## 二、 研究計畫內容（以 10 頁為限）：

### (一) 摘要

在這份專題中，我們將 1996~2013 年之健保資料輸入資料庫，進行疾病關聯性分析。以共病組成為依據設定距離公式，再利用機器學習方法對共病分群，繪製親緣關係圖。

共病群又可分為兩個時期討論，一為發病前，一為發病後。探討前者有助於推測早期病徵，後者有助於預測未來可能誘發的病症。最後將結果呈現於平台，供醫師患者參考。

### (二)研究動機與研究問題

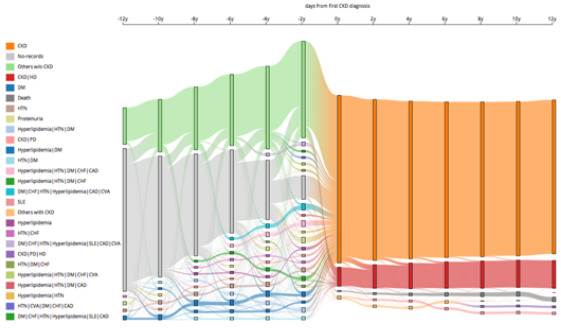
由於數據處理的便利性，醫療及資訊的結合蔚然成風。機器學習於預測系統的應用與日俱增，除了系統優化，也不乏影像判讀、處方/診斷推薦系統等等。健保資料庫的建立為以上提供了良好且大量的資料來源，如何從中獲取有用的訊息，是我覺得十分有價值的探討方向。

別於上述用藥推薦系統，這份專題旨在使用門診處方及治療明細(CD)檔進行共病/併發症分析，實作疾病評估平台。

### (三)文獻回顧與探討

1. A Visual Analysis Approach to Cohort Study of Electronic Patient Records[1]

這份研究建立一互動式平台，由使用者輸入因子(factor)與時期，平台將每個時期的病人根據因子分類，反饋視覺化結果如(圖一)。分類方法為頻率分群法及階層式分群法。觀察疾病在不同時期的佔比可協助判斷因果關係。



(圖一)個案研究:慢性腎臟病

2. Information categorization approach to literary authorship disputes[2] 這份研究以詞頻分析比較文本相似度，並提出距離公式

$$D(T_1,T_2)=\frac{1}{N_{12}}\sum_{k=1}^{N_{12}}|R_1(w_k)-R_2(w_k)|F(w_k)$$

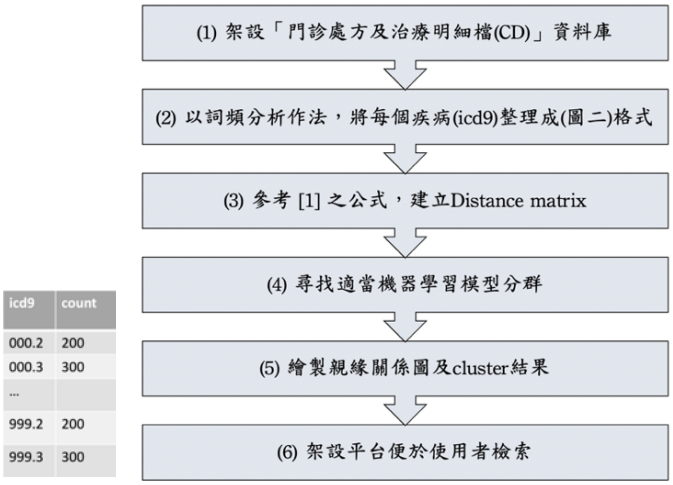
$$F(w_k)=[-p_1(w_k)\log(p_1(w_k))-p_2(w_k)\log(p_2(w_k))]/Z$$

文本 $T_1$ 與 $T_2$ 的距離越小，表示兩者越相似。

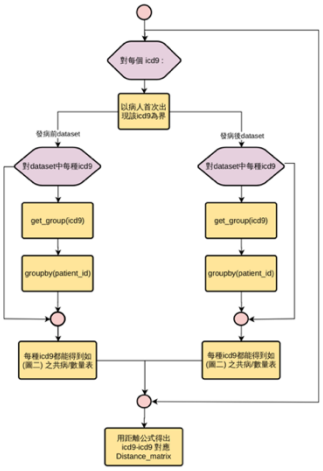
$N_{12}$ 代表疾病 $T_1$ , $T_2$ 共同詞彙的總量，$F(w_k)$代表機率質量函數，為 $w_k$ 在 $T_1$、 $T_2$ 中 Shannon's entropy 的和除以標準化參數 Z。$R_1(w_k)$代表 $w_k$ 在 $T_1$ 中的權重，對於每個詞彙 $w_k$，$w_k$ 在 T 中出現頻率越高，$R(w_k)$越低。

## (四)研究方法及步驟

1. 研究步驟(圖三)




| icd9 | count |
|---|---|
| 000.2 | 200 |
| 000.3 | 300 |
| ... | |
| 999.2 | 200 |
| 999.3 | 300 |




(圖二)　　(圖三)　　(圖四)

2. 研究方法

(1)原始資料處理流程

如上(圖四)所示，得到所有 icd9 碼的(圖二)形式後，即可套用文獻[2]中的公式，兩兩計算距離並記錄到 Distance matrix 中。

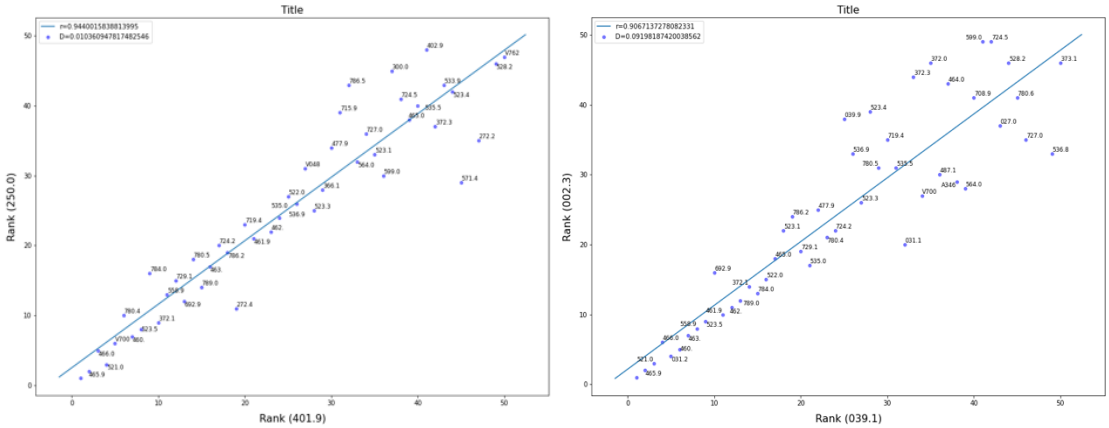
(2) 距離計算

參考公式[2]：$D(T_1,T_2)=\frac{1}{N_{12}}\sum_{k=1}^{N_{12}}|R_1(w_k)-R_2(w_k)|F(w_k)$

$N_{12}$ 代表疾病 $T_1$ , $T_2$ 的共同共病數，$F(w_k)$為機率質量函數。

對於每種共同的共病 $w_k$，$R(w_k)$ 代表 T 中 $w_k$ 的權重，例如 465.9 急性上呼吸道感染(感冒)在 T 中最常見，因此 R(465.9)=1。

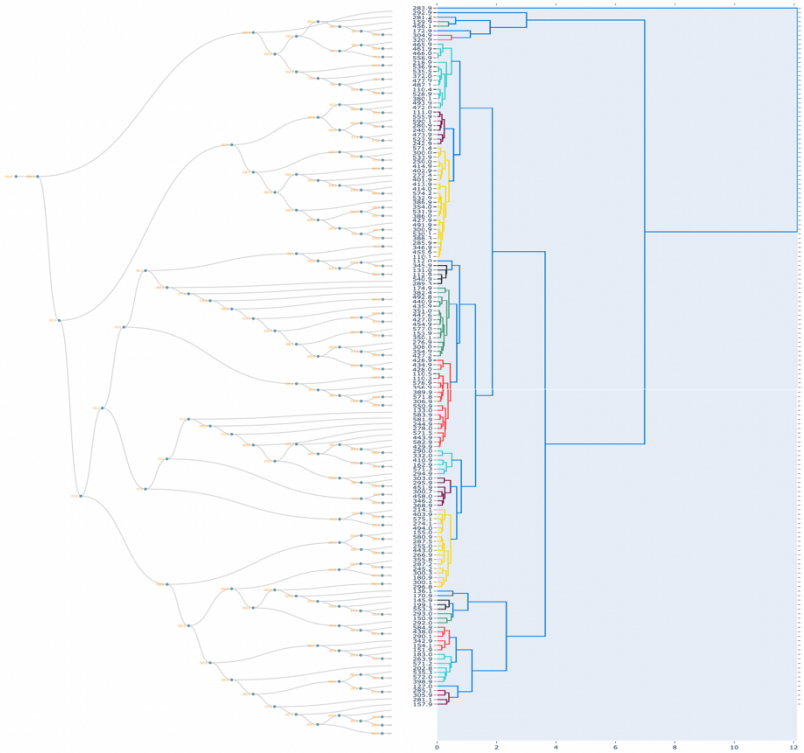
如圖(五)所示，兩疾病的共同共病群，也可以繪製成散佈圖反應相似程度。左圖中 250.0(糖尿病)和 401.9(本態性高血壓)的共病群分佈趨近對角線，且兩者計算結果約為 0.01，表示兩疾病呈高度相關; 右圖中 002.3(C 型副傷寒)和 039.1(放射線菌感染所致肺疾病)距離結果約為 0.09，表示兩疾病關聯性較低。



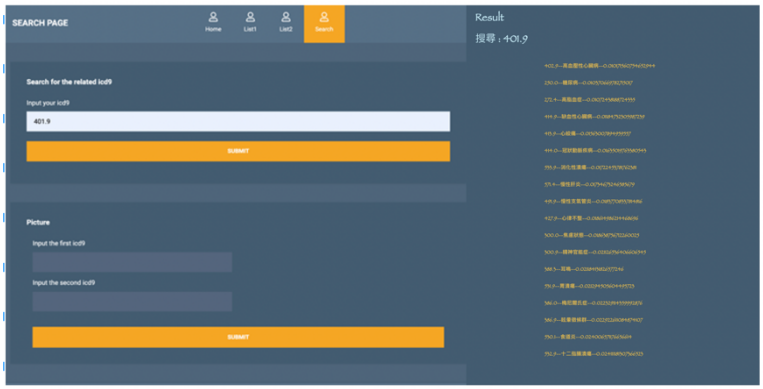

(圖五)

## (五)預期結果

(圖六)內科常見代碼的試做樹狀圖，未來希望能參考其他模型進行更精確的分類分群。並建立互動式查詢平台，讓病患能查詢其他疾病的在未來併發的機率，草圖如(圖七)所示。



(圖六)




(圖七)

## (六)參考文獻

[1] A Visual Analysis Approach to Cohort Study of Electronic Patient Records. Chun-Fu Wang, Jianping Li, Kwan-Liu Ma, Chih-Wei Huang, Yu-Chuan Li. In Proceedings of BIBM 2014; pp. 521-528.

[2] Yang, A. C. C., Peng, C. K., Yien, H. W., & Goldberger, A. L. Information categorization approach to literary authorship disputes. Physica A 2013: Statistical Mechanics and its Applications, 329(3), 473-483.

[3] Yang AC, Hseu SS, Yien HW, Goldberger AL, Peng CK. Linguistic analysis of the human heartbeat using frequency and rank order statistics. Phys Rev Lett. 2003 Mar 14; 90(10):108103.

[4] Profiling phenome-wide associations: a population-based observational study Shabbir Syed-Abdul, Max Moldovan, Phung-Anh Nguyen, Ruslan Enikeev, Wen-Shan Jian, Usman Iqbal, Min-Huei Hsu, Yu-Chuan Li. Journal of the American Medical Informatics Association 2015, 22(4), 896-899.

[5] A probabilistic model for reducing medication errors: A sensitivity analysis using Electronic Health Records data. Huang CY, Nguyen PA, Yang HC, Islam MM, Liang CW, Lee FP, Jack Li YC. Comput Methods Programs Biomed 2019; 170: 31-38.

## (七)需要指導教授指導內容

1. 資料庫空間及健保資料提供
2. 伺服器架設
3. 模型建構指導
4. 理論與實作方法檢討
5. 從醫學與使用者面向評估結果合理與實用性

表 C802