運用資料探勘建構共患疾病關聯模型-以痤瘡患者為例

林伊亭[a] 王綺嫻[bc] 蕭嘉士[a] 顏永泰[a] 許怡欣[b] 劉立[ad]

[a]台北醫學大學醫學資訊所研究所 [b]台北醫學大學醫務管理研究所 [c]財團法人恩主公醫院
[d]台北醫學大學附設醫院

sgto5@ms24.hinet.net

摘要

本研究以痤瘡患者為例，運用資料探勘技術建立共同患病的關聯模型。根據研究統計，痤瘡為皮膚科中常見疾病，約佔皮膚科門診人次的20%。而痤瘡患者大多數因共患疾病提高疾病複雜度，導致使用醫療資源的頻率上升，造成醫療資源耗用。為降低耗用，本研究探討痤瘡患者共患疾病的現況，利用資料探勘中的關聯法則分析全民健康保險學術研究資料庫，從中發現痤瘡與共患疾病的隱性關聯。經過歸納資料並分析結果後，針對探勘的結果進行焦點團體座談，了解痤瘡患者共患疾病之臨床病徵及影響共病的相關因素。本研究為共患疾病之探究，醫療領域中屬於預防性醫學一環，其研究結果可輔助分析疾病危險因子，未來可發展於用藥決策系統及整合診斷決策系統，提供臨床醫師診斷及治療方面的參考。

關鍵字：痤瘡、共患疾病、資料探勘

壹、 緒論

在國外學者研究中指出，痤瘡患者同時罹患的共患疾病導致臨床治療的困難度增加，將造成治療結果的差異(Cornoni-Huntley, Foley & Guralnik, 1991)。因此本研究欲建立痤瘡共患疾病之關聯模型，然而針對目前國內痤瘡患者共患疾病的研究仍顯不足，無法對共患疾病有全盤性的深入了解。所以本研究以全民健康保險資料庫的資料為分析目標，利用資料探勘的技術，從龐大的醫療資料與文獻資訊中擷取出有效用的資訊，資料探勘為近年來常運用於醫學資料庫分析的演算法之一，對於協助萃取有效益資訊是極重要的研究方法，以關聯法則進行運算與比對，呈現痤瘡患者常出現的共患疾病種類與組合，並以其關聯程度進行探討與分析。

貳、 文獻探討

2.1. 痤瘡介紹及共患疾病相關研究

痤瘡(Acne)此疾病常發生於九歲至三十五歲人口的疾病，好發於青少年，如10-17歲的女性及14-19歲的男性。痤瘡的形成原因一般認為與雄性荷爾蒙有關，隨著毛囊脂腺分泌增多、毛囊口角化及狹窄，以及痤瘡桿菌增殖菌感染引起粉刺、丘疹、膿胞、結節及囊腫（劉正義,2001)。痤瘡情況嚴重或若未適當治療會產生疤痕，包括：色素型和紅斑型的表淺型痘疤、冰鑿型和車廂型的凹洞型痘疤及蟹足腫。痤瘡發生的部位為皮脂腺分布較密集及數量較多的位置，如：臉部、前胸、臀部及後背。

共患疾病(comorbidity)是指研究對象在罹病的同時或其臨床病程中，除了患有研究的指標疾病之外，共同患有其他臨床的疾病。本研究中所稱的共患疾病，定義為研究對象患有指標疾病的同時或前後一個月內，所罹患的疾病組合，為事件性共患疾病。相關文獻中有提到的痤瘡患者的共病包括：多囊胞性卵巢症候群(polycystic ovary syndrome) (Buggs C &Rosenfield RL ,2005)、SAPHO症候群(SAPHO syndrome) (Iqbal &Kolodney, 2005)，及PAPA 症候群(PAPA syndrome) (Stichweh , Punaro &Pascual , 2005)。

上述症候群中的疾病共同發生的機會較高，致病機轉相近，又因病患可能在不同科別就診，無法獲得整合性的治療，經由大型資料庫的分析，期能更適切的探討其共病現象。

2.2. 大型資料庫探討共患疾病的現況

在國內外學術文獻網站中，以資料探勘(data mining)、共患疾病(comorbidity)，及醫療資訊系統(medical information system, MIS) 為關鍵字交叉組合查詢後，得知相關的文獻數量不多，顯示目前資料探勘及資料庫的分析應用於共患疾病的研究不足。

本研究針對痤瘡患者的共患疾病進行探討，分析一年之內的健保學術研究資料庫。醫療界中大型資料庫的應用，多以關聯法則分析民眾的就醫資料，瞭解疾病之間相互關係及時間序列關係。黃昱銘（2004）研究中指出，醫療資料庫中的診斷記錄存在許多隱藏性資訊規則，找出症狀及疾病之關聯性後即可建立診斷系統輔助就診(黃昱銘,2004)。

2.3. 資料探勘方法及關聯規則演算法介紹

資料探勘技術(Data Mining)是資料探索 (KDD, Knowledge Discovery in Database)之一，主要是從含有巨量資料的大型資料庫中，運用快速的電腦運算萃取有價資訊、關聯過程及隱藏事件。Data Minings已廣乏應用於商業之中，藉由資料探勘可分析顧客區分群集及購物行為，依此設計行銷方式以獲效益。應用於醫學領域中，探勘可將大量的病歷資料進行數據分析，尋找與指標疾病有相關性的疾病，以互相關聯的程度來做為定義共患疾病的參考，輔助醫療診斷。由於醫學診斷領域上要求演算法具備較高的準備性及處理雜亂資料的能力，並能減少演算法測試樣本數，資料探勘技術便是符合此要求的演算法。(吳國禎,2000)

資料探勘處理的方法採用關聯法則。關聯法則由Agrawal等學者提出Apriori演算法，主要用於分析大型商業交易資料庫中，商品項目之中的關聯性。

其描述如下：關聯規則$X \rightarrow Y$： X、Y為交易項目的集合，且 $X \cap Y = \varnothing$。令$I=\{i1, i2, \ldots, im\}$，I為項目(Items)所組成的集合。D 為所有交易的集合，T為一筆交易的集合，*若$T \subseteq X$則稱T交易中包含X集合。*

若$X \subset I$, $Y \subset I$, $X \cap Y = \varnothing$則滿足最小支持度及最小信賴度。

支持度指X此資料項目在資料庫D所佔的比例，如方程式(1)所示。而s(X $\rightarrow$Y)形式為P(X∪Y)表示同時發生X與Y交易項目機率。信賴度指發生某事件的情況下，生另一事作的機率，可視為關聯的強度。如方程式(2)所示。

支持度 support s：

X在D出現次數/ $|D|$ = s % (0<s≦1) (1)

*信賴度*confidence c：

$X \cup Y$ 在D出現的次數/ X在D出現次數= c % ((0<c≦1) (2)

Apriori 演算法的流程可分為兩大階段：於資料庫中挖掘符合或大於訂定支持度的高頻項目集(Large itemsets)後，依此再產生關聯規則並運算。

✓ 高頻項目集 (Large itemsets)

先計算每單一項目在資料庫中出現的次數，若出現的次數大於或等於研究設定的最小支持度(minimum support)，則能藉此決定出高頻項目集合Large I-itemsets (Li)。進行的方法為先針對單一項目是否滿足最小支持度進行合併和刪除，以產生候選項目集合(candidate itemsets)，再從候選項目集合中，針對兩兩組合的項目是否滿足最小支持度進行合併和刪除，以產生較大的候選項目集合，再從此候選項目集合，針對每三個組合的項目是否滿足最小支持度進行合併和刪除，如此反覆進行直到無法產生新的候選項目集合為止。(Agrawal, R., Imielinski, T., &Swami, A. 1993;Ramakrishnan Srikant,Quoc Vu &Rakesh Agrawal,1997)

✓ 產生關聯規則

將每個高頻項目集合中計算出信賴度，若達到研究設定的最小信賴度，則關聯規則成立。設定X與Y是交易項目的集合，令T為一筆交易，資料庫中有c%交易包含X也包含Y，且若支持度大於指定之最小信度(Minimum Condidence)，則關聯規則指X.Y在c%的信賴度下成立。

參、 材料與方法

3.1. 研究對象

本研究欲探討痤瘡患者共患疾病，痤瘡之國際疾病分類碼為 706.0 及 706.1，前者為痘樣痤(Acne varioliformis)，後者為尋常性痤瘡(Other acne)，其分類為型態及嚴重度的不同。資料來源採用國家衛生研究院全民健康保險學術研究資料庫，以 2002 年特定主題分檔之「醫學中心西醫門診及處方治療明細檔」中，四月至九月的就醫資料中，國際疾病分類碼三個欄位中任一欄位為 706.0 及 706.1 的就醫資料，並將病患依年齡分組，進行資料探勘處理。因此曾在 2002 年四月至九月以痤瘡診斷就醫的病患，為本研究的研究對象。這段期間病患資料依 ID 歸戶後共有 18,272,523 人，因痤瘡疾病就醫的人數歸戶後為 541,255 人。

3.2. 研究設計與研究流程

✓ 研究設計

本研究運用資料探勘技術，欲分析出痤瘡共患疾病所進行步驟如下:

■將健保資料庫中將研究對象進行資料萃取的步驟。

■將資料萃取匯入資料庫儲存。

■利用 Index Mininer 資料探勘軟體，將研究分析範圍資料匯入至程式。

■進行資料探勘中關聯組合演算法（Apriori）分析。

其關聯組合演算法的演算參數設定如下：

Support 值 minSupport：0.01 maxSupport:1

Confidence 值 minConfidence:0.6

■程式進行資料分析。

■運用資料匯整技術進行分析結果。

■將分析的結果進行質性研究以求研究之精準透徹。

應用於本研究的資料分析，支持度為以有就醫人數中，同時患有痤瘡及某項特定共患疾病的人數列入計算；信賴度以所有痤瘡病患人數中，同時患有痤瘡及某項特定共患疾病的人數列入計算。因此，由支持度可知痤瘡和某項特定疾病共同發生的普及程度；由信賴度可知在痤瘡病患中，同時罹患痤瘡和某種特定疾病的準確程度。

✓ 研究流程

本研究分為量性及質性兩大部分。量性方面：以資料探勘方法中的關聯法則，分析痤瘡患者因痤瘡就醫前後一個月內之就醫紀錄中，和痤瘡共同出現的疾病診斷以代表痤瘡患者的共患疾病。質性方面：針對探勘的結果，邀請相關科別的專科醫師進行焦點團體座談，以了解臨床實務和探勘結果的相關及影響。使用資料探勘的方法須注意結果是否具臨床意義，仍須從臨床的角度加以驗證。研究流程如Figure 1所示。

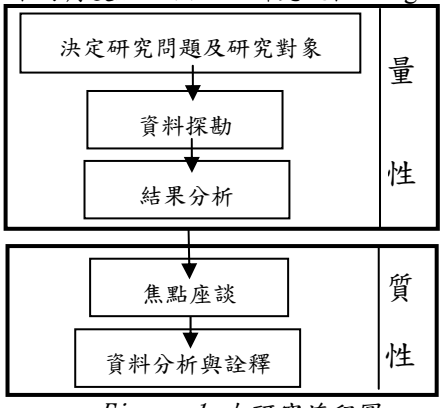

Figure 1 本研究流程圖

3.3. 關聯規則演算與分析

本研究使用Apriori演算法來分析疾病與疾病之間的隱含關係，用於分析資料中會同時出現的資訊，運算各項目與其屬性之間共同發生的關聯。關聯法則有兩個參考依據：包括表示各項目同時發生機率的支持度與表示關聯強度的信賴度，用來衡量計算出的關聯是相有參考效益。

本研究設定minimal support為40%，由疾病資料庫中計算每一疾病代碼出現的次數，共有1.2.3.4.5五種疾病，5×40%＝2，因此出現次數小於2者將刪除，大於者2則納入large itemset 1，再將large itemset 1中的疾病代碼兩兩組合成candidate itemset 2，出現次數小於2的組合加以刪除，出現次數大於2的則列入large itemset 2，依此原則得到large itemset 3接者計算large itemset中關聯規則的信賴度，本研究設定的minimal confidence為100%，因此可以得到3→5，2→5，2.3→5等關聯規則，由於指標疾病為3，因此得到疾病代碼5為指標疾病之共患疾病的規則。

肆、 結果

本研究以 Excel 分析部分資料庫中痤瘡共病的概況，抽樣 Excel 可接受的最大值 65,536 筆就醫資料來進行測試，此人數超過 10%的總就醫人數(完整資料庫)，分析的因子包括各年齡分組的人數及共患疾病總數，可得知痤瘡患者集中於 10-39 歲的年齡，再加上隨著年齡增加，慢性的共患疾病大幅增加。 其資料結果如下：

Table 1 部分痤瘡患者年齡分布及共患疾病數目

Age	Patient number	Percent (%)	共患疾病總數	平均每人共病數(種)
0~9	477	0.7	1577	3.3
10~19	20149	**30.7**	33427	1.7
20~29	28585	**43.6**	53551	1.9
30~39	9916	**15.1**	25570	2.6
40~49	4456	6.8	14527	3.3
50~59	1229	1.9	5492	4.5
60~69	427	0.7	2453	5.7
70~79	239	0.36	1825	7.6
80~89	56	0.08	423	7.6
90~99	2	0.003	17	8.5

擷取年齡為 10-39 歲的三組病人進行詳細的關聯法則分析。以 Indtx Miner 軟體分析此部分的資料，可知隨著年齡增加女性患者的比例逐漸增加。資料分析結果如下：

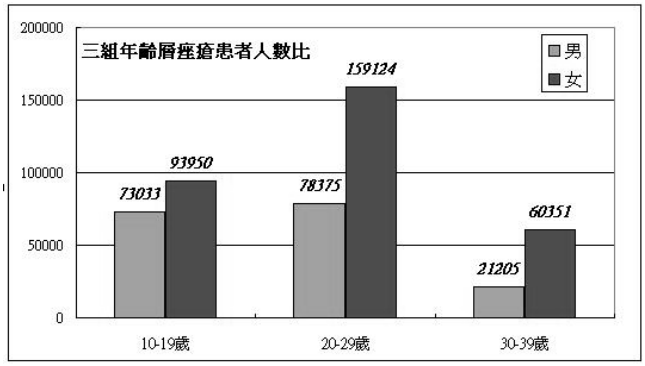

Figure 2 三組年齡層痤瘡患者人數比

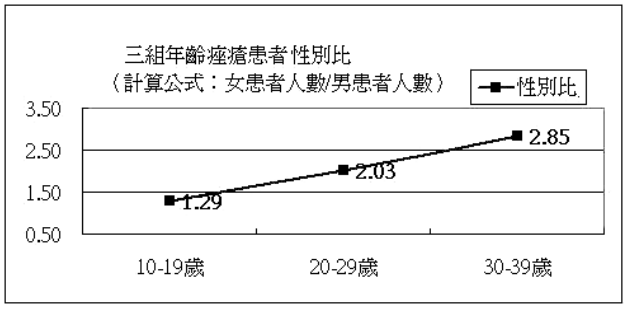

Figure 3 三組年齡層痤瘡性別比例

由資料庫關聯分析得知 10~39 歲痤瘡病患常見共患共疾病分別為上呼吸道感染、濕疹及上呼吸道疾病等，在信賴度為 1 的強度下，支持度大於 0.01。資料結果如 Table 2 所示。

Table 2 10~39 歲痤瘡病患之共患疾病支持度表

(信賴度皆為 1)

**為該年齡層支持度最高的共患疾病*

年齡分佈 / 支持度 / 共患疾病	10-19 歲	20-29 歲	30-39 歲
上呼吸道感染 (Upper respiratory tract infection)	***0.09**	0.06	0.06
濕疹 (eczema)	0.05	***0.07**	***0.08**
上呼吸道疾病 (Diseases of the respiratory system)	0.02	0.02	0.02
急性扁桃腺炎 (Acute tonsilitis)	0.02	0.01	0.01
尋常性疣 (Verruca vulgaris)	0.02	0.01	0.01
急性鼻竇炎 (Acute sinusitis)	0.02	0.01	0.01
近視 (Myopia)	0.01		
蟹足腫等增生性疤痕 （Keloid）		0.01	

伍、 討論

本研究依據關聯法則探勘痤瘡患者共患疾病的結果，進行質性驗證，由焦點團體座談方式討論發生共病的原因並詮譯結果。由於痤瘡患者集中於 10-39 歲的年齡，研究內容指出此階段年齡層都有耳鼻喉科及皮膚科的相關疾病，而在 10-19 歲的組別還有眼科的疾病，因此焦點團體座談邀請耳鼻喉科、皮膚科及眼科共四位臨床專科醫師，共同討論探勘結果。

✓ 耳鼻喉科

在耳鼻喉科的相關疾病方面，上呼吸道感染為常見的疾病，和痤瘡一同發生的原因，有可能因為兩者皆為好發率高的疾病，提高共病的發生機會。亦有文獻報告指出，痤瘡病患接受長達數週的口服抗生素，導致喉嚨發炎、呼吸困難、氣喘等和耳鼻喉科相關的副作用。

✓ 皮膚科

和痤瘡共病的皮膚科其他疾病，包括濕疹、毛囊炎、色素性疾病、尋常性疣、蟹足腫等增生性疤痕及足癬。以濕疹為例，濕疹是皮膚科相當常見的疾病，除了盛行率高及同時就醫的方便性導致和痤瘡共病之外，痤瘡相關的口服及外用治療藥物亦常導致濕疹的發生，包括刺激性皮膚炎及口唇炎等。另外，因治療痤瘡的口服抗生素會導致光敏感等副作用，增加痤瘡與皮膚疾病共病的機率。

✓ 眼科

眼科的近視疾病在 10-19 歲的年齡組別和痤瘡有共病現象，近視是青少年相當常見的疾病，和痤瘡共病的

原因除了盛行率高之外還有兩者發生年齡較為相近。本研究的研究期間為四月至九月，而 10-19 歲學生於新學期初(九月份)會在學校接受相關視力檢查，固有近視的診斷碼比率增加，若研究期間調整，也許不會有近視的診斷碼。在臨床治療方面，痤瘡和近視的治療並無相關影響。

陸、 結論

本研究使用資料探勘技術中的Apriori演算法，產生高頻項目集及關聯規則，探討痤瘡及其共患疾病的隱性關聯，由研究結果得知痤瘡患者的共患疾病多為上呼吸道感染及濕疹等病徵，並邀請痤瘡共病的專家含耳鼻喉科、皮膚科及眼科醫師進行座談，分析痤瘡共病的發生原因，對於預防醫學產生極大的影響。共病的資料探勘研究成果可提供醫師做全盤性的醫療診斷，降低痤瘡病患於不同科別重覆醫療的支出。尚能提供民眾對於痤瘡共患疾病的了解及建立正確的醫療保健觀念。未來可結合醫療團隊，以資料探勘技術探討其它疾病的共病現象，運用在實證醫學上，提高醫療品質。

致謝

感謝台北醫學大學醫學資訊研究所 蔣以仁老師提供資料探勘軟體

柒、 文獻參考

[1]吳國禎(2000)，資料探索在醫學資料庫之應用，未出版碩士論文，桃園，pp10-15。

[2]黃昱銘(2004)，有效率地探勘疾病和病症之複合項關聯規則，未出版碩士論文，台南，pp 29-45。

[3]劉正義(2001)，某醫學中心皮膚科青少年門診常見皮膚疾病的分析，未出版碩士論文，台中。

[4]Agrawal, R., Imielinski, T., and Swami, A. (1993), "Mining Association Rules between Sets of Items in Large Databases," ACM SIGMOD Conference on Management of Data, pp2-5.

[5]Buggs C, Rosenfield RL (2005), "Polycystic ovary syndrome in adolescence," Endocrinol Metab Clin North Am,34(3), pp 677-705.

[6]Feinstein, A. R. (1970). "The pre-therapeutic classification of co-morbidity in chronic disease ,"Journal of chronic diseases. 23, pp 445-468.

[7] Iqbal , Kolodney (2005). "Acne fulminans with synovitis-acne-pustulosis-hyperostosis-osteitis (SAPHO) syndrome treated with infliximab,".Journal of the American Academy of Dermatology. 52,S118-20.

[8] Ramakrishnan Srikant and Quoc Vu and Rakesh Agrawal (1997), " Mining Association Rules with Item Constraints",pp3-5.

[9] Stichweh, Punaro , Pascual (2005), "Dramatic improvement of pyoderma gangrenosum with infliximab in a patient with PAPA syndrome,"Pediatr Dermatol.,22(3), pp 262-265.