯規則」實施分析處理。運用群集分析、關聯規則演算法與 Apriori 演算法等。經過統計分析該院最多人就診的科別為心臟內科與腸胃內科，患者也以 40 歲以上人群有關。關聯規則發現，腸胃內科病患多為 40-80 歲，心臟內科的病患年齡層提高為 50-80 歲。經過演算結果了解，就醫人口以戶籍地人口為多、老人為多數，因此個案醫院拓展服務項目可以置重點於老人心臟及腸胃內科患者的醫療服務。

關鍵字：資料過濾 關聯規則 群聚分析

1. 研究動機與目的

隨著人口老化與經濟 M 型化，預防醫學的觀念日趨受到重視，加上目前醫療機構均極力發展非健保給付的自費醫療，進而演變高階健康檢查或健康管理服務已經蔚為潮流。因此，許多醫機構均開始致力分析目標族群的組成並針對其需求特性發展出適切的服務。透過電子化病歷工作的推廣，使研究者可透過科學的資訊工具與方法找出目標客群，分析其特性並進一步執行客戶關係管理。

多數就醫人口之調查研究中，策重於疾病分佈與健保資源使用概況的描述，對病人來源與居住地之分析結果，較侷限於該院或醫院聯盟間應用，因此無法初窺單一醫院就醫人口在行政區域的分佈情形，更難以比較顧客就醫屬性及其就醫情形的差異性。因此，本研究之目的主要在以個案醫院 97 年的門診服務量，建立就診病患分析模式，了解就診顧客的就醫屬性，進一步探討行政區域分佈與服務社區內民眾的就醫情形，以做為個案醫院擴展服務對象的參考。

2. 文獻探討

2.1 資料探勘

資料探勘是指從巨大的資料倉儲找尋隱藏在資料中的訊息的一種過程與技術¹。如趨勢(Trend)、特徵(Pattern)、相關性(Relationship)、資料樣型分析(Data Pattern Analysis)、功能相依分析(Functional Dependency Analysis)等²，其主要目的是方便使用者使用資料探勘的工具，不一定要具有統計背景；本次將運用資料探勘的群集分析探索癌症病患的特質。

2.2 關聯規則演算法

Agrawal³等學者首先提出一個數學模式，用以找出項目之間的關聯性，並以關聯規則表示，令 $I = \{i_1, i_2, \dots, i_n\}$ ， I 即是所欲討論的項目(items)所組成的集合，在此 I 可視為百貨店或超市內所有商品組成的集合， D 是一個交易的集合，亦可視為一特定資料庫，其中每一個交易 T 是項目的集合，注意，每筆交易中商品項目購買的數量是不考量的，另外每筆交易皆有一個交易序號 TID 做為識別，如表 1 所示。

表1 交易資料庫

| TID | 交易項目 |
|-----|------|
| 100 | ACD |
| 200 | BCE |
| 300 | ABCE |
| 400 | BE |

我們說明關聯規則的定義；假設 I 是交易資料庫中所有項目的集合， T 是全部交易資料的集合，一筆交易資料 T_j ， $T_j \in T$ ，是由一些項目所形成的集合，稱之為項目組(itemsets)，項目組的項目個數 $|T_j| \geq 1$ 。在項目組 X 與 Y 之間有一關聯規則被表示 $X \rightarrow Y$ ， $X, Y \subseteq I$ 且 $X \cap Y = \emptyset$ 。有兩個參數 s 與 c 分別為支持度(support)與信賴度(confidence)，用來決定關聯規則是否為有效規則(strong rules)；關聯規則 $X \rightarrow Y$ 的支持度 s 表示為：在所有的交易集

合中，同時包含有 X 與 Y 的比率值，即 $s=(\text{同時包含有 } X、Y \text{ 的交易數量})/(\text{總交易數量})$ ；而關聯規則 $X \rightarrow Y$ 的信賴度 c 表示為：在包含有 X 的交易集合中，也同時包括有 Y 的比率值，即 $c=(\text{同時包含有 } X、Y \text{ 的交易數量})/(\text{包含有 } X \text{ 的交易數量})$ 。擷取出來的關聯規則，其支持度與信賴度必須大於或等於所指定的最小支持度與最小信賴度，這樣的關聯規則才有意義。

擷取關聯規則最具代表性的方法為 Apriori 演算法⁴，其探勘過程主要分成兩個階段：第一階段先找出滿足最小支持度的所有項目組，也就是在交易資料庫中包含此項目組的交易數量必須大於或等於所指定的最小支持數量。然後，這些滿足最小支持數量的項目組就稱之為高頻項目組(frequent itemsets)，若一個項目組包含有 k 個項目，稱之為 k 項目組(k -itemsets)；若某 k -項目組滿足最小支持數量，即稱之為高頻 k 項目組(frequent k -itemsets)。第二階段就根據前階段所找出的高頻項目組及以最小信賴度為條件，計算出所有符合的關聯規則；例如 ABC 為高頻 3 項目組， $A、B、C \subseteq I$ ，假如關聯規則 $AB \rightarrow C$ 滿足最小信賴度，則此關聯規則成立。

2.3 Apriori 演算法

根據表 1，利用 Agrawal 等學者所提出的 Apriori 演算法運作過程如圖 1 所示(假設 Support 為 50%)，其演算法說明如下：

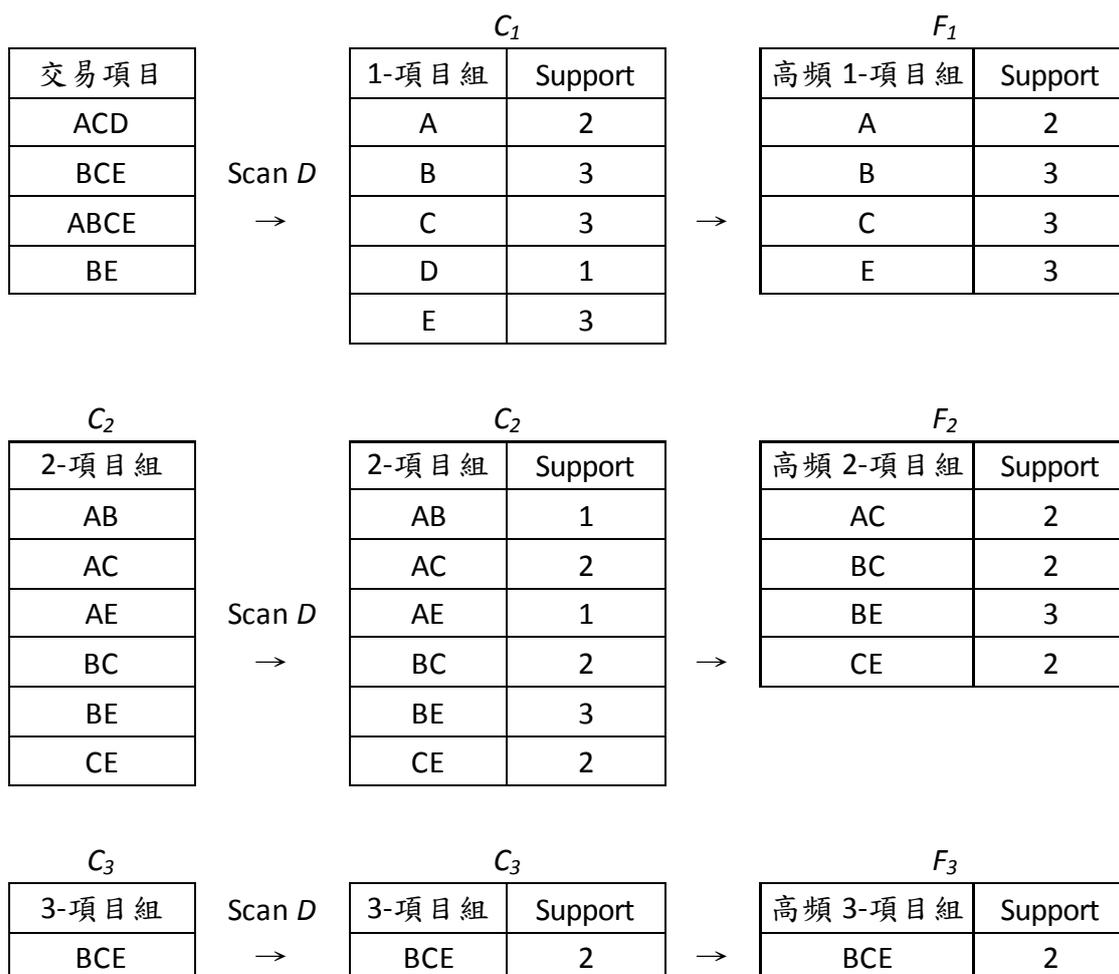


圖 1 高頻項目組之產生過程

首先掃描整個資料庫，對每一個商品項目出現在所有記錄的次數進行計數，得到表格 C_1 中每一個 itemset 的計數值，若某件商品項目出現次數少於所訂定最小頻率係數，則便會被刪除，得到表格 F_1 ，在此 itemset 內的項目，皆經過排序。

在搜尋高頻 k -項目組，第二階段一開始先依據表格 F_1 中每一個 itemset 做為參考，做兩件商品項目的組合，得到 C_2 表格，再依據 C_2 表格內的 itemset 來掃描整個資料庫，對每 itemset 的商品組合出現在所有記錄的次數進行計數，得到表格 C_2 中每一個 itemset 的計數值，若兩件商品項目組合出現次數少於所訂定最小頻率係數，則便會被刪除，得到表格 F_2 ，而表格 F_2 中的 itemset 皆為高頻 2-項目組。接下來由表格 L_2 中求得三件商品項目組合，此三件商品項目組合必須符合其次一階商品項目組合（兩件商品項目組合），皆需包含在 F_2 中，例如要得到三件商品項目組合 BCE，則在 F_2 中的 itemset 必須包含 BC、BE、CE，如此便可得到表格 C_3 ，再依據 C_3 表格內的 itemset 來掃描整個資料庫，對每 itemset 的商品組合出現在所有記錄的次數進行計數，得到表格 C_3 中每一個 itemset 的計數值，若三件商品項目組合出現次數少於所訂定最小頻率係數，則便會被刪除，得到表格 F_3 ， F_3 內的 itemset 為高頻 3-項目組，在此例中， F_3 內的 BCE 為唯一的高頻-3 項目組。此一階段會一直進行直到表格 F_k 中的高頻 k -項目組，無法求出任何一項高頻 $(k+1)$ -項目組為止。

2.4 顧客關係管理

張瑋倫(2008)⁴認為顧客關係管理真正價值在於轉換策略、操作流程與企業功能來保留顧客並創造顧客忠誠度與利潤。因此，在顧客關係管理操作上可以依據 Winer(2001) 所提出的 7 個基本運作流程：顧客活動的資料庫、資料庫分析、決定目標顧客群、爭取目標顧客群的工具、與目標顧客群建立關係、隱私權考量及衡量指標建立。

2.5 病患就醫行為考慮因素

陳尚志(2009)⁶認為在國內許多研究中指出，病患就醫選擇考量的要點可以歸納文 5 個因素：可用性、可近性、適應性、負擔力與接受性，以舊衣考慮因素為出發點，分析即有顧客群選擇到院就診的特性再加以分析顧客特質有利於醫院對病患的關析管理工作。

3. 研究方法

3.1 資料來源

某醫院 97 年門急診及住院資料。

3.2 研究對象

本次是針對醫院門急診掛號資料，分析就醫患者人口特徵包括(1)年齡：區分 80 歲者以上、70-80 歲者、60-70 歲者、50-60 歲者、40-50 歲者、30-40 歲者、20-30 歲者及 20 歲者以下等 8 階層。(2)居住地區：實施初步資料分析後步，依據病患就醫選擇的可近性效益，該院就診病患多來自於台南市、台南縣、高雄市及高雄縣等 4 行政地區(67%)，為凝聚分析中心將就醫病患限制於高雄與台南縣市的就診病患。(3)就診季節：以門診就診時間結合農曆季節區分，將病患的就醫時間區分為春、夏、秋、冬等四季。

3.3 研究工具

(1)以醫院門診的掛號資料為資料庫查閱病歷表。(2)資料探勘分析：本研究將所收集的病患資料，前置以 PASW Statistics 18 進行資料過濾，再以 SPSS Clementine for windows 12.0 資料探勘套裝軟體，採用「關聯規則」實施分析處理。

4. 結果與討論

經收集、過濾(剔除屬性重複與離值項目)與整理看診日期、患者性別、患者生日、戶籍代碼、戶籍名稱、科別代碼、科別名稱、主診斷碼及診斷名稱等 9 項資料後，選用就診季節、年齡層、患者性別、戶籍代碼、科別代碼等 5 項資料，其中戶籍代碼設定為台南與高雄縣市患者，因為經初步描述分析，台南與高雄縣市患者就醫距離的可近性遠大於其他地區，為凝聚分析重點故僅留用該地區就醫患者資料(圖 2、表 2)。

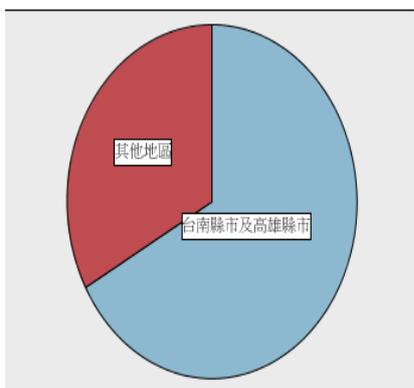


圖 2 就診病患居住地分布

| 表 2 就診病患居住地統計表 | | |
|--------------------------------|---------|-----|
| 地區別 | 就診人數 | 百分比 |
| 台南市東區、南區 仁德鄉、歸仁鄉 湖內鄉、路竹鄉 | 516,731 | 67 |
| 其他地區 | 254,974 | 33 |
| 合計 | 771,705 | 100 |

經過次數分配統計分析，該醫院 97 年台南與高雄縣市的門急診就醫人數計有 771,705 人次，就診季節上沒有明顯差異，病患主要年齡層為 40 歲以上老年人居多(圖 3)，居住地區以台南市為主且其中又以東區及西區較多(表 3)，就診性別上也沒有明顯差異，就診科別中以心臟內科與腸胃內科就診人次較多(表 4)。

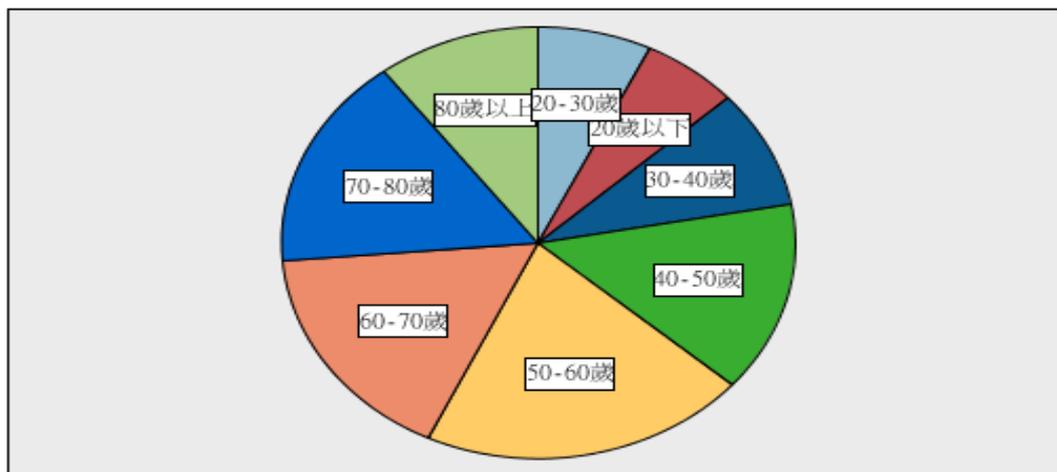


圖 3 病患主要年齡層分布情形

表 3 就診病患居住地統計表

| 戶籍-代碼 | 就診人數 | 百分比 |
|-------------|---------|-----|
| 台南市東區 2101 | 205,418 | 40 |
| 台南市南區 2102 | 122,083 | 23 |
| 台南縣仁德鄉 4127 | 103,801 | 20 |
| 台南縣歸仁鄉 4128 | 37,254 | 7 |
| 高雄縣湖內鄉 4216 | 34,001 | 7 |
| 高雄縣路竹鄉 4215 | 14,174 | 3 |

表 4 97 年度各科別就診人次統計表

| 科別 | 人次 | 百分比 | 科別 | 人次 | 百分比 |
|--------|-------|-------|---------|------|------|
| 心臟內科 | 54263 | 10.50 | 子宮頸抹片 | 4373 | 0.85 |
| 腸胃內科 | 51723 | 10.01 | 神經外科 | 3522 | 0.68 |
| 復健科 | 37980 | 7.35 | 血液腫瘤科 | 3464 | 0.67 |
| 家庭醫學科 | 30638 | 5.93 | 成人健檢 | 3153 | 0.61 |
| 中醫 | 30043 | 5.81 | 預約開刀 | 1994 | 0.39 |
| 內分泌 | 29836 | 5.77 | 整型外科 | 1665 | 0.32 |
| 內科 | 29150 | 5.64 | 心臟血管外科 | 1524 | 0.29 |
| 神經內科 | 28616 | 5.54 | 居家照護 | 1428 | 0.28 |
| 外科 | 26804 | 5.19 | 急診外科門診 | 1204 | 0.23 |
| 骨科 | 21561 | 4.17 | 急診醫學科 | 1183 | 0.23 |
| 身心科 | 16652 | 3.22 | 精神科居家照護 | 787 | 0.15 |
| 泌尿科 | 15870 | 3.07 | 胸腔食道外科 | 647 | 0.13 |
| 內科、老年科 | 15088 | 2.92 | 感染科 | 446 | 0.09 |
| 洗腎 | 14720 | 2.85 | 安寧照護科 | 209 | 0.04 |
| 眼科 | 12043 | 2.33 | 風濕免疫科 | 173 | 0.03 |
| 胸腔內科 | 11912 | 2.31 | 腹膜透析 | 100 | 0.02 |
| 耳鼻喉科 | 10260 | 1.99 | 體外震波 | 62 | 0.01 |
| 體檢 | 10253 | 1.98 | 營養諮詢 | 51 | 0.01 |
| 婦產科 | 9853 | 1.91 | 麻醉科 | 44 | 0.01 |
| 牙科 | 8768 | 1.70 | 愛滋篩檢 | 8 | 0.00 |
| 皮膚科 | 6769 | 1.31 | 出生證明 | 2 | 0.00 |
| 腎臟科 | 6500 | 1.26 | 住院候床 | 1 | 0.00 |
| 小兒科 | 6225 | 1.20 | 放射科 | 1 | 0.00 |
| 其他項目收入 | 5163 | 1.00 | | | |

依據到院病患的特徵，到院就醫的人群多數為周邊地區民眾，且就醫人群性別與就診季節上的無差異，可顯示醫院目前提供的醫療服務可滿足各時段就診的男性或女性病

患，就門診設置而言，該院最多人就診的科別為心臟內科與腸胃內科，較一般外科或內科就診人數為多，與到院的患者偏向為 40 歲以上人群有關(圖 4)。

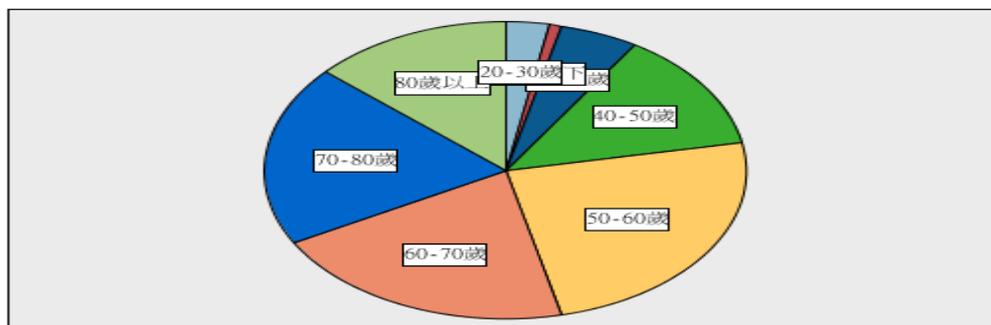


圖 4 心臟內科與腸胃內科就診年齡區分

將科別代碼萃取出心臟內科與腸胃內科後，與就診季節、年齡層、患者性別、戶籍代碼等 4 項資料製作網路圖，依據關聯強度可以發現，腸胃內科病患多為 40-80 歲，且多來自於台南市東區、南區與台南縣仁德鄉(圖 5)，而心臟內科的病患除年齡層提高為 50-80 歲外餘均與腸胃內科相同(圖 6)。

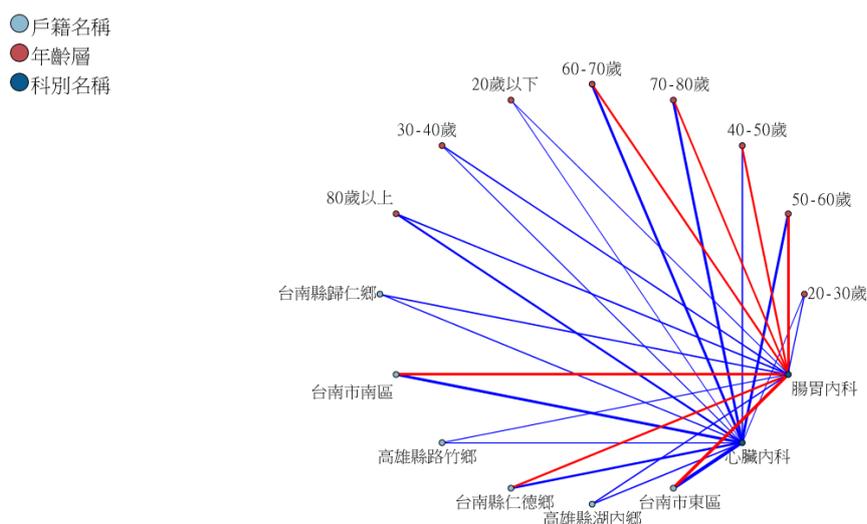


圖 5 以腸胃內科與年齡層和戶籍地關聯圖

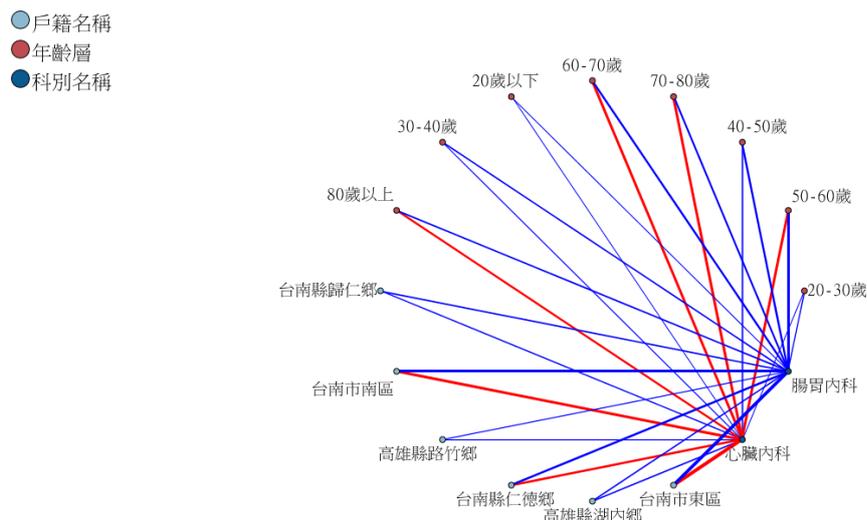


圖 6 以心臟內科與年齡層和戶籍地關聯圖

將科別代碼萃取出心臟內科與腸胃內科後投入後項，就診季節、年齡層、患者性別、戶籍代碼等 4 項資料投入前項，設定最小支持度設定為 1%，最小置信度設定為 80%，以關聯規則 apriori 方式運算，可以獲得 13 項規則(如表 5)，規則項目最多第 13 條規則而言，以台南市東區 30-40 歲的男性民眾(1.12%)有可能看腸胃科(83.29%)。

表 5 Apriori 關聯規則運算之結果

| 後項 | 前項 | 支持度 % | 置信度 % | 規則支持 % |
|-------------|---------------------------------|-------|-------|--------|
| 科別名稱 = 腸胃內科 | 年齡層 = 20-30 歲 | 2.86 | 89.33 | 2.55 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 | 5.31 | 82.41 | 4.38 |
| 科別名稱 = 腸胃內科 | 年齡層 = 20-30 歲 and 患者性別 = 1 | 1.55 | 89.44 | 1.38 |
| 科別名稱 = 腸胃內科 | 年齡層 = 20-30 歲 and 患者性別 = 2 | 1.3 | 89.23 | 1.16 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 戶籍名稱 = 台南縣仁德鄉 | 1.01 | 84.89 | 0.86 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 季節 = 秋季 | 1.42 | 83.69 | 1.19 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 季節 = 春季 | 1.22 | 82.92 | 1.01 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 患者性別 = 1 | 3.09 | 82.75 | 2.55 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 季節 = 冬季 | 1.37 | 82.13 | 1.13 |

| | | | | |
|----------------|---|------|-------|------|
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 戶籍名稱 = 台南市東區 | 1.94 | 81.97 | 1.59 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 患者性別 = 2 | 2.23 | 81.95 | 1.82 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 季節 = 夏季 | 1.31 | 80.84 | 1.06 |
| 科別名稱 = 腸胃內科 | 年齡層 = 30-40 歲 and 戶籍名稱 = 台南市東區 and 患者性別 = 1 | 1.12 | 83.29 | 0.94 |

5. 結論

依據該院 97 年門急診病患資料，就診病患多數為 40 歲以上進入高年齡層的族群，而該院開設的老人門診亦佔總門診量的 2.92%(第 13 順位)，因此可依推測，該院就診病患已受到人口老化的影響，提高就診病患的年齡層；另就醫人口特質上戶籍地多位於醫院周邊，因此，個案醫院拓展服務項目可以置重點於老人醫療服務，亦可以進一步將對象設定為心臟及腸胃內科的患者。

個案醫院要拓展服務可以運用郵寄門診時間表、參與公益活動、提供醫療折扣等方式(潘大永、陳城箴, 2009) 加強病患的忠誠度與維持病患關係，在永續經營的理念下，亦可將拓展目標擴大至 30-40 歲青壯年族群，以辦理衛生教育等公益活動維持醫病關係。

6. 參考文獻

1. David Olson, Yong Shi 著;郭志隆, 張芳菱譯(2008), 資料探勘, 台北市, 麥格羅希爾出版,高立總經銷
2. 廖述賢、溫志皓(2009), 資料採礦與商業智慧, 台北市, 雙葉書廊
3. R. Agrawal, T. Imielinski, and A. Swami, "Database mining: a performance perspective," IEEE Tran. Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, Dec., 1993.
4. 張瑋倫(2008), 顧客關係管理-理論與實務第二版, 台北市, 學貫行銷股份有限公司
5. Winer R.S. "Customer Relationship Management: A Framework, Research Directions, and Future," Haas School of Business, University of California at Berkeley, April 2001
6. 陳尚志. (2009). 以 KANO 二維模式探討大台北地區血液透析病患就醫行為. 臺灣大學. Available from Airiti GCScholar database.
7. 潘大永, & 陳城箴. (2009). 醫院病患資料庫行銷應用分析. 嘉南學報 (人文類) (35), 531-544.

An Application of Association Rule Techniques to Investigate Patients' Characteristics in a Case Hospital

Jo Hsien Tseng¹ Chia Hung Lin² Hui Feng Wu³ Yu-Cheng Chen⁴

¹ Department of Hospital and Health Care Administration, Chia Nan University of Pharmacy and Science and angelar52007@yahoo.com.tw

² Department of Hospital and Health Care Administration, Chia Nan University of Pharmacy and Science and s582580@hotmail.com

³ Department of Hospital and Health Care Administration, Chia Nan University of Pharmacy and Science and k55991@yahoo.com.tw

⁴ Department of Hospital and Health Care Administration, Chia Nan University of Pharmacy and Science and ycchen@mail.chna.edu.tw

Abstract

Physical examination and health management has been becoming gradually important in current years. Using scientific methods and information technology to analyze characteristics of customer relationship is a tendency. To establish an analytical model for understanding properties of customers in a case study is important for extending hospital services. In this study, we took a data filtering after collecting patients' data and applied to the cluster analysis, association rules algorithm and the Apriori algorithm methods to establish "association rules". Results showed that the most popular medical divisions were cardiology and gastrointestinal medicine and these patients were over age 40. Association rules showed that most gastrointestinal medicine patients were 40-80 years old, and cardiology patients were 50-80 years old. Results indicate patients received medical treatment in the case hospital were local residents and the majority of population was the elderly. We recommend that extending services of the hospital should focus on medical services of cardiology and gastrointestinal medicine departments for the elderly residents.

Key word : Data Filtering, Association Rules, Cluster Analysis