國立台灣大學管理學院資訊管理學系

碩士論文

Department of Information Management

College of Management

National Taiwan University

Master Thesis

健康風險預測

Predicting the Risk of Individual Health

沈書靜

Shu-Jing Shen

指導教授：盧信銘博士

Advisor: Hsin-Ming Lu, Ph.D.

中華民國 104 年 9 月

September, 2015

誌謝

時間匆匆的流逝,研究所的兩年一下就過了,在完成學業的過程中跌跌撞撞,有時覺得難熬,卻也因為身旁的人們伸出手,或推或拉的陪著我前進,讓我終於完成了碩士論文,謝謝幫助過我的所有人,因為有你們,我才能走到此刻的里程碑。

首先,要感謝指導教授盧信銘老師的耐心教導,總是不厭其煩的對我解說許多基本概念,在研究的過程中,也適時給予我指導並幫助我解決遇到的困難,讓我朝著正確的方向前進,進而完成碩士論文。除了盧老師之外,也謝謝口試委員孔令傑老師與曹承礎老師寶貴的建議與指導,使我的論文更加完善。

再來,要感謝實驗室的學長和好夥伴們,給予我的論文及學業許多的幫助,還有貼心的學弟妹們,謝謝你們的支持和鼓勵。感謝親愛的好友們,在奮鬥的日子中與你們互相打氣加油,我才有堅持下去的動力。感謝台大資管系給予我充足的資源與良好的學習環境。謝謝一路上幫助我的每個人!你們都是我的小幸運,這兩年的回憶,不論是開心和難過,都因為有你們而顯得溫暖而閃閃發亮。

最後,給我親愛的家人們,謝謝你們給我精神和實質上的幫助,因為有你們的關心照顧以及滿滿的愛,我才能把學業完成。謝謝上天和土地公爺爺的保佑,讓我遇見這些美好的人們,以及在我最艱難的時刻都給予轉折點,發現柳暗花明又一村。

這兩年的時光中,我面臨人生至今最大的挑戰,一路走來並不順遂,但都踏實的踩著每一步,謝謝大家對我的提攜照顧,你們真的對我很重要!祝福每個人都能度過生命中的困難,而我將帶著你們的祝福,邁向下個旅程。期許未來我也能扮演好協助者的角色,幫助身旁的人。

摘要

現代人對健康的重視逐漸提升，未來健康狀況的預測是現今相當重要的議題，其中，住院為重大的醫療事件，因照護成本高，不論對個人、醫院、全民健康保險都是一項較沉重的負擔，因此，未來是否會住院成為許多人所關心的議題，本研究將聚焦於次年與兩年後的住院率及長期住院率的預測。

過去的研究中，預測未來健康事件時，經常使用合併症指標(Comorbidity index)將病情嚴重性量化，現在主要有以診斷基礎合併症指標和以藥物處方為基礎合併症指標。由於合併症指標能用來表現病情的嚴重程度，因此常被許多研究者作為預測未來健康狀況的特徵值之一。因此，本研究將合併症指標做為預測特徵值，並同時使用兩種合併症指標作為預測特徵值，提升預測表現。除了使用合併症指標當作預測特徵值，也使用機器學習(Machine learning)的方式，例如：特徵值選擇(feature selection)和主題模型(Topic model)，尋找能預測患者未來健康狀況的特徵值。本研究希望透過不同的特徵值及模型，改善未來健康風險的預測。

從實驗結果來看，本研究所使用的特徵值皆具有預測未來住院的能力，而結合不同的合併症指標，比單獨使用其中一種的表現來得佳。另外，透過機器學習方法找出的特徵值，在預測未來住院的表現上，比合併症指標表現來得好，未來除了使用合併症指標預測未來健康風險，也能參考機器學習方法找出的特徵值，改善預測結果。

關鍵字：健康風險、住院預測、合併症指標、特徵值選取、主題模型

Abstract

With increasing attention on individual health, the prediction of individual health risk is an important issue today. One of the important health risks is the hospitalization. The hospitalization is costly for individual, society and health insurance, and many people concern this issue. Therefore, this study focuses on predicting subsequent-year and subsequent-2year hospitalizations.

In the past, many research use comorbidity index to quantify the diseases and present the risk of individual health. They predict the risk of individual health by evaluating one's comorbidity score. For this reason, we use comorbidity as our feature to predict future hospitalization. We also combine different comorbidity index to improve performance. In addition, we use machine learning method, such as feature selection and topic model to find the factors which affect individual's future health status. Our research expects to improve predicting performance of the risk of individual health by using different feature combination and model.

Our results show that both comorbidity index and the feature found by machine learning methods have good performance. In addition, the performance predicted by the feature found by machine learning methods is better than comorbidity index.

Keywords：Individual Health Risk、hospitalization prediction、comorbidity index、feature selection、Topic model

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景與動機

未來健康狀況的預測是現今相當重要的議題，對於個人而言，若能準確預測未來健康狀況，可以提升自身對健康狀況的警覺性，並能嘗試改善目前健康狀況，預防風險的發生，此外也能事先做好未來健康醫療規劃，例如：籌措未來的醫療費用、投保相關的醫療險等；對醫院或政府來說，若能掌握總體的未來健康狀況，可以作為未來醫療資源配置、醫療成本估算與預算編列，以及未來醫療健康照護與衛生安全相關政策的參考。

在眾多醫療事件之中，住院屬於相當重要的健康事件之一，因住院的高醫療照護成本，不論是對個人、醫院、全民健康保險都是一項較沉重的負擔(Klaus W. Lemke et al.，2012)，而長期住院的成本更是所費不貲，因此，住院的醫療健康照護成為許多人所關心的議題，也讓住院率的預測也更為重要，本研究將聚焦於次年與兩年後的住院率及長期住院率的預測。

過去的研究中，預測未來健康事件時，經常透過合併症(Comorbidity)了解個體的基本健康狀況，合併症的定義為：個體除了主要疾病外，同時患有一種或一種以上的疾病。為了使用合併症衡量病患的健康狀況，發展了合併症指標(Comorbidity index)，合併症指標將病情嚴重性量化，依照疾病的嚴重程度分類，並予以不同分數，嚴重程度越高，分數越高，最後加總分數，當作病患健康的風險因子，現在文獻主要有以診斷基礎合併症指標和以藥物處方為基礎合併症指標。由於合併症指標能用來表現病情的嚴重程度，因此常被許多研究者作為預測未來健康狀況的特徵值之一，過去的研究許多都著重於發展診斷基礎合併症指標和以藥物處方為基礎合併症指標，比較此兩種指標的預測表現，找到較有預測力的特

徵值(Yaa-Hui Dong et al.,2013)。本研究除了分別使用以診斷基礎合併症指標和以藥物處方為基礎合併症指標外，也結合兩種合併症指標做為預測特徵值，期望透過同時考慮兩種合併症指標提升預測表現。

合併症指標是以整合的方式，將疾病依嚴重程度分類給予分數值，最後加總分數做為風險因子，但在大量的資料中，可能只有部分診斷碼或藥物代碼是與未來住院相關的。因此，本研究除了使用合併症指標當作預測特徵值，也使用機器學習(Machine learning)的方式，例如：特徵值選擇(feature selection)和主題模型(Topic model)，尋找能預測患者未來健康狀況的特徵值。

本研究對患者的診斷及用藥資料做特徵值選取(feature selection)，計算資訊獲利(Information gain)與卡方檢定(chi-square)，從病歷記錄中找出與住院相關的特徵值。此外，也結合主題模型，將患者過去一年的診斷及用藥紀錄整合視為一篇文章 (Document)，使用 Latent Dirichlet Allocation (簡稱LDA)從中萃取出主題，期望透過潛藏患者健康狀況的主題預測未來健康風險，提升預測未來健康狀況的能力。

研究資料的使用上選擇台灣全民健康保險研究資料庫，因行政申報資料內容包含各區域、人口群等，且取得成本較低，過去許多研究將行政申報資料作為主要醫療服務研究資料。而台灣的健保納保率高，行政申報的資料含有用藥、診斷、住院等完整的醫療相關資料，因此本研究使用台灣全民健康保險研究資料庫做為主要研究資料，透過次級資料測量個體健康狀況。

本研究期望透過使用了醫療領域的合併症及機器學習方法找出的特徵值，並嘗試不同的特徵值組合，找出能提升未來健康狀況預測能力的特徵值，提升預測能力，讓預測結果能做為患者及醫療相關單位進行未來規劃的參考。

1.2 研究目的

基於上述的背景與動機，本研究透過患者的醫療紀錄，產生住院預測的特徵值，預測未來是否住院及長期住院的可能性。在資料的部分，利用台灣全民健康保險研究資料庫進行分析，期望透過合併症指標、特徵值選取、主題模型 (topic model)生成的特徵值等，找出具有預測未來健康狀況能力的特徵值，增進預測未來健康狀況的能力，亦即改善住院率的預測能力，獲得較好的預測結果。

因此，主要探討的問題有以下幾點：

- 結合以診斷為基礎之合併症指標和以藥物處方為基礎之合併症指標，是否有較佳預測未來住院的能力？
- 探討機器學習方式找出的特徵值是否有預測未來住院的能力？
- 探討不同模型的表現是否有顯著差異性？

第二章　文獻探討

未來健康狀況是許多人所關心的事，台灣 1995 年開始實行全民健康保險，納保率達 99%以上，國民的診斷、用藥、醫療費用申報等皆儲存於其中，大量的行政資料取得成本低，且具有高分析價值，因此，本研究使用健保資料庫進行，希望藉由將主題模型 (topic model)萃取出的健康狀況 (health condition)加上合併症指標一起作為預測特徵值，提升次年住院率預測表現。透過潛藏狄利克雷分配(Latent Dirichlet allocation，簡稱 LDA)模型，將診斷與用藥資料類比成文章(Document)，健康狀況類比於主題 (Topic)，找出健康狀況的機率分佈，與合併症指標結合加入預測模型中預測次年住院率，期望住院率預測的準確度能更進一步提升。

以下將未來健康狀況預測研究分為三部分：合併症測量及相關研究、特徵值選取、模型方法。

2.1 合併症測量及相關研究

Klaus W. Lemke et al. (2012)提到，對於保險和社會，住院需要高醫療照護成本，若能預測住院風險，能幫助醫療資源規劃及照護管理，因此，許多研究針對住院率進行預測，而過去許多的研究使用合併症指標 (comorbidity index)預測未來健康狀況，主要有以診斷基礎合併症指標和以藥物處方為基礎合併症指標，多數研究著重於發展此兩種合併症指標及探討此兩種指標的預測表現，藉此找到較有預測力的特徵值。

本節將先分別介紹以診斷為基礎的合併症指標及以藥物處方為基礎合併症指標，再介紹使用此兩種指標最為預測特徵值的相關研究。

查爾森合併症指標（Charlson comorbidity index，簡稱 CCI）為使用的以診斷為基礎的合併症，為 Charlson et al. (1987) 以 1984 年紐約州立醫院的患者為研究對象，採病歷回顧的資料蒐集方式，研究一個月內住院的 559 位內科病人，使用存活分析方法探討合併症與一年內死亡的關係，最後依相對危險性，分別給予 19 類合併症 1、2、3、6 的權重值，依患者擁有的合併症類別，將合併症權重累加做為 CCI 總分。之後 CCI 也發展出許多不同的版本，例如：Deyo、Romano、D'Hoore，國內較常採用的方法為 Deyo et al. (1992) 所發展的版本，CCI/Deyo 使用 icd9-9-CM 定義原始 CCI 的疾病類別，在合併症類別和名稱的部分，將原版本的「惡性腫瘤」、「白血病」、「淋巴癌」等三類，合併成「惡性腫瘤，包含白血病與淋巴癌」，修正後 CCI/Deyo 共包含 17 類合併症。

常見的以診斷為基礎合併症指標除了 CCI 之外， Elixhauser 合併症指標 (Anne Elixhauser et al.，1998)為 Elixhauser 等人於 1992 年使用加州急性醫院 1,779,167 位病人發展出的合併症指標，共有 30 類合併症，此合併症的研究排除 18 歲以下的病人，除了 CCI 原有的合併症種類，還增加了肥胖、體重下降、嚴重精神病、酒精濫用、藥物濫用、憂鬱症等合併症，此合併症指標與住院天數、住院費用、院內死亡相關。

Yu-Tseng Chu (2010)的研究中，利用台灣健康保險資料，使用三種不同的診斷基礎合併症指標當預測特徵值：Elixhauser (Anne Elixhauseret al.，1998)、CCI/Deyo、CCI/Romano，預測急性心肌梗塞 (Acute Myocardial Infarction，簡稱 AMI)和慢性阻塞性肺病 (Chronic Obstructive Pulmonary Disease，簡稱 COPD)的短期與長期死亡率。Yu-Tseng Chu (2010)的研究結果為使用 Elixhauser 合併症指標當特徵值的表現比 CCI/Deyo、CCI/Romano 來得好，此研究驗證了 Elixhauser 合併症指標對短期和長期死亡率皆有足夠的預測能力。

在以藥物處方為基礎合併症指標的部分，Michael Von Korff(1992)提出慢性疾病分數 (Chronic Disease Score，簡稱 CDS)，藉由門診病人過去一年所使用的門診用藥資料，判斷患者的慢性病狀況，是美國第一個以藥物處方為基礎的合併症衡量方法，最初的版本只包含 17 類疾病，透過控制年齡、性別、看診次數，CDS 對次年住院和死亡具有預測能力。Kathleen G. Putnam et al.(2002)使用 CDS 作為預測住院的特徵值之一，驗證 CDS 具有預測住院率的能力，能作為預測的指標之一。

Yaa-Hui Dong et al. (2013) 發展以藥物處方為基礎的合併症指標 (Pharmacy-Based Disease Indicator，簡稱 PBDI)，該研究根據 CDS 的框架與解剖治療化學 (Anatomical Therapeutic Chemical，簡稱 ATC)分類系統發展 PBDI。因 CDS 是依據美國的疾病及用藥情形所制定，然而疾病與用藥會因為地理區域與種族而有所差異，因此，該研究依照台灣地區主要疾病調整 CDS 的合併症種類，例如：加入南亞常見的 B 型肝炎病毒感染，最後制定出 37 類對應台灣主要疾病的藥物類別。該研究透過比較 PBDI 和 CCI/Deyo 兩種不同類型的合併症指標預測未來住院的能力，評估 PBDI 是否能當作預測未來健康風險的特徵值。在模型的部分，使用羅吉斯回歸進行預測，資料集使用 2005 年與 2006 年健保資料，兩組對照的參數分別為 PBDI 分數、年齡、性別與 CCI/Deyo 分數、年齡、性別。預測次年住院的表現為，在 2005 年，PBDI 模型的 AUC 為 0.72，優於 Deyo 模型的 0.69；2006 年與 2005 年結果相同，證實此研究發展的 PBDI 在預測次年住院率的表現優於診斷為基礎的合併症指標。

Joshph P. Parker (2003)的研究中，探討藥物資料是否能改善再入院的機率預測，此研究比較 CDS 和 CCI/Deyo 預測再入院的能力，結果顯示兩種合併症指標對於再入院機率的預測都有不錯的表現，此外，將 CDS 加入 CCI/Deyo 的基準線模型，結合兩者合併症指標共同預測，比單獨使用其中一種合併症指標有更好的預測表現。

從合併症指標相關研究來看，可發現以藥物處方為基礎的合併症指標及以診斷為基礎的合併症指標對於未來健康風險都有足夠的預測能力，因此可作為預測未來健康狀況的特徵值。

2.2 特徵值選取

資訊科技的進步，讓大量的電子健康記錄能加以應用分析。Saria et al. (2010)從出院總結報告使用語言模型，找出語言特徵值 (Language feature)中與結果正相關、負相關與無關的字詞當作特徵值，結合結構化資料如：藥物、診療資料等，預測早產嬰兒是否會於院內發生併發症，預測結果 F1 值為 88.3，顯示非結構的文字型態診斷資料，具有提升未來健康狀況的預測能力。

加護病房(Intensive Care Units)內的患者通常病情嚴重，需要較多的照護，若能找出影響病人的相關因子，有助於提升加護病房的照護品質，Marzyeh Ghassemi et al. (2014)使用診斷紀錄當作預測特徵值之一，預測加護病房的死亡率。診斷記錄與病人的病情相關，但過去的預測研究多使用結構化的特徵值，例如：年齡、性別、衡量疾病嚴重程度指標 (例如：SAPS、SOFA)等，非結構化的文字記錄鮮少使用，該研究使用 LDA 從診斷紀錄萃取主題 (Topic)分佈，推論出 50 個主題，將診斷紀錄轉換成可分析的結構化特徵值，並使用主題當作預測特徵值之一，預測加護病房的死亡率，包含：醫院內死亡、出院後 30 天內死亡及出院後一年內死亡，此外計算 50 個主題的加護病房死亡率，推論出個別主題的特性，找出與死亡或存活相關的主題。最後，研究結果顯示，透過 LDA 擷取文字診斷紀錄的主題面向資訊並結合結構化的特徵值，能提升死亡率預測能力。

LDA 模型的架構為一完整的生成模型(generative model)和主題模型，主題模型假設每篇文件可能包含一個或多個主題，且每個字詞可能在不同的主題出現，因此，各主題內的字詞有其機率分佈(probability distribution)。LDA 假設每一篇

文件是由多個主題組成，不同的主題在文件中比重也不同，因此每篇文件有自己的主題機率分佈，而在不同的主題下出現，可能會有相同的字詞出現，但這些字詞的比重在不同的主題下也有所不同，所以每個主題下有特定的詞彙機率分佈。每一個文件都是由主題機率分佈、字詞機率分佈決定文件的內容。

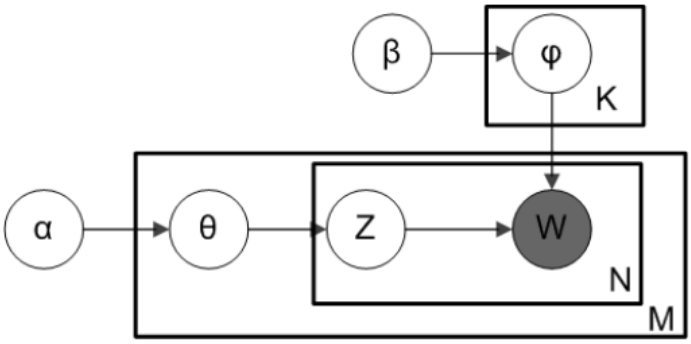


圖 2.2.1 LDA 模型架構圖(http://en.wikipedia.org/)

圖 2.2.1 為 LDA 模型的架構。K 為主題數量，M 為文件總數，N 為文章長度，W 向量為組成文集(corpus)的所有詞彙，Z 向量為文件中的詞彙對應到的主題，例如：$Z_{m,n}$ 表在第 m 篇文章中的第 n 個詞所對應的主題，因此 Z 向量的長度為 N，α 與 β 皆為狄利克雷先驗參數，其中 α 控制每個文件下主題的分佈，θ 為文件的主題先驗機率分佈向量，向量長度為主題數量 K；而 β 控制每個主題下字詞的分佈，φ 為主題中的詞彙先驗機率分佈向量，詞彙w_i在第 k 個主題下為 $\mathrm{P}(w_i|z_i = \mathrm{k})$。而由所有文件組成的文集生成機率(generative probability)為：

$$P(W,Z,\theta,\varphi|\alpha,\beta) = \prod_{k=1}^{K} P(\varphi^{(k)}|\beta) \prod_{d=1}^{D} P(\theta^{(d)}|\alpha) \prod_{i=1}^{N_d} P(z_i{}^{(d)}|\theta^{(d)})\, P(w_i{}^{(d)}|\varphi^{(z_i{}^{(d)})})$$

(公式 2.2.1)

我們所要推算的目標是後驗(posterior)機率$P(Z|W)$，但 θ 和 φ 的估計牽涉到潛藏變數，使得$P(Z|W)$不易做到精確估算，目前較多使用概似估計的方式來估算潛藏變數，例如：變形概似估計(Variational Approximation)、馬可夫鏈蒙特卡羅法(Markov chain Monte Carlo)等方法(Blei et al.,2003)。

在許多的研究中，透過吉布斯抽樣(Gibbs Sampling)推算 LDA 的參數θ和φ，吉布斯抽樣法是馬可夫鏈蒙特卡羅法的一種實作，在母體機率分布未知，但個別樣本機率已知的情況下，對主題變數進行抽樣建構馬可夫鏈，在迭代之後樣本分佈會逐漸收斂至接近母體的分佈。

因此，假設有一些文件希望用一些主題來描述，並用一些字詞描述文件的主題，可以使用吉布斯抽樣完成 LDA 學習，簡單描述學習的流程：在決定主題的數量 K 之後，首先歷經所有文件的詞彙，初始給訂每個詞彙一個隨機主題$z^{(0)}$，接著每一次會排除目前位置的詞彙，計算兩件事：第一，計算所有文件中，詞彙 w 在主題 z 中所出現的次數，也就是文集中屬於主題 z 的詞彙 w 比例；第二，在文件 d 中，屬於主題 z 的詞彙出現次數，亦等於文件中各主題的比例。-i 表示去除位置 i 的詞彙，計算除了位置 i 之外的上述兩件事後，可得到詞彙 w 在文件 d 中被指定成主題 k 機率，可表示成$P(z_i = k|z_{-i}, d, w)$。當計算完目前詞彙屬於各主題 z 的機率後，透過此分佈為詞彙 w 隨機抽取新的主題$z^{(1)}$，之後每個詞都重複使用相同方式迭代更新下個詞彙的主題$z^{(1)}$，直到所有文件的主題分佈θ和所有主題下的詞彙分佈φ收斂為止，輸出在主題下詞彙機率分佈矩陣φ和文件中主題分佈矩陣θ。

目標的機率$P(Z|W)=\frac{P(W,Z)}{P(W)} \propto P(W,Z)$，而$P(W,Z) = P(W|Z)P(Z)$，分別將前項$P(W|Z)$對$\varphi$積分和後項$P(Z)$對$\theta$積分分開積分，透過前述的方式完成吉布斯抽樣，使得馬可夫鏈的主題變數逐漸趨於穩定最後收斂，藉此估計φ和θ參數：

$$\widehat{\varphi^{(k)}} = \frac{n_{(.)\ (k)}^{(.)\ (v)} + \beta}{n_{(.)\ (k)}^{(.)\ (.)} + V\beta} \qquad \text{(公式 2.2.2)}$$

$$\widehat{\theta^{(d)}} = \frac{n_{(.)\ (k)}^{(.)\ (v)} + \alpha}{n_{(.)\ (.)}^{(d)\ (.)} + K\alpha} \quad \text{(公式 2.2.3)}$$

在上述對φ和θ的估計式中，n 表示計數，左上標表示文集範圍，(.)表全部範圍，即為整個文集；左下標表在文件內的位置，若(.)表示文件中所有的位置，若為(- i)表排除文件中第 i 個位置；右上標為要記數的詞彙，若為(.)表示將所有的詞彙納入計數，若為(v)表計算詞彙 v 的數量；右下標為要計數的主題，若為(.)表示將所有的主題納入計數，若為(k)表計算主題 k 的數量，因此，$n_{(.)\ (k)}^{(.)\ (v)}$為計算詞彙 v 在主題 k 出現的次數。完成 LDA 的學習後，得到每個文件的主題分佈以及文件中的字詞分佈，將這些生成的主題作為預測特徵值，用於之後的預測。

除了從記錄中尋找預測未來健康狀況的特徵值外，患者住院的即時資訊，是預測短期病情變化的重要特徵值，如同前述提到加護病房內的病人因病情嚴重，因此，需要較多的醫療照護，Yi Maoet al.(2012)研究說明即時偵測能提早得知病情惡化的資訊，避免患者病情更加惡化，對於加護病房或一般病房的病患都十分重要。若能事先預測嚴重醫療事件的發生，就能增加救治病人的機會，該論文提出一個整合性的資料探勘方法，使用無線感測裝置收集病患的即時重要資訊：心跳率 (heart rate)和血氧飽和度比率 (oxygen saturation rate)做為特徵值之一，用於預測病情惡化，即時對醫療人員送出警示，提升醫院內醫療照護的品質。

2.3 模型方法

本研究中使用以下四個分類模型預測次年住院率：支持向量機 (Support Vector Machine，簡稱：SVM)、羅吉斯回歸 (Logistic Regression)、隨機森林 (random forest)、類神經網路 (neural-net)。

2.3.1 支持向量機 (Support Vector Machine，簡稱：SVM)

SVM 為一種監督式學習，是常見的分類方法之一，廣泛的應用於統計分類與回歸分析上。SVM 利用支持向量(support vector)和邊界(margin)找尋超平面，透過超平面分割資料，SVM 分類器的特性為找到最佳化的超平面，使總誤差最小化並最大化幾何邊。線性資料和非線性的資料皆能透過超平面去分割，SVM 能將原始資料透過非線性轉換放置到較高的維度，在高維度中尋找最佳化的超平面。

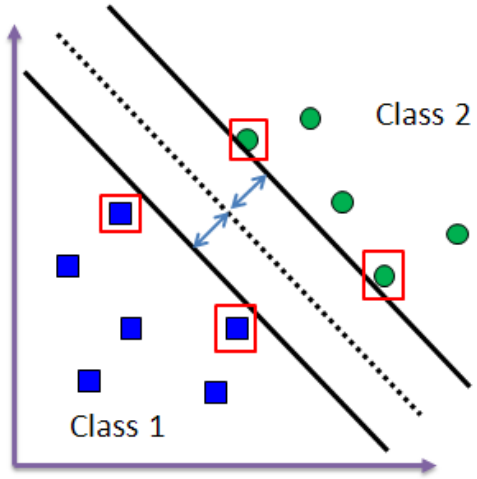


圖 2.3.1.1 最小邊界示意圖

2.3.2 羅吉斯回歸模型 (Logistic Regression)

羅吉斯回歸為預測事件機率的模型，適用於分析應變數為二元的機率統計方法。於次年住院率的預測中，藉由羅吉斯回歸產出 (Output)的次年住院機率，可以看出患者病情的嚴重程度。

Yaa-Hui Donget al. (2013)研究以藥物處方為基礎的合併症指標，即是使用羅吉斯回歸做為預測次年住院機率的模型，比較人口統計參數、人口統計參數搭配

CCI/Deyo、人口統計參數搭配 PBDI 三組特徵值的預測表現。公式如下所示，β^T為模型回歸參數，X為模型輸入參數：

$$\frac{e^{\left(\beta^T X\right)}}{\left[1+e^{\left(\beta^T X\right)}\right]} \quad \text{(公式 2.2.4)}$$

2.3.3 隨機森林 (random forest)

隨機森林為多決策樹的分類器，為一種組合學習 (assemble learning)方法，利用重複抽樣 (resampling)建立多棵決策樹，最後使用投票 (voting)的方式，選擇次數最多的類別當作結果。

隨機森林透過以下的演算法建立每顆決策樹：

1. 定義 N 為訓練用樣本數目，M 為特徵值數目。
2. 輸入：使用 m 個特徵值，用於決策樹節點的結果判定，m < M。
3. 從訓練用的 N 個樣本中，重複取樣 N 次，形成一個樣本數為 N 的訓練集。
4. 對於每個節點，隨機選 m 個特徵值，決策樹上每個節點基於這些特徵值決策結果。
5. 每顆決策樹都不會剪枝(Pruning)。

2.3.4 類神經網路 (neural-net)

類神經網路為模仿生物神經元網路排序及訊號傳遞的數學模型，可解決非線性複雜型的架構，透過反覆學習修正權重達到高準確率。

類神經網路架構可以分成輸入層(Input layer)、隱藏層(Hidden layer)、輸出

層(Output layer)，輸入層和輸出層的數目皆為 1，隱藏層的數量則不一定，並位於輸入層和輸出層的中間，因未和外界接觸而稱為隱藏層，中間處理過程是個黑盒子。倒傳遞類神經網路（back propagation）是目前類神經網路中較為普遍使用的一種，架構一樣為輸入層、隱藏層、輸出層，其中隱藏層的數量不一定，也可以沒有隱藏層。在隱藏層之間及隱藏層和輸出層之間，通常使用非線性轉換。透過計算均方誤差（mean square error，簡稱 MSE），比較輸出層的目標向量與輸出向量的差距，調整權重，達到最小化 MSE。

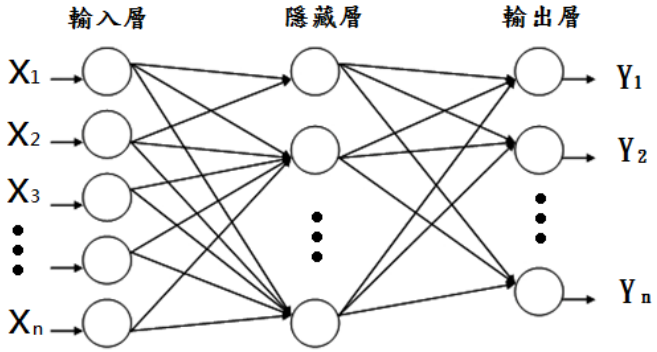


圖 2.3.4 類神經網路架構圖

2.4 小結

由上面幾個小節可以看出，目前已有許多未來健康狀況預測相關研究，除了以醫學上常使用的合併症指標作為特徵值，也有研究也從非結構性的文字資料中，萃取出具有預測性的特徵值幫助預測。

針對次年住院率的預測，許多研究持續發展以合併症指標作為預測特徵值，且預測表現不斷進步，本研究希望使用機器學習方法找出與住院相關的特徵值，並導入主題模型的概念，透過 LDA 模型萃取出的相關主題，找出與次年住院率相關的特徵值，進一步的提升次年住院率的預測表現。

第三章　資料及方法

本章將描述實驗中所使用的資料集，並說明實驗中使用的特徵值與預測模型，幫助理解實驗的進行。患者未來是否住院與其目前的身體健康狀況有關，透過患者的診斷及用藥資料，可以推斷患者目前的身體健康狀況。因此，本研究將整合患者的診斷與用藥資訊，產生相關的預測特徵值，進行次年及兩年後的住院預測。

3.1 資料

本實驗使用全民健康保險研究資料庫 (National Health Insurance Research Database，簡稱 NHIRD)，NHIRD 為供研究使用的健保資料，由健保署將前一年健保資料篩選出，並將敏感性欄位加密，再交由國衛院製成全民健康保險研究資料庫與相關加值資料。

全民健康保險研究資料庫的資料內容包含：門診病患及住院病患的承保資料、診斷及處方用藥資料等，診斷資料編碼使用國際疾病分類碼第九版(全民健康保險研究資料庫，簡稱：ICD-9-CM)，藥物資料使用全民健康保險自定義的編碼，於本研究中，在資料處理時會將此編碼對應到 WHO 的 ATC 編碼，對應表來自健保局網站。

本實驗使用 2005 年至 2008 年的資料，預測次年及兩年後住院率。使用的資料集包含：承保資料檔（ID）、住院醫療費用清單明細檔（DD）、住院醫療費用醫令清單明細檔(DO)、門診處方及治療明細檔(CD)、門診處方醫令明細檔(OO)、特約藥局處方及調劑明細檔（GD）、特約藥局處方醫令檔（GO）等檔案。

下表為實驗及預測主要使用的資料表及欄位：

表 3.1.1 實驗及預測主要使用的資料表及欄位

資料表及年度	資料欄位	目的
承保資料檔（**ID2005~ID2007**）	ID_BIRTHDAY ID_SEX ID_IN_DATE ID_OUT_DATE	判別當年度是否為合格的投保人及取得患者基本資料，包含年齡、性別
住院醫療費用清單明細檔（**DD2005~DD2008**）	ICD_9CM_CODE ICD_9CM_CODE_1 ICD_9CM_CODE_2 ICD_9CM_CODE_3 ICD_9CM_CODE_4	取得患者的診斷資料
	E_BED_DAY S_BED_DAY	計算患者的住院天數，用以判斷患者是否為長期住院
門診處方及治療明細檔（**CD2005~CD2007**）	A_CODE_ICD9_1 A_CODE_ICD9_2 A_CODE_ICD9_3	取得患者的診斷資料
門診處方醫令明細檔（**OO2005~OO2007**）	DRUG_NO	取得患者的用藥資料
特約藥局處方及調劑明細檔（**GD2005~GD2007**）	DRUG_NO	取得患者的用藥資料
特約藥局處方醫令檔（**GO2005~GO2007**）	DRUG_NO	取得患者的用藥資料

3.2 前處理

本實驗在進行前，會做一些前處理，篩選所要的資料和建立特徵值。在合併症指標的特徵值部分，使用的合併症指標為 CCI/ Deyo 與 PBDI(Yaa-Hui Dong et al.，2013)，分別將診斷碼與藥物代碼依對應到 17 類 CCI/Deyo 和 37 類 PBDI 的合併症指標。在機器學習的特徵值部分，利用患者記錄中的診斷碼及藥物代碼，計算資訊獲利和卡方檢定，找出對住院有影響力的特徵值，並嘗試放入不同數量的特徵值至模型中，觀察模型的預測表現，找出讓模型預測能力最佳的特徵值數量。同時，也將患者過去一年的診斷紀錄及用藥紀錄整合成文章形式，使用 LDA 找出潛藏的健康風險主題。

在前處理，主要可以分成以下幾個部分：

3.2.1 投保者基本資料處理

- 實驗個體篩選：根據文獻(Yaa-Hui Dong et al.，2013)所述的兩個條件，透過 ID 檔篩選出合格的實驗個體：

  (1). 在當年度滿 18 歲，且次年度於健保至少投保一天

  (2). 需符合持續投保，如有退保應於 60 天內重新加保

  第一個條件，可透過投保者的生日(ID_BIRTHDAY)計算出年齡，再進行篩選；第二個條件，需透過加退保記錄篩選，可由當年度的加退保日期(IN_DATE 和 OUT_DATE) 及次年的加退保日期(IN_DATE 和 OUT_DATE) 判斷是否有持續投保。最後，保留合格實驗個體的年齡、性別等基本資料。

- 住院記錄：

  透過 DD 檔找出年度內的住院患者，將當年度合格個體資料和次年度的住院檔案進行串檔，並計算患者的住院天數，若患者住院天數大於所有住院記錄的平均住院天數，則判定為長期住院。

- 取得患者的診斷資料及用藥資料：

  需將 CD、OO，DD、DO， GD、GO，分別串檔成 CDOO、DDDO、GDGO，取出患者的診斷碼 icd9 和用藥代碼 drug_no。

- 抽樣：

  在住院資料統計中，沒住院的人口數遠大於住院人口數，為不平衡資料(un-balance data)；而長期住院的資料統計中，不平衡的狀況更加嚴重，因此透過抽樣來平衡資料。每年從沒住院的患者資料中，抽出與住院患者相等的數量，讓住院及沒住院的患者數為一比一，提升預測能力。

- 人口統計：

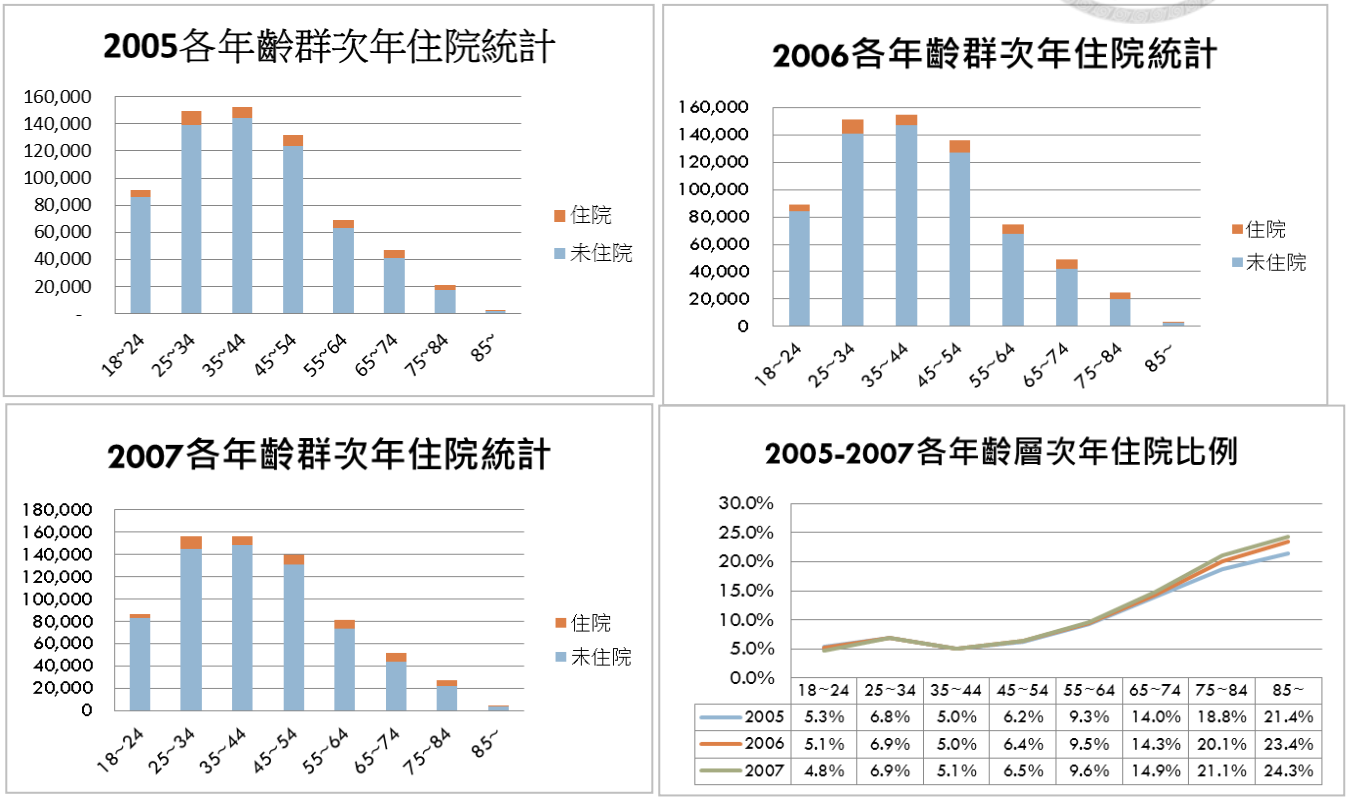
  篩選完後，將各年度的人口做簡單的統計，各年度的總人數列於下表。

表 3.2.1.1 各年度合格個體總人數

年度	2005	2006	2007
總人數	665,073	683,467	702,095

人口統計如下列前三張圖所示，長條圖中，橘色為次年住院人口，藍色為次年未住院人口，從圖中可看出青壯年的人口是最多的，45 歲之後的年齡群逐漸下降。右下的折線圖，是各年齡層中次年住院

人口佔該年齡層總人口的比例，從折線圖的趨勢可看出，從 45 歲的年齡群開始，次年住院人口逐漸上升，85 歲以上的年齡群，次年住院的比例高達 20%。



	18~24	25~34	35~44	45~54	55~64	65~74	75~84	85~
2005	5.3%	6.8%	5.0%	6.2%	9.3%	14.0%	18.8%	21.4%
2006	5.1%	6.9%	5.0%	6.4%	9.5%	14.3%	20.1%	23.4%
2007	4.8%	6.9%	5.1%	6.5%	9.6%	14.9%	21.1%	24.3%

圖 3.2.1.1 各年齡層住院統計(左上為 2005 年人口統計，右上為 2006 年人口統計，左下為 2007 年人口統計，橘色為住院人口，藍色為未住院人口；右下為各年齡層住院比例)

性別人口統計如以下三張圖所示，長條圖中，綠色為次年長期住院人口，橘色為次年短期住院人口，藍色為次年未住院人口，其中，女性人數多於男性人數。右下的圓餅圖，是所有人口中男女次年住院及未住院人數佔總人口數的比例，2005 到 2007 年的比例皆相近，男生住院人口比例與女生住院人口比例皆占該年度總人口數的 3%及 4%。

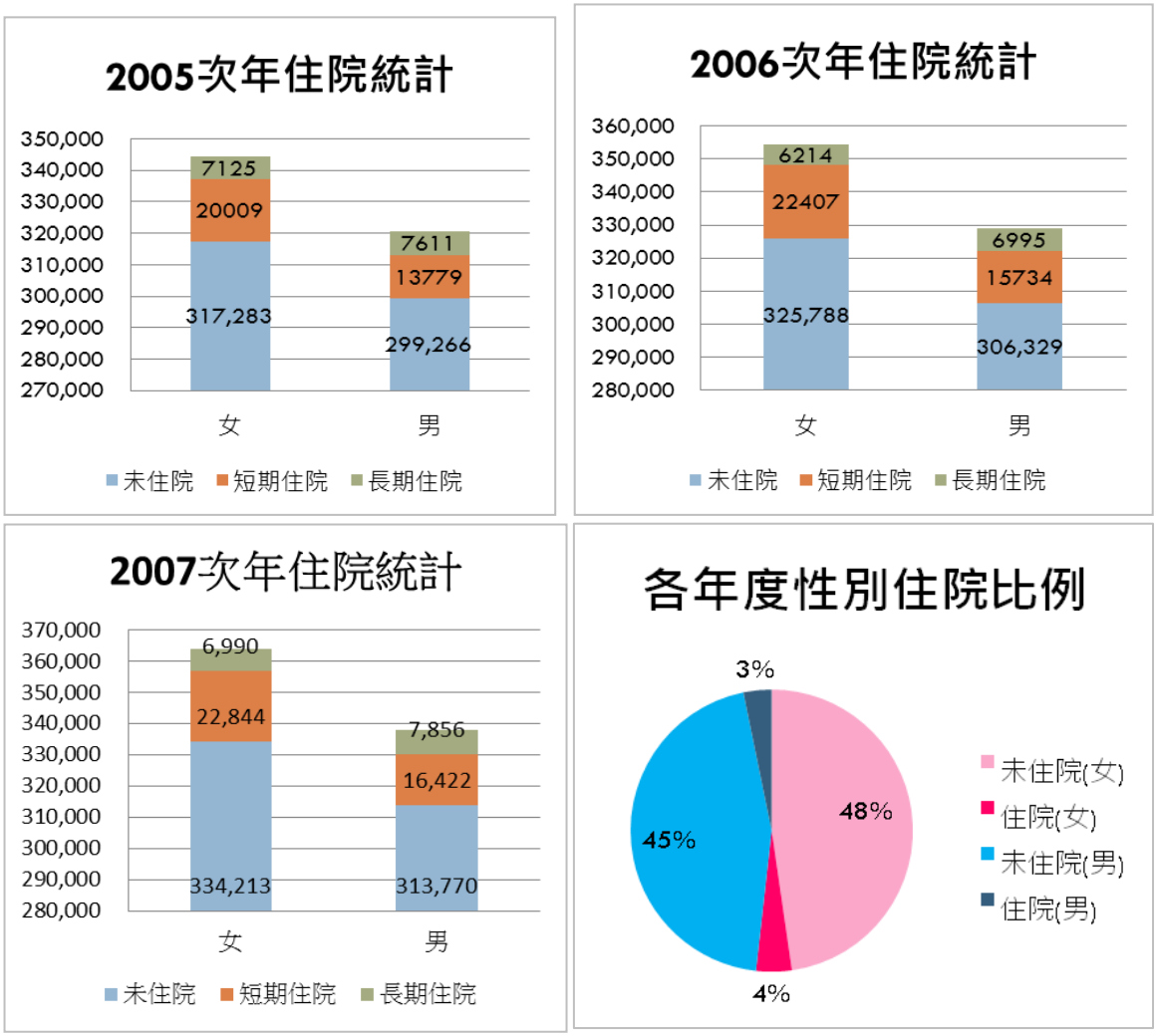


圖 3.2.1.2 性別住院統計(左上為 2005 年性別住院人口統計，右上為 2006 年性別住院人口統計，左下為 2007 年性別住院人口統計，其中綠色為長期住院人口，橘色為短期住院人口，藍色為未住院人口；右下為各年度性別在總人口中的住院比例)

3.2.2 預測特徵值

預測特徵值主要有人口統計變數、合併症指標、診斷碼、藥物代碼、LDA 主題，本小節將介紹如何產生這些預測特徵值。

- 年齡、性別：

  在前述實驗個體的篩選及資料處理後，即可得到患者的年齡及性別，根據文獻將年齡以 10 歲為一個區間，總共分成八個年齡群，八個年齡群分別為 18~24、25~34、35~44、45~54、55~64、65~74、75~84 及 85 歲以上。

- CCI / Deyo：

  CCI/Deyo 為以診斷為基礎的合併症指標，為國內最常用的查爾森合併症指標，共有 17 類合併症。CCI/Deyo 合併症對照表內包含疾病名稱、疾病診斷碼及該類別疾病的查爾森分數，若一病患同時有多個診斷碼屬於其中一類疾病，查爾森分數也只算一次。以疾病心肌梗塞(Myocardial Infarction)為例，此類疾病相關的診斷碼區間為 410 – 410.9，查爾森分數為 1 分，若該患者同時有 410 和 410.1 兩個診斷碼，都符合此區間的診斷碼，所以仍只算一分。舉例：若一位患者同時有兩診斷碼，410.1、410.2，這個個診斷碼皆屬於 CCI/Deyo 心肌梗塞(Myocardial Infarction)的診斷碼區間，因此，Deyo score 為 1 分。

  透過對照 CCI/Deyo 合併症對照表和患者的 icd9 診斷碼，加總計算患者得到的 CCI/Deyo 分數，完成以診斷為基礎合併症特徵值。

表 3.2.2.1 CCI/DEYO 對照表

ICD-9 CM	診斷類別	查爾森分數
410 – 410.9	心肌梗塞(Myocardial Infarction)	1
428 – 428.9	充血性心臟衰竭(Congestive HeartFailure)	1
433.9, 441 – 441.9, 785.4, V43.4	周邊血管疾病(Peripheral Vascular Disease)	1
430 – 438	腦血管疾病(Cerebrovascular Disease)	1
290 – 290.9	失智症(Dementia)	1
490 – 496, 500 – 505, 506.4	慢性肺部疾病(Chronic Pulmonary Disease)	1
710.0, 710.1, 710.4, 714.0 – 714.2, 714.81, 725	風濕性疾病(Rheumatologic Disease)	1
531 – 534.9	潰瘍性疾病(Peptic Ulcer Disease)	1
571.2, 571.5, 571.6, 571.4 – 571.49	輕度肝臟疾病(Mild Liver Disease)	1
250 – 250.3, 250.7	糖尿病(Diabetes)	1
250.4 – 250.6	糖尿病伴隨末端器官衰竭 (Diabetes with Chronic Complications)	2
344.1, 342 – 342.9	半身麻痺或下身麻痺(Hemiplegia or Paraplegia)	2
582 – 582.9, 583 – 583.7, 585, 586, 588 – 588.9	腎臟疾病(Renal Disease)	2
572.2 – 572.8	中度或重度肝臟疾病(Moderate or Severe Liver Disease)	3
140-172.9,174-195,200-208.9	惡性腫瘤，包括白血病與淋巴癌(Any malignancy,including leukemia andlymphoma)	3
196-199.1	轉移性腫瘤(Metastatic solid tumor)	6
042 – 044.9	後天免疫缺乏症候群(AIDS)	6

- PBDI：

PBDI 為以藥物處方為合併症指標，Dong YH (2013)依據台灣地區主要疾病調整 CDS 的合併症種類所制定的以藥物處方為基礎合併症指標，共對應 37 類台灣主要疾病的藥物類別，使用 ATC 的藥物編碼。

ATC 是世界衛生組織所制定的官方藥物分類系統，ATC code 共有七位，由英文跟數字組成，第 1、4、5 位，必為大寫英文字母，剩餘的為數字。此分類系統共分為五層，第一層為第 1 個字母，代表解剖學上作用的位置，共使用 14 個字母，第二層為 2 位的數字，為治療學的分類，第三層為 1 位的英文字母，是藥理學上的分類，第四層為 1 位的英文字母，表化學結構的分類，第五層為 2 位數字，表示化合物的分類。

透過將 drug_no 和 ATC 的對照表，將 drug_no 轉換成 ATC code，確認患者是否有使用該類別用藥。因 PBDI 無原始權重，本研究使用類別方式，將 PBDI 的每個類別作為一個特徵值，當某類別藥物使用多種，該類別仍只算一次。

表 3.2.2.2PBDI 與 ATC 對照表

Drug categories	**ATC codes of Drug classes/drugs**
Anti-platelet agents(抗血小板劑)	B01AC (except 06, 08, 15), C04AD03
Anti-coagulants(抗凝血劑)	B01AA-AB, B01AD-AE, B01AX
Anti-arrhythmics(抗心律失常)	C01BA-BD, C01BG , C01EB10
Cardiac stimulants(強心劑)	C01CA07, C01CE, C01CX
Digitalis glycosides(強心劑)	C01AA
ACEIs/ARBs*(降血壓、改善蛋白尿)	C09AA, C09BA-BB, C09CA, C09DA-DB
β blockers(β 阻斷劑)	C07AA-AB, C07AG, C07BA-BB, C07BG, C07CA-CB, C07CG, C07DA-DB, C07FA-FB
Ca2+ blockers(鈣離子阻斷劑)	C08CA, C08CX, C08DA-DB, C08EA, C08EX, C08GA
Loop diuretics(環管利尿劑)	C03CA-CC, C03CX , C03EB
Anti-hypertensives(抗高血壓藥)	C02AA -AC, C02BA-BB, C02CA, C02CC, C02DA-DD, C02DG, C02KA-KD, C02KX, C02LA-LC, C02LE-LG, C02LK-LL, C02LX, C03AA-AB, C03AH, C03AX, C03BA-BD, C03BX, C03DA-DB, C03EA, C03XA
Anti-angina agents(抗心絞痛劑)	C01DA, C01DX16
Anti-diabetic agents(抗糖尿病藥)	A10AB-AF, A10BA-BD, A10BF-BH, A10BX, A10XA
Anti-lipemic agents(抗血脂藥)	C10AA-AD, C10AX, C10BA, C10BX

Agents for thyroid disorder(甲狀腺失調藥物)	H03AA, H03BA-BB
Agents for respiratory illness(呼吸系統疾病藥物)	R03AC, R03AH, R03AK, R03BA-BC, R03BX, R03CC, R03DA-DC, R03DX
Anti-tubercular agents(抗結核藥)	J01GA01, J01GB04, J04AA-AD, J04AK, J04AM
Agents for gastric acid disorder(胃酸失調藥物)	A02BA-BD, A02BX02
Agents for inflammatory bowel disease (腸躁症藥物)	A07EC
Agents for hepatitis(治療肝炎藥物)	J05AB04, J05AF05, J05AF08, J05AF10-12, L03AB04-05, L03AB09-11
Agents for liver failure(治療肝功能衰竭藥物)	A06AD11, H01BA01, H01BA04, H01BA06, H01CB01-02
Agents for renal disease(治療腎臟疾病藥物)	V03AE, A11CC
Erythropoetins(醣蛋白激素)	B03XA
Anti-epileptics(抗癲癇藥)	N03AA-AG, N03AX
Anti-parkinsonian agents(抗帕金森劑)	N04BA-BD, N04BX
Agents for anxiety and tension(治療焦慮和緊張藥物)	N05BA-BE, N05BX
Lithium(鋰)	N05AN01

Anti-depressants(抗憂鬱藥物)	N06AA-AB, N06AF-AG, N06AX
Agents for malignancies(治療惡性腫瘤藥物)	L01AA-AD, L01AG, L01AX, L01BA-BC, L01CA-CD, L01CX, L01DA-DC, L01XA-XE, L01XX-XY, L03AA, L03AC, A04AA
Anti-psychotics(抗精神病藥)	N05AA-N05AH, N05AL, N05AX
Opioids(鴉片類藥物)	N02AA-AG, N02AX
Agents for rheumatic disease(治療風濕性疾病藥物)	L04AA13, L04AA24, L04AB, L04AC03, L04AC07 , M01CB, M01CC01, P01BA02
Glucocorticosteroids(糖皮質激素)	H02AB, H02BX
NSAIDs*(非類固醇抗發炎藥)	M01AA-AC, M01AE, M01AG-AH, M01AX
Anti-gout agents(抗痛風劑)	M04AA-AC, M04AX
Agents for transplant(器官移植藥物)	L04AA02-04, L04AA06, L04AA10, L04AA18-19, L04AC01-02, L04AD, L04AX01
Anti-glucoma agents(青光眼治療劑)	S01EA, S01EB (except 06), S01EC-EE
Anti-HIV* antivirals(抗 HIV* 病毒藥物)	J05AE, J05AF01-04, J05AF06-07, J05AF09, J05AG, J05AX07- 09, J05AR

- 疾病碼：

因病歷紀錄中的診斷碼數量多，因此會先設定門檻值，篩掉出現數量過低的疾病，減少特徵值的數量。若一年內 icd9 在當年度患者診斷紀錄出現次數少於 10 次，則刪除該診斷碼及相關記錄。

本研究透過資訊獲利 (Information Gain)及卡方值(Chi-square)選擇重要的特徵值。在選擇的過程中，我們藉由訓練集依照人數切 10 等分，其中一份當驗證集，做 10-folds cross validation，根據當年度的診斷資料與目標年度的住院結果，計算每個 icd9 特徵值對住院結果的資訊獲利及卡方值(Chi-square)，依序照資訊獲利、卡方值的大小排序特徵值對住院的影響力。

- 藥物代碼：

患者病歷紀錄中使用的藥物處方碼數量多，因此和診斷碼相同，會先設定門檻值，篩掉使用數量過低的藥物，減少特徵值的數量。因藥物紀錄的數量遠多於診斷碼，因此，若某藥物在當年度的患者處方用藥紀錄中出現數量少於 100 次，則刪除該藥物碼(drug_no)及相關資料，並將 drug_no 對應至 ATC code，方便之後解讀研究結果和比較不同特徵值的差異性。

接著同樣透過資訊獲利 (Information Gain)及卡方值(Chi-square)選擇重要的特徵值。根據用藥資料與目標年度的住院結果，計算每個 ATC code 特徵值對住院結果資訊獲利及卡方值(Chi-square)，並依序照資訊獲利和卡方值的大小排序特徵值對住院的影響力。

- LDA 特徵值：

將患者所有的診斷碼和使用的藥物代碼整理成文章的形式，每一個患者的醫療資料當作一篇文章，接著將這些患者的用藥及診斷資料，放入 LDA 主題模型學習，萃取出健康風險相關主題作為預測特徵值，最後會輸出所有患者的每個主題分佈，每一位患者所有主題的值加總為 1。

設定值如下：設定主題數量為 20、54、100、200、250、300 個主題，迴圈數設定為 1,500，在跑第 500 次之後，每 50 次會輸出一次文章中的主題分佈，最後將這些輸出的主題分佈平均，作為特徵值進行預測。

輸出主題分佈矩陣，如下所示意：

	1	2	…	第 N 位患者
Topic 1	0.28	0.2	…	0.21
Topic 2	0.4	0.1	…	0.066
…	…	…	…	…
Topic 20	0.12	0.08	…	0.11

3.3 預測特徵值組合

使用的人口統計變數的特徵值 (demographic features)，包含：年齡、性別，分別搭配以診斷為基礎的合併症指標、以藥物處方為基礎的合併症指標、診斷碼特徵值、藥物處方特徵值、LDA 萃取出的主題，並將上述五種特徵值互相搭配。其中，以人口統計變數加入最常用的合併症指標 CCI/Deyo 當作基準線模型 (baseline model)。

使用的特徵值組合有以下七種：

- 人口統計變數+CCI/Deyo：年齡、性別、CCI/Deyo
- 人口統計變數+PBDI：年齡、性別、PBDI
- 人口統計變數+以藥物處方為基礎合併症指標+以診斷為基礎合併症指標：年齡、性別、PBDI、CCI/Deyo
- 人口統計變數 + 診斷碼特徵值：年齡、性別、診斷碼特徵值(icd9)
- 人口統計變數 + 藥物特徵值：年齡、性別、藥物特徵值(ATC code)
- 人口統計變數 + 診斷碼特徵值+ 藥物特徵值：年齡、性別、診斷碼特徵值(icd9)及藥物特徵值(ATC code)
- 人口統計變數 + LDA：年齡、性別、LDA 主題

本實驗一共使用四個模型及 7 個特徵值組合進行次年(2005~2007)及兩年後(2005~2006)住院率預測，預測結果分成住院和住院期間大於平均住院天數的長期住院。

3.4 預測模型

本實驗主要使用以下四個分類模型預設次年住院率：支持向量機 (Support Vector Machine，簡稱：SVM)、羅吉斯回歸 (Logistic Regression)、隨機森林 (random forest)、類神經網路 (neural-net)，分別使用以下的 R package 實做：

- 支持向量機模型：e1071
- 羅吉斯回歸模型：glmnet
- 隨機森林：randomForest
- 類神經網路：nnet

在每次預測中，模型會依照人數切成 10 等份，每次會留一份當測試集，剩下的為訓練集，做 10-fold cross-validation，最後將 10 次的平均值做為結果。

在評估的部分，因不同立場注重的指標不同，對於個體來說，預測的 precision 高，較有做為參考的價值，能做為患者未來醫療規劃方向的參考，但從社會資源規劃的角度，recall 值高，代表真的會住院的人有確實被找出來，能做為政府機關或醫療單位未來資源配置的參考，進而提供個體適當的醫療服務，因此本研究採用了綜合性的指標 AUC （Area under the Curve），AUC 為 ROC（Receiver operating characteristic curve）下的面積，而 AUC 的值越大，正確率越高。

第四章　結果

本研究使用不同的特徵值預測未來住院機率，除了使用合併症指標外，也透過機器學習的方式，利用特徵值選取找出影響住院的疾病碼和藥物代碼，並透過 LDA 從患者的藥物及診斷紀錄中，找出影響住院的 Topic。本章將詳細說明本實驗的結果，觀察不同特徵值組合的預測未來住院表現，以及透過機器學習方式選出的特徵值與合併症指標是否具有差異性。

4.1 住院預測

- 特徵值數量選擇

  在第三章前處理的預測特徵值部分，已透過資訊獲利及卡方值排序特徵值對住院的影響程度，在進行模型預測前，需挑選預測能力最佳的特徵值數量。

  在選擇特徵值數量時，同樣做 10-folds cross validation，將訓練集分成十份，拿其中一分當驗證集，剩於維持當訓練集。在訓練模型的過程中，逐次增加放入的特徵值數量，觀察模型的預測表現，找到 AUC 最佳或開始平滑的點，選出預測能力最佳或最適當的數量。

  圖 4.1.1 為挑選羅吉斯回歸預測能力最佳的特徵值數量時，在放入不同數量診斷碼特徵值時 AUC 表現，逐次增加特徵值單位量為 100。由圖中可看出，在放入約前 1000 名的特徵值後，AUC 趨近於平滑，因此，羅吉斯回歸模型選擇 1000 個診斷碼特徵值當預測次年住院的特徵值數量。

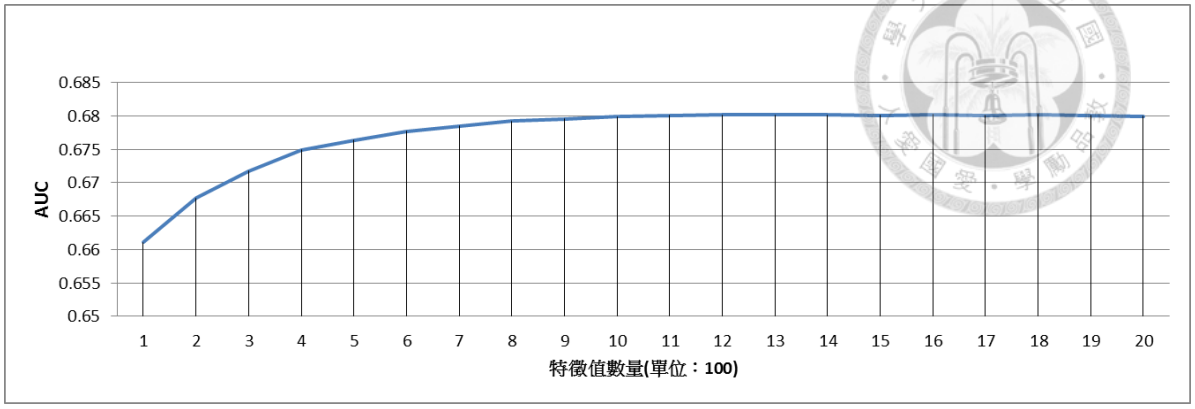


圖 4.1.1 羅吉斯回歸模型在不同診斷碼特徵值數量下的 AUC 表現

挑選最適當的藥物特徵值數量作法與前述相同，透過 10-folds cross validation，逐次增加放入的特徵值數量，從模型的預測表現找出 AUC 最佳或開始平滑的點作為要放入特徵值數量。

圖 4.1.2 為羅吉斯模型在放入不同數量藥物代碼特徵值的 AUC 表現，逐次增加的特徵值數量單位為 50，在放入約前面 300 個特徵值後，AUC 趨近於平滑，因此，選擇 300 個藥物代碼特徵值當作羅吉斯回歸預測次年住院的特徵值數量。

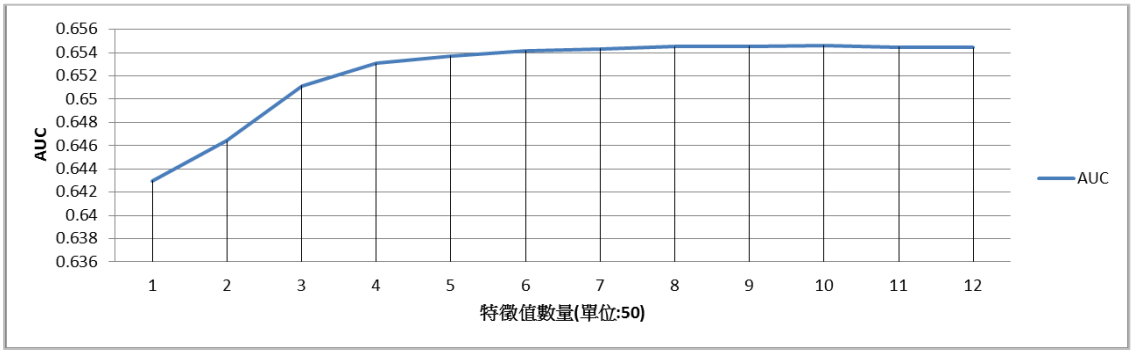


圖 4.1.2 羅吉斯回歸模型在不同藥物特徵值數量下的 AUC 表現

其他模型尋找最適合的診斷碼及特徵值數量，作法皆與上述所說方法相同，因此不再逐一說明。

挑選最適當的 LDA 數量，作法與前述相似，同樣透過 10-folds cross validation，分別放入 LDA 模型萃取出的 20、54、100、200、250、300 個主題，找出具有最佳預測能力的主題數量，做為預測特徵值數量，因做法類似，故不再舉例說明。

- 模型預測結果

透過合併症指標、挑選出的預測特徵值與主題進行預測後，下面為不同住院情形的預測結果，共分為預測次年住院、次年長期住院、兩年後住院、兩年後長期住院來討論。

- 預測次年住院：

下圖為在四個模型中，不同預測特徵值組合的 AUC 表現，長條圖中，由左到右的特徵值組合分別：CCI/Deyo、PBDI、CCI/Deyo+PBDI、ATC code、icd9、ATC code+icd9、LDA topics，前三項為合併症指標的特徵值組合、後四項為機器學習的特徵值組合。

在各模型中，機器學習的特徵值組合表現優於合併症指標的特徵值組合，在合併症指標的特徵值組合中，AUC 的表現的優劣順序為：CCI/Deyo+PBDI> PBDI > CCI/Deyo，在機器學習的特徵值組合中，相同的優劣順序為：ATC code + icd9> icd9 > ATC code，而 LDA topics 在羅吉斯回歸和類神經網路的表現介於 ATC code 和 icd9 間，在隨機森林和支持向量機的表現是機器學習的特徵值組合中最佳的。

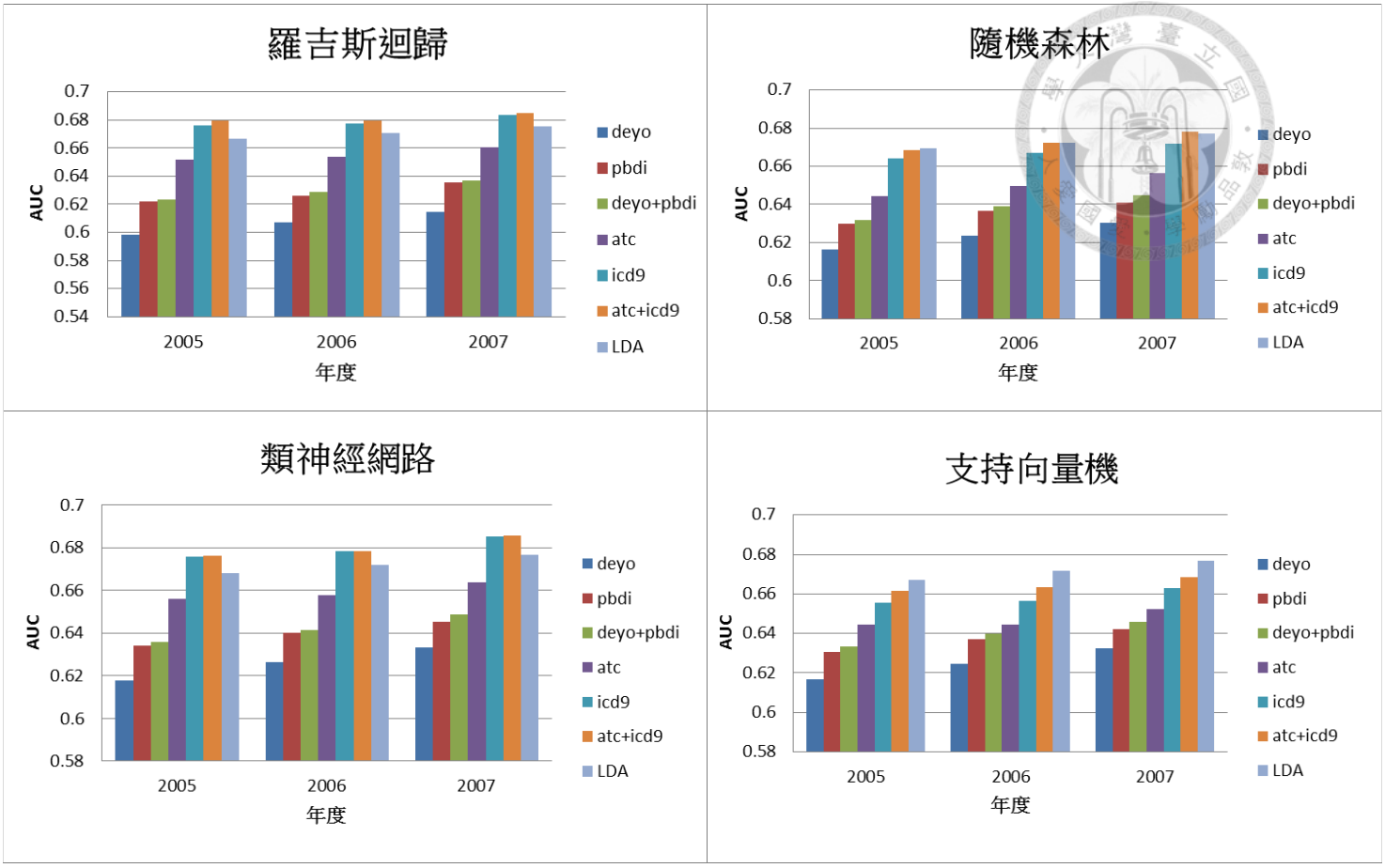


圖 4.1.3 各特徵值組合於各模型預測次年住院的 AUC 表現(時間：2005～2007 年)

針對每個模型，我們將不同特徵值組合預測的 AUC 值對基準線 CCI/Deyo 做 t-test，檢測有無顯著差異。在實驗中使用 10-folds cross validation，因此每年皆會有 10 個 AUC 值，我們依各特徵值組合將每年的 10 個值整合起來，與基準線 CCI/Deyo 做 paired t-test 的雙尾檢定，檢驗水準為 $\alpha=0.05$。

表 4.1.1 為各模型使用不同特徵值組合預測次年住院率的平均 AUC 表現及對基準線 CCI/Deyo 的顯著性，檢測的結果顯示，PBDI、CCI/Deyo+PBDI、ATC code、icd9、ATC code+icd9、LDA topics 皆顯著優於 CCI/Deyo。

表 4.1.1 各模型使用各特徵值組合預測次年住院率的 AUC 平均值及 t-test 顯著性 (時間為 2005~2007 年)

	羅吉斯回歸	隨機森林	類神經網路	支持向量機
CCI/Deyo	0.607	0.623	0.626	0.625
PBDI	0.628***	0.636***	0.640***	0.637***
Deyo+PBDI	0.630***	0.638***	0.642***	0.640***
ATC	0.655***	0.650***	0.659***	0.647***
icd9	0.679***	0.668***	0.680***	0.658***
ATC+icd9	0.681***	0.673***	0.680***	0.664***
LDA	0.671***	0.673***	0.672***	0.672***

*p<0.05, **p<0.01, ***p<0.001

針對預測次年住院的結果，我們將各模型的最佳表現挑出，透過 t-test 檢測各模型間預測次年住院的最佳結果是否有顯著差異性，表 4.1.2 將具顯著差異的模型組合列出，表內的值為各模型最佳預測表現的 AUC，從結果中可看出羅吉斯回歸和類神經網路的表現較好，這兩者顯著優於隨機森林和支持向量機的最佳表現，但羅吉斯回歸與類神經網路之間則是無顯著差異。

表 4.1.2 使用 t-test 檢測模型間預測次年住院的最佳 AUC 具顯著差異性的組合

(表格內的值為模型最佳表現的 AUC 值)

	最佳 AUC
羅吉斯回歸	0.681***
隨機森林	0.673

	最佳 AUC
羅吉斯回歸	0.681***
支持向量機	0.672

	最佳 AUC
類神經網路	0.680***
隨機森林	0.673

	最佳 AUC
類神經網路	0.680***
支持向量機	0.672

*p<0.05, **p<0.01, ***p<0.001

■ 預測次年長期住院：

在各模型中，機器學習的特徵值組合表現同樣優於合併症指標的特徵值組合，在合併症指標的特徵值組合中，AUC 表現的優劣順序與預測次年住院同為：CCI/Deyo+PBDI> PBDI > CCI/Deyo，在機器學習的特徵值組合中，整體的優劣順序為：ATC code + icd9> icd9 > ATC code，而 LDA topics 的表現較不穩定，在羅吉斯回歸和隨機森林的整體表現與 ATC code + icd9 相差不遠，不同的年間各有勝負，在隨機森林和支持向量機中 LDA topics 表現是機器學習中最佳的，但在類神經網路中的表現只接近 ATC code。但就合併症指標及機器學習方法找出的特徵值，整體的趨勢仍是機器學習方法優於合併症指標。

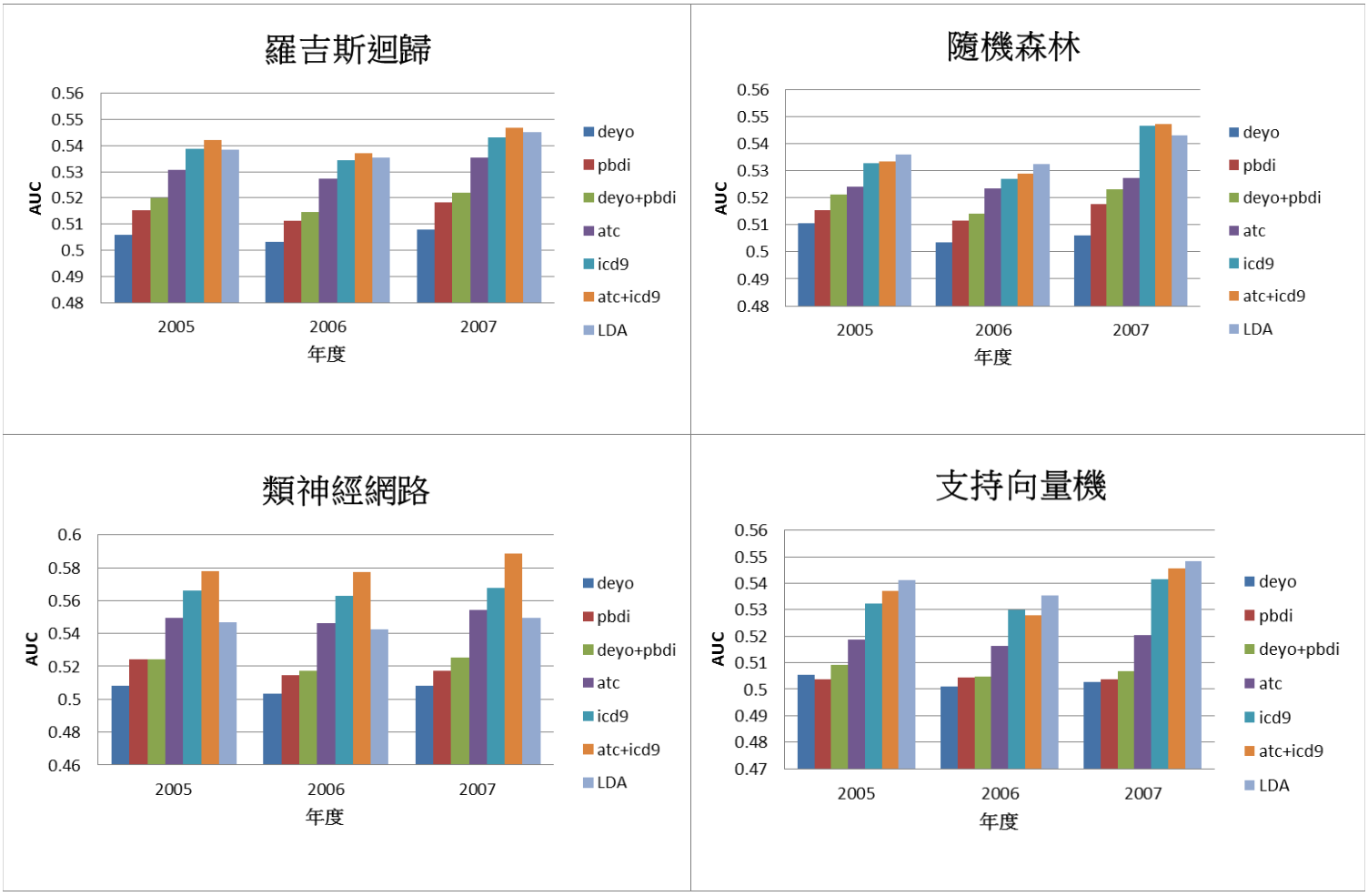


圖 4.1.4 各特徵值組合於各模型預測次年長期住院 AUC 表現(時間：2005～2007 年)

針對每個模型，使用不同特徵值組合預測的 AUC 值，檢測各組合對基準線 CCI/Deyo 有無顯著差異。與預測次年住院相同，我們依特徵值組合將每年的 10 個值整合起來，將各特徵值組核對 CCI/Deyo 做 paired t-test 的雙尾檢定，檢驗水準為 α=0.05。

表 4.1.3 檢測的結果顯示，除了利用支持向量機中以 PBDI 特徵值預測次年長期住院的結果不顯著外，其餘各模型中的不同特徵值組合預測表現皆顯著優於 CCI/Deyo。

表 4.1.3 各模型使用各特徵值組合預測次年長期住院率的 AUC 平均值及 t-test 顯著性

(時間：2005~2007 年)

	羅吉斯回歸	隨機森林	類神經網路	支持向量機
CCI/Deyo	0.506	0.507	0.507	0.503
PBDI	0.515 ***	0.515***	0.519***	0.504
Deyo+PBDI	0.519***	0.519***	0.522***	0.507***
ATC	0.531***	0.525***	0.550***	0.518***
icd9	0.539***	0.536***	0.566***	0.535***
ATC+icd9	0.542***	0.537***	0.581***	0.537***
LDA	0.540***	0.537***	0.546***	0.542***

*p<0.05, **p<0.01, ***p<0.001

針對預測次年長期住院的結果，我們將各模型的最佳表現挑出，透過 t-test 檢測各模型間預測次年長期住院的最佳結果是否有顯著差異性，表 4.1.4 將具顯著差異的模型組合列出，表內的值為各模型最佳預測表現的 AUC，從結果中可看出類神經網路的表現最好，且顯著優於其他模型。

表 4.1.4 使用 t-test 檢測模型間預測次年住院的最佳 AUC 具顯著差異性的組合

(表格內的值為模型最佳表現的 AUC 值)

	最佳 AUC
類神經網路	0.581***
羅吉斯回歸	0.542

	最佳 AUC
類神經網路	0.581***
隨機森林	0.537

	最佳 AUC
類神經網路	0.581***
支持向量機	0.542

	最佳 AUC
支持向量機	0.542***
隨機森林	0.537

	最佳 AUC
羅吉斯回歸	0.542***
隨機森林	0.537

*p<0.05, **p<0.01, ***p<0.001

■ 預測兩年後住院：

在羅吉斯回歸、隨機森林及類神經網路中，機器學習的特徵值組合表現同樣優於合併症指標的特徵值組合，在合併症指標的特徵值組合中，AUC 的表現的優劣順序與前面部分皆相同：CCI/Deyo + PBDI> PBDI > CCI/Deyo，在機器學習的特徵值組合中，整體的優劣順序為：ATC code + icd9 > icd9 > ATC code，而 LDA topics 的表現較不穩定，在羅吉斯回歸和隨機森林的表現與 ATC code + icd9 相差不遠，但在類神經網路中，表現只接近 ATC code。

但在支持向量機中，合併症指標的特徵值組合中，AUC 的表現的優劣順序雖與前面部分同為：CCI/Deyo + PBDI> PBDI > CCI/Deyo，但 PBDI 和 CCI/Deyo + PBDI 的表現相差不大，此外，機器學習的特徵值組合 ATC code + icd9 、 icd9 、 ATC code 的表現與合併症指標差不多，LDA topics 的表現優於其他的特徵值組合。推測 ATC code + icd9 、 icd9 、 ATC code 差異不大的原因為選取要放入模型中的特徵值數量時，參數未調整適當，導致結果與其他模型的趨勢不同。

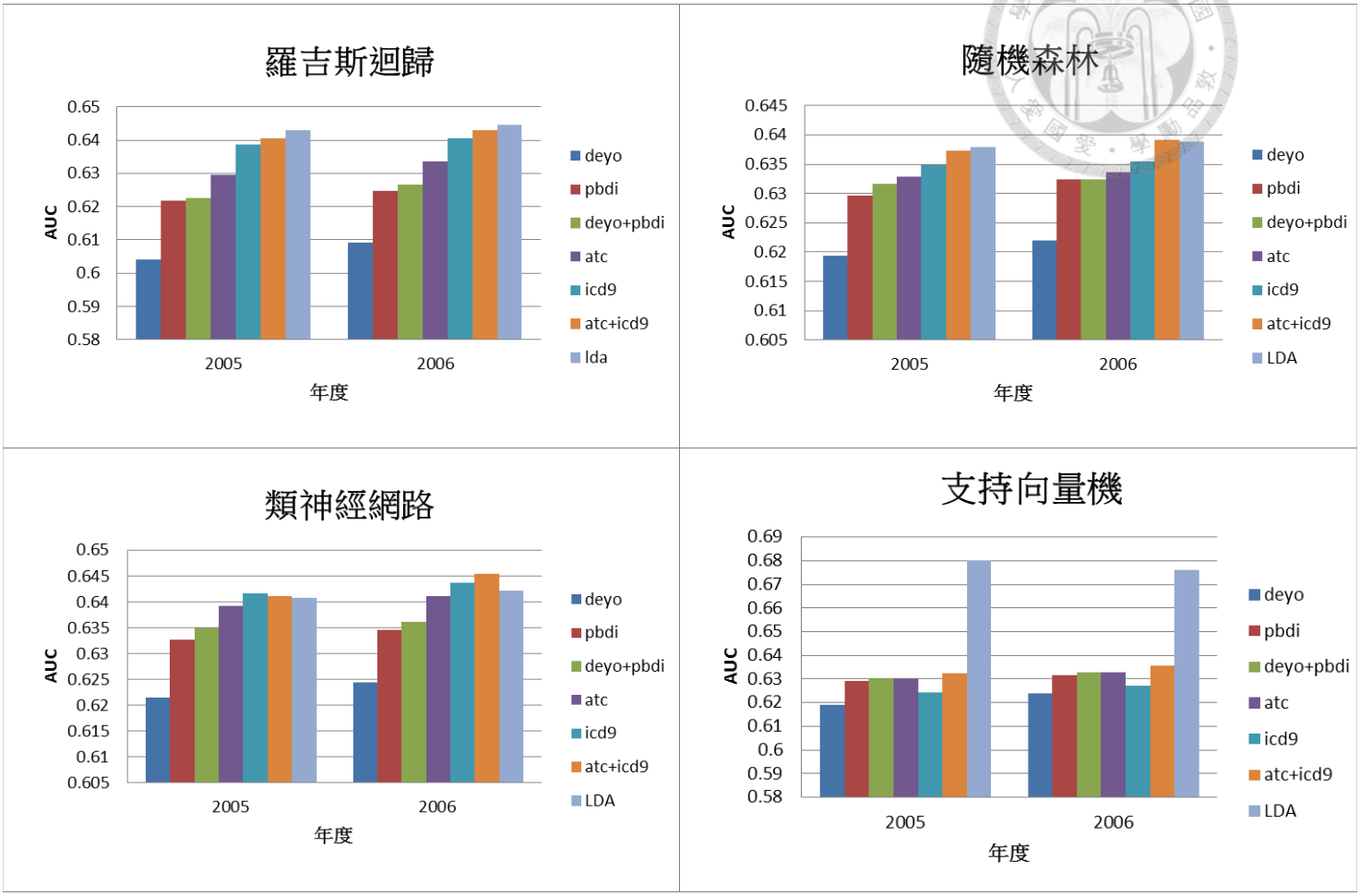


圖 4.1.5 各特徵值組合於各模型預測兩年後住院的 AUC 表現(時間：2005～2006 年)

針對各模型，使用不同特徵值組合預測的 AUC 值，檢測各組合對基準線 CCI/Deyo 有無顯著差異。與前述做法相同，我們依各特徵值組合將每年的 10 個值整合起來，做 paired t-test 的雙尾檢定，檢驗水準為 $\alpha=0.05$。檢測的結果顯示，各模型中的不同特徵值組合皆顯著優於 CCI/Deyo。

表 4.1.5 各模型使用各特徵值組合預測兩年後住院率的 AUC 平均值及 t-test 顯著性

(時間：2005~2006 年)

	羅吉斯回歸	隨機森林	類神經網路	支持向量機
Deyo	0.607	0.621	0.623	0.622
PBDI	0.623***	0.631***	0.634***	0.626***
Deyo+PBDI	0.623***	0.632***	0.636***	0.632***
ATC	0.632***	0.633***	0.640***	0.631***
icd9	0.640***	0.635***	0.643***	0.626 **
ATC+icd9	0.642***	0.638***	0.643***	0.634***
LDA	0.644***	0.638***	0.642***	0.678 **

*p<0.05, **p<0.01, ***p<0.001

我們同樣透過 t-test 檢測各模型間預測兩年後住院的最佳結果是否有差異性，下表將具顯著差異的模型組合列出，從結果中可看出支持向量機的表現顯著優於其他模型。

表 4.1.6 使用 t-test 檢測模型間預測兩年後住院的最佳 AUC 具顯著差異性的組合

(表格內的值為模型最佳表現的 AUC 值)

	最佳 AUC
支持向量機	0.678***
羅吉斯回歸	0.644

	最佳 AUC
支持向量機	0.678***
隨機森林	0.638

	最佳 AUC
支持向量機	0.678***
類神經網路	0.643

	最佳 AUC
類神經網路	0.643***
隨機森林	0.638

	最佳 AUC
羅吉斯回歸	0.644***
隨機森林	0.638

*p<0.05, **p<0.01, ***p<0.001

■ 預測兩年後長期住院：

在各模型中，機器學習的特徵值組合表現優於合併症指標的特徵值組合，在合併症指標的特徵值組合中，AUC 的表現的優劣順序與前面同為：CCI/Deyo+PBDI> PBDI > CCI/Deyo，在機器學習的特徵值組合中，整體的優劣順序為：ATC code + icd9> icd9 > ATC code。LDA topics 的表現優於合併症指標，但在各模型中的表現結果有所差異，在羅吉斯回歸的表現，與 ATC code + icd9 相差不遠，在隨機森林和類神經網路中，表現接近加入 ATC code 的特徵值組合，而支持向量機中，LDA topics 的表現是此模型中表現最好的，但整體的趨勢仍是機器學習方法優於合併症指標。

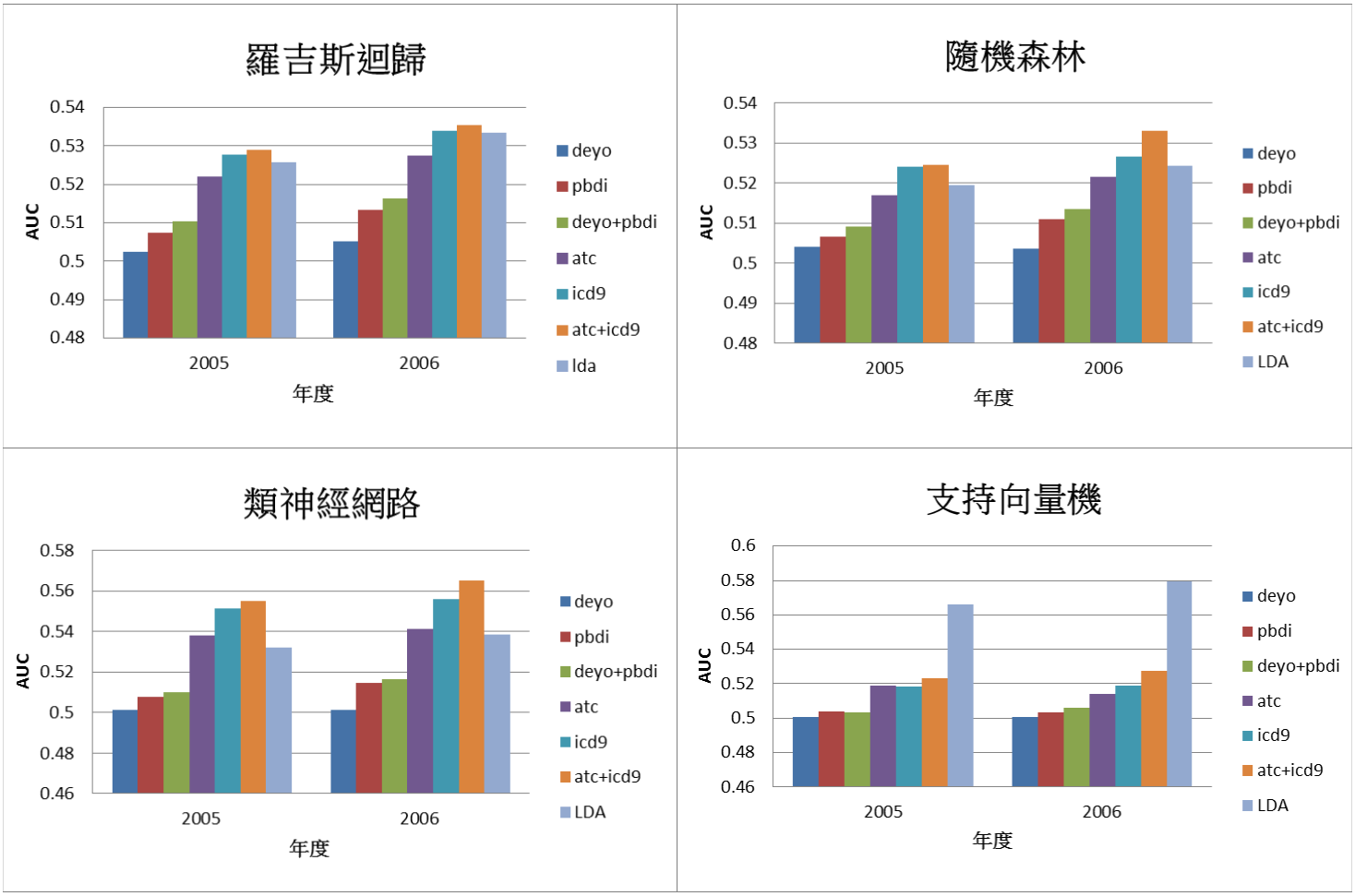


圖 4.1.6 各特徵值組合於各模型預測兩年後長期住院的 AUC 表現(時間：2005～2006 年)

針對每個模型，使用不同特徵值組合預測的 AUC 值，檢測各組合對基準線 CCI/Deyo 有無顯著差異。我們使用 t-test 做檢測，方法與設定和前面相同。檢測的結果顯示，各模型中的不同特徵值組合的預測結果皆顯著優於 CCI/Deyo。

表 4.1.7 各模型使用各特徵值組合預測兩年後長期住院率 AUC 平均值及 t-test 顯著性

(時間：2005～2006 年)

	羅吉斯回歸	隨機森林	類神經網路	支持向量機
Deyo	0.504	0.504	0.501	0.501
PBDI	0.510***	0.509***	0.511***	0.504***
Deyo+PBDI	0.513***	0.511***	0.513***	0.505***
ATC	0.525***	0.519***	0.540***	0.516***
icd9	0.531***	0.525***	0.554***	0.519***
ATC+icd9	0.532***	0.529***	0.560***	0.525***
LDA	0.530***	0.522***	0.535***	0.573***

*p<0.05, **p<0.01, ***p<0.001

同樣透過 t-test 檢測各模型間預測兩年後長期住院的最佳結果是否有顯著差異，我們將具顯著差異性的組合列於下表，從結果中可看出支持向量機的表現顯著優於其他模型，類神經模型的表現也不錯，優於羅吉斯回歸和隨機森林。

表 4.1.8 使用 t-test 檢測模型間預測兩年後長期住院最佳 AUC 具顯著差異性的組合

(表格內的值為模型最佳表現的 AUC 值)

	最佳 AUC
支持向量機	0.573**
羅吉斯回歸	0.532

	最佳 AUC
支持向量機	0.573*
隨機森林	0.529

	最佳 AUC
支持向量機	0.573**
類神經網路	0.560

	最佳 AUC
類神經網路	0.560***
羅吉斯回歸	0.532

	最佳 AUC
類神經網路	0.560***
隨機森林	0.529

*p<0.05, **p<0.01, ***p<0.001

- 小結：

觀察各特徵值組合預測不同住院結果的表現，可觀察到以下幾個趨勢，在合併症指標的部分，結合診斷和藥學基礎合併症指標的預測表現，優於單獨使用其中一種合併症指標。在機器學習的部分，透過機器學習方法找出的特徵值具有預測未來住院的能力，且表現優於合併症指標，根據實驗的結果應是能做為預測住院風險的特徵值。

在模型的預測能力部分，我們檢測模型的最佳預測表現是否有差異，從結果可看出，整體而言，類神經網路的表現還不錯。在預測兩年後住院，支持向量機表現較佳。

4.2 特徵值選取

本研究透過機器學習方式找出與住院相關的特徵值，本小節將重要特徵值列出，探討與未來健康風險相關性高的特徵值。

- 診斷碼特徵值選擇：
  - 預測次年住院：

    從表 4.2.1 中可觀察到，在預測次年住院的特徵值中，主要分為以下三部分，第一種為女性懷孕相關診斷碼，包含：Supervision of other normal pregnancy(產前檢查)、Supervision of normal first pregnancy(首次正常妊娠的監測)、Threatened abortion, antepartum condition or complic(迫切流產)等診斷碼，懷孕後次年因生產而住院，十分直觀；第二種為慢性病，包含 Essential hypertension, unspecified(本態性高血壓;自發性高血壓)、Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)、Constipation(便祕)、Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)、Senile cataract (老年性白內障)等，顯示慢性疾病與次年住院有一定的相關性，從原始資料觀察，罹患慢性病且住院的多為年長者；第三種為高齡者盛行率高的疾病，例如：Hypertrophy (benign) of prostate(良性攝護腺肥大前列腺增生)、Urinary tract infection, site not specified(泌尿道感染)、Dizziness and giddiness(眩暈)等，因年紀導致生理結構及免疫系統改變，若症狀輕微而忽視治療，當未來抵抗力下降也存在住院風險。此外特徵值選取中出現了 Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)，從原始資料查看後，發現接踵疫苗多為高齡者，55 歲以上的接踵者佔所有接踵人數的九成以上，比對疫苗接踵規定，接踵者有年齡或特殊疾病限制，因此導致住院原應為年齡而非疫苗有問題。

表 4.2.1　2005 到 2007 年預測次年住院診斷碼特徵值(前 16 名)

	2005		2006		2007	
	icd9	症狀	icd9	症狀	icd9	症狀
1	V22.1	Supervision of other normal pregnancy(產前檢查)	V22.1	Supervision of other normal pregnancy(產前檢查)	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)
2	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)	250.00	Supervision of other normal pregnancy(產前檢查)
3	V22.0	Supervision of normal first pregnancy(首次正常妊娠的監測)	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)	V04.8	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)
4	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)	486	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)
5	V048	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)	V22.0	Supervision of normal first pregnancy(首次正常妊娠的監測)	496	Supervision of normal first pregnancy(首次正常妊娠的監測)
6	564.0	Constipation(便祕)	564.0	Constipation(便祕)	564.0	Constipation(便祕)
7	640.03	Threatened abortion, antepartum condition or complic(迫切流產)	780.4	Dizziness and giddiness(眩暈)	402.90	Abdominal pain, unspecified site(腹痛 (嬰兒腹絞痛), 下腹痛)
8	780.4	Dizziness and giddiness(眩暈)	789.00	Abdominal pain, unspecified site(腹痛 (嬰兒腹絞痛), 下腹痛)	438.9	Dizziness and giddiness(眩暈)
9	789.00	Abdominal pain, unspecified site(腹痛 (嬰兒腹絞痛), 下腹痛)	640.03	Threatened abortion, antepartum condition or complic(迫切流產)	428.0	Threatened abortion, antepartum condition or complic(迫切流產)
10	414.9	Chronic ischemic heart disease, unspecified(缺血性心臟病)	600.0	Hypertrophy (benign) of prostate(良性攝護腺肥大前列腺增生)	585	Urinary tract infection, site not specified(泌尿道感染)
11	599.0	Urinary tract infection, site not specified(泌尿道感染)	599.0	Urinary tract infection, site not specified(泌尿道感染)	780.4	Unspecified hypertensive heart disease without conge(高血壓性心臟病)
12	643.13	Hyperemesis gravidarum with metabolic disturbance(妊娠劇吐（未生產）)	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)	599.0	Hypertrophy (benign) of prostate(良性攝護腺肥大前列腺增生)
13	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)	414.9	Chronic ischemic heart disease, unspecified(缺血性心臟病)	518.81	Chronic ischemic heart disease, unspecified(缺血性心臟病)
14	715.90	Osteoarthrosis, unspecified whether generalized or localized, site unspecified(骨性關節炎;骨關節病變)	49.6	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)	600.0	Senile cataract, unspecified(老年性白內障)
15	366.10	Senile cataract, unspecified(老年性白內障)	366.10	Senile cataract, unspecified(老年性白內障)	V70.0	Chronic kidney disease(慢性腎衰竭; 尿毒症)
16	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)	782.3	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)

■ 預測次年長期住院：

從表 4.2.2 中可觀察到，在預測次年長期住院的部分排除了懷孕的狀況，因生產的住院天數通常小於平均住院天數。預測長期住院的特徵值主要分為以下幾個兩部分，第一種為慢性病，這些慢性病在年長者的比例較高，包含 Essential hypertension, unspecified(本態性高血壓;自發性高血壓)、Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)、Chronic kidney disease(慢性腎衰竭; 尿毒症)、Constipation(便祕)、Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)、Senile cataract (老年性白內障)等，顯示慢性疾病與次年長期住院有一定的關聯性，且住院者多為年長者;第二種為高齡者盛行率高的疾病，例如：Hypertrophy (benign) of prostate(良性攝護腺肥大前列腺增生)、Urinary tract infection, site not specified(泌尿道感染)、Dizziness and giddiness(眩暈)、Pneumonia, organism unspecified(肺炎)等，年長者因年紀導致的生理結構及免疫系統改變，抵抗力也隨之下降，染病後可能因初期症狀輕微而忽視治療或因抵抗力不足，導致最後住院。

此外，在長期住院的特徵值出現了 Unspecified schizophrenia, unspecified(精神分裂症)，從相關文獻中得知，嚴重的精神性疾病通常會有長期住院與重複性入院的狀況，原因為精神性疾病可能會有幻覺、妄想、行為舉止怪異等狀況，程度嚴重的話可能會有自殘或傷人的危險性，因此入院的觀察及治療時間會較長，且出院後可能因無法融入社會或無定期回診治療以及其他原因而復發導致再入院，因此精神性疾病在醫療資源的消耗也較多，近期精神性疾病病患的治療照護及人權等相關議題仍在討論中，從精神性疾病出現在長期住院的特徵值中，推論此疾病與未來長期住院具相關性，需要社會上多加注意。

表 4.2.2　2005 到 2007 年預測次年長期住院診斷碼特徵值(前 16 名)

	2005		2006		2007	
	icd9	症狀	icd9	症狀	icd9	症狀
1	401.9	Essential hypertension, unspecified(本態性高血壓; 自發性高血壓)	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)	401.9	Essential hypertension, unspecified(本態性高血壓; 自發性高血壓)
2	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)
3	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)
4	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)	564.0	Constipation(便祕)	486	Pneumonia, organism unspecified(肺炎)
5	715.90	Osteoarthrosis, unspecified whether generalized or localized, site unspecified(骨性關節炎;骨關節病變)	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)
6	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)	564.0	Constipation(便祕)
7	564.0	Constipation(便祕)	486	Pneumonia, organism unspecified(肺炎)	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)
8	780.4	Dizziness and giddiness(眩暈)	438.9	Unspecified late effects of cerebrovascular disease (中風後遺症;腦血管疾病後遺症; 陳舊性腦血管意外)	438.9	Unspecified late effects of cerebrovascular disease (中風後遺症;腦血管疾病後遺症;陳舊性腦血管意外)
9	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)	585	Chronic kidney disease(慢性腎衰竭; 尿毒症)	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)
10	366.10	Senile cataract, unspecified(老年性白內障)	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)	585	Chronic kidney disease(慢性腎衰竭; 尿毒症)
11	295.90	Unspecified schizophrenia, unspecified(精神分裂症)	295.90	Unspecified schizophrenia, unspecified(精神分裂症)	780.4	Dizziness and giddiness(眩暈)
12	780.52	Insomnia, unspecified(失眠)	V70.0	Routine general medical examination at a health care facility(在保健機構的一般常規醫學檢查成人健檢健康檢查)	599.0	Urinary tract infection, site not specified(泌尿道感染)
13	486	Pneumonia, organism unspecified(肺炎)	715.90	Osteoarthrosis, unspecified whether generalized or localized, site unspecified(骨性關節炎;骨關節病變)	518.81	Acute respiratory failure(呼吸衰竭)
14	721.3	Lumbosacral spondylosis without myelopathy(腰椎、骨、腰部退化性脊椎炎腰骶骨椎關節黏連)	782.3	Edema(水腫)	600.0	Hypertrophy (benign) of prostate(良性攝護腺肥大前列腺增生)

15	438.9	Unspecified late effects of cerebrovascular disease (中風後遺症;腦血管疾病後遺症;陳舊性腦血管意外)	366.10	Senile cataract, unspecified(老年性白內障)	V70.0	Routine general medical examination at a health care facility(在保健機構的一般常規醫學檢查成人健檢健康檢查)
16	V70.0	Routine general medical examination at a health care facility(在保健機構的一般常規醫學檢查成人健檢健康檢查)	780.4	Dizziness and giddiness(眩暈)	782.3	Edema(水腫)

- 預測兩年後住院及長期住院：

特徵值多與預測長期住院的特徵值相似，包含慢性病及年長者盛行率高的疾病，例如：高血壓、糖尿病、慢性阻塞性肺病、白內障、心臟病、失眠等等，這之中許多特徵值也是國人的十大死因，近代被關注的精神分裂症也出現於兩年後長期住院，由此可驗證這些疾病對未來健康風險確實有威脅性，因此，對於個體或社會來說，這些疾病需要被多加注意。

- 藥物處方特徵值選擇：

- 預測次年住院：

因 2005 至 2007 年的藥物特徵值內容相似，因此只將 2005 年的前幾名特徵值列出作為說明的參考，其餘各年的特徵值將收於附錄。在 2005 年到 2007 年的藥物特徵值第一名為 SODIUM CHLORIDE(生理食鹽水注射液)，推測應為今年住院次年再住院的機率高。

從藥物的適應症，可看出許多藥物為治療慢性病，像是治療便秘、糖尿病、失眠、狹心症等，除此之外，藥物特徵值也包含許多一般性常用藥物，包含胃藥、消炎止痛等常用藥物，這些藥物適用於許多常見症狀，例如：發燒、發炎、腸胃不適等許多疾病常見的症狀，許多年長者盛行的疾病都出現這些常見症狀，因此，從適應症來看與疾病特徵值具有相似性。但在藥物特徵值的部分，即排除了懷孕的情形，因懷孕要檢查，但正常的狀況下不需要吃藥，推測此原因為藥物特徵值預測次年住院表現較差的原因。

表 4.2.3　2005 年預測次年住院藥物處方特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
2	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
3	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
4	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
5	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
6	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
7	A03FA01	METOCLOPRAMIDE HCL MONOHYDRATE(胃斯妥持效膠囊)	消化器機能異常(噁心、嘔吐、腹部膨滿感)
8	A02AX	STACAINE(胃必康)	急慢性胃炎、胃痛、胃灼熱、胃酸過多、胃部不適感
9	N02BE01	ACETAMINOPHEN (普除痛錠)	解熱、止痛(關節痛、肌肉痛、風濕痛、神經痛、月經痛、牙痛、頭痛之舒緩)
10	A03AX13	DIMETHYLPOLYSILOXANE(瓦斯康錠)	適合空氣嗤下症及腹部具有膨滿感、鼓腸等。
11	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
12	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
13	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
14	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
15	N07CA91	DIPHENIDOL HCL(敵芬尼朵糖衣錠)	止吐
16	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀

■ 預測次年長期住院：

在預測次年長期住院的部分，連續三年的特徵值內容都十分相似，因此同樣只將 2005 年的前幾名特徵值列表當說明的參考，其餘各年的重要特徵值請見附錄。在預測次年長期住院的特徵值中，前三名中有兩個藥品類型皆為注射液，一是 SODIUM CHLORIDE(生理食鹽水注射液)，另一為治療高血壓、急性肺水腫的 FUROSEMIDE(樂泄靜脈注射液)，通常注射液類型的藥品多為院內所使用，因此與預測次年住院的結果的解讀一樣是患者今年住院，次年再住院的機率也高。在預測長期住院的部分同樣也出現許多治療慢性病的藥物，例如：便秘、糖尿病、失眠、類風溼性關節炎、狹心症、帕金森氏症等疾病，此外也包含許多用於常見症狀的一般性藥物，從藥物的適應症可看出與疾病特徵值的相似性。

與預測次年住院特徵值比較不同的是，癲癇、帕金森氏症在長期住院中的排名較前面，取代了腸胃性疾病的用藥。

表 4.2.4　2005 年預測次年長期住院藥物處方特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
2	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
3	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充
4	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生
5	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
6	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
7	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
8	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
9	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
10	N03AE01	CLONAZEPAM(癲可錠)	癲癇
11	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
12	N04AA01	TRIHEXYPHENIDYL HCL(顫立靜錠)	帕金森氏症
13	M01AH01	CELECOXIB(希樂葆膠囊)	緩解成人類風濕性關節炎之症狀與徵兆
14	M01AC06	MELOXICAM(美樂錠)	類風濕性關節炎、骨關節炎及僵直性脊椎炎之症狀治療
15	B01AC07	DIPYRIDAMOLE(保心丁注射液)	對急、慢性狹心症之治療可能有效
16	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠

■ 預測兩年後住院及長期住院：

在預測兩年後住院的部分，內容與次年住院出現的特徵值相似，從藥物的適應症來看，可分為以下幾種，第一種為治療慢性病，像是便秘、糖尿病、失眠及狹心症等，除此之外，藥物特徵值也包含許多一般性常用藥物，包含胃藥、消炎止痛等常用藥物，這些用藥可能為年長者盛行的疾病所使用。第二種為注射液類型藥劑，包括：SODIUM CHLORIDE(生理食鹽水注射液)、治療高血壓和急性肺水腫的 FUROSEMIDE(樂泄靜脈注射液)、治療狹心症的 DIPYRIDAMOLE(保心丁注射液)，與前面的解讀同樣為今年住院的患者，明年在入院的可能性也較高。

而預測兩年後長期住院的特徵值，內容與長期住院相似，包括：慢性疾病用藥、再入院型的患者，但多了精神性疾病，推測原因為精神性疾病的影響是較長遠的，需要較長的治療時間或無法根治。

● 與合併症指標比較

■ 與以診斷為基礎合併症比較：

我們將診斷碼特徵值於羅吉斯回歸的係數較高的前幾名與 CCI/Deyo 疾病類別做比較，從四個不同的結果：次年住院、次年長期住院、兩年後住院、兩年後長期住院，發現以下重疊的部分，包含：失智症(Dementia)、糖尿病(Diabetes)、轉移性腫瘤 (Metastatic solid tumor)、惡性腫瘤，包括白血病與淋巴癌(Any malignancy, including leukemia and lymphoma)等。

比較時也有發現相異的部分，這部分在羅吉斯回歸的係數較高，應是影響住院的重要特徵值，但未包含於診斷碼特徵值中，例如：精神性疾病，其中，器質性幻覺徵候群、單純型精神分裂症、情感型性精神分裂症等在預測未來長期性住院的羅吉斯回歸係數較高，需要較為注意。

此外，影響住院程度高的特徵值還出現許多意外傷病相關診斷碼，例如：

燒傷、骨折等相關診斷碼，推測這些傷害需要的復原時間較長，次年可能還需要進行手術，例如：骨折患者若有植入鋼釘，次年可能須進行手術拆除。

■ 與以藥物處方為基礎合併症比較：

我們將藥物處方特徵值於羅吉斯回歸的係數較高的前幾名與 PBDI 內容做比較，發現重疊的部分有： Anti-hypertensives(抗高血壓藥) 、 Anti-diabetic agents(抗糖尿病藥)、 Agents for respiratory illness(呼吸系統疾病藥物)、 Anti-psychotics(抗精神病藥)、Agents for transplant(器官移植藥物)等。

此外，相異的部分為人工受孕相關用藥，此藥物在羅吉斯回歸的係數也是前幾名，但造成住院的原因是孕婦生產，並非藥物有不良影響。

■ 小結：

特徵值選擇找出的診斷碼和藥物與合併症，有許多類似的慢性疾病，例如：高血壓、糖尿病、心臟相關疾病等。

從與合併症比較中可看出，PBDI 包含的藥物範圍較廣，舉例來說，精神疾病用藥及預防器官移植排斥的用藥都是藥物特徵值中係數較高的藥物，這兩個部分都有包含在 PBDI 中，但診斷碼特徵值，係數較高的診斷碼有出現精神疾病，但 CCI/Deyo 並未包含精神疾病。推測這是 PBDI 表現優於 CCI/Deyo 的原因之一。

- LDA topics

我們從 LDA topics 特徵值於羅吉斯回歸的係數較高的前幾名，討論會影響住院的主題。

■ 次年住院：

影響次年住院的主題主要有三個。

第一個是 topic 78，此主題與懷孕相關，主要的診斷碼與藥物處方有：V22.1(Supervision of other normal pregnancy，產前檢查)、626.0(Absence of menstruation，無月經症)、643.0(Mild hyperemesis gravidarum，妊娠輕度嘔吐)、G03DA04(PROGESTERONE，黃體素注射液)不孕症等，次年因生產而住院，所以此主題十分合理。

第二個是 topic 231，此主題與慢性疾病相關，包含以下藥物及診斷碼：571.5 (Cirrhosis of liver without mention of alcohol，慢性肝病及肝硬化) 、C03CA01 (FUROSEMIDE，來適泄注射液)高血壓、C03DA01(SPIRONOLACTONE，使佳通膜衣錠)高血壓、 N05BA06(LORAZEPAM，悠然錠)焦慮失眠症等，從內容可看出這些慢性病許多都出現在合併症指標及找出的特徵值中，是現代人健康的隱憂。

第三個是 topic 173，此主題與精神性疾病相關，內容包含：N05AL01 (SULPIRIDE，舒復寧膜衣錠)精神病狀態、 N05CD04 (ESTAZOLAM，艾斯樂錠) 失眠、 N05AX08 (RISPERIDONE ，悠寧膜衣錠)精神分裂症之相關症狀、295.90(Unspecified schizophrenia, unspecified，精神分裂症) 、N04AA01(TRIHEXYPHENIDYL HCL，顫立靜錠)帕金森氏症等，精神性疾病也出現在 PBDI 和診斷碼及藥物特徵值中，因此，精神性疾病需要多加重視。

■ 次年長期住院：

次年長期住院的主題主要也有三個，其中一個是 topic 158 (精神性疾病)，因內容與次年住院的主題 topic 173 相似，所以不多加贅述。

topic 2 的內容為肝疾病相關診斷碼及藥物，包含：571.40(Chronic hepatitis，慢性肝炎)、574.20(Calculus of gallbladder without mention of cholecystitis, without mention of obstruction，膽囊結石未提及膽囊炎)、A05BA03(FRUCTUS CARDUI MARIAE EXTRACT，倍利肝膠囊) 肝硬化、571.5 (Cirrhosis of liver without mention of alcohol，慢性肝病及肝硬化)，PBDI 在發展的時候，有依台灣狀況加入本地常見的肝疾病，此主題驗證肝疾病為本地重要的預測特徵值。

最後，是 topic 143，此主題內容看似與住院相關藥物，包含：B05XA03 (SODIUM CHLORIDE，生理食鹽水注射液)、J01DB04(CEFAZOLIN，賜爾寧注射劑)細菌引起之感染症、M01AB15(KETOROLAC TROMETHAMINE，抑痛能靜脈注射液)疼痛之短期療法等，因注射藥物多使用於醫院內，推測為曾入院的患者，會入院的患者健康狀況通常較不好，因此也有較高的健康風險，因此未來再入院的可能性也高。

■ 兩年後住院：

兩年後住院的部分，年齡的影響比找出的主題來得大，係數最高的四個主題分別為 55 歲以上的四個高齡群。其餘疾病相關的主題主要有以下兩個。

第一個是 topic 87，內容多為年長者常見疾病及用藥，例如：564.0(Constipation，便秘)、A06AB06(SENNOSIDEs，仙塞落糖衣錠，治療緩解便秘)、N06BX03(PIRACETAM，腦福膠囊，治療治療腦血管障礙及老化所引起的智力障礙)、N04BA02(MADOPAR，美道普錠，治療巴金森氏症)、

332.0(Paralysis agitans，續發性巴金森病態)、290.0(Senile dementia, uncomplicated，老年癡呆症)等，這些疾病好發於年長者，此結果與合併症指標內容和藥物及診斷碼特徵值相似，隨著年齡增長銀髮族的保健需多注意。

第二個是 Topic 77，此主題與精神性疾病相關，內容與次年住院的精神性疾病主題用藥及診斷碼相似，精神性疾病除了是次年住院的重要因素，更是會影響到兩年後的健康，與慢性疾病一樣影響較長遠，因此需多加注意。

■ 兩年後長期住院：

在預測兩年後長期住院，許多主題與預測次年長期住院的結果相似，包含慢性疾病、精神性疾病、曾經住院，例如：topic 207(慢性肺病)、 topic 232 (慢性肝病)、topic 157(精神性疾病)、topic 202(再入院)等。

比較不同的是以下的主題，topic 128 內容為癱瘓相關的診斷碼和用藥，包含：344.1(Paraplegia，下身麻痺，截癱)、 707.0(Pressure ulcer，糖尿病伴有下肢潰瘍)、 907.2 (Late effect of spinal cord injury，顱內損傷之後遺症)、952.9 (Unspecified site of spinal cord injury without evidence of spinal bone injury，脊髓病灶損傷)、 536.9(Unspecified functional disorder of stomach，胃腸功能性障礙)、 A03FA01(METOCLOPRAMIDE HCL MONOHYDRATE，胃斯妥持效膠囊，治療消化器機能異常)，除了意外傷害，年長者因中風、糖尿病等疾病導致癱瘓，身體的健康程度也隨之下降，且癱瘓患者的需要較多的照料，若照料不佳可能導致健康狀況惡化，引發其他併發症，因此未來長期住院的機率也較大。

■ 小結：

從 LDA 主題中可觀察到，慢性疾病和精神性疾病也是影響住院的重要因素，針對兩年後住院，年齡的影響較大，而影響住院的主題也和年齡有關，例如：年長者常見疾病的主題(內容包含：巴金森氏症、老化所引起的智力障礙及年齡老化智力障礙等)。

PBDI 在發展時，依照台灣地區的常見疾病調整合併症內容，加入的亞洲地區常見的肝炎，LDA 主題中也出現了肝疾病，驗證 PBDI 的調整是適當的，未來若其他地區想依照當地疾病特性調整合併症指標內容時，或許可以透過 LDA 尋找相關疾病主題。

此外，從主題中也可以發現一些非特定疾病，卻是未來住院風險的族群，例如：當年度曾住院、癱瘓等，這些族群的健康風險也相對高，在醫療照護需要更多的關照。

第五章　結論

5.1 實驗結論與貢獻

本研究希望透過不同的特徵值及模型，改善未來健康風險的預測。就預測的表現而言，合併症指標和機器學習方法找出的特徵值，都具有預測未來住院的能力，可以當作個體及總體未來健康規劃的參考。

首先從合併症指標的結果來看，PBDI 大多的表現略優於 Deyo/CCI，原因可能為 PBDI 在發展的時候，考慮到亞洲常見的疾病，依據地區性做了調整，且包含較多的疾病，因此比 Deyo/CCI 符合台灣的狀況。而結合 Deyo/CCI 和 PBDI 當預測特徵值時，互補所缺，比單獨使用其中一種特徵值表現來得佳。

機器學習方法主要分成兩部分：特徵值選取和 LDA Topic，兩者在預測未來住院的能力，皆優於合併症指標，因此，本研究認為機器學習方式所找出的特徵值，具有預測未來住院的能力。

根據文獻說明，各地區的疾病特性有所差異，因此不同地區使用合併症指標預測未來健康風險的表現也有所不同，而本研究使用台灣健康保險資料庫，透過機器學習方式選取出的特徵值，會符合台灣當地的疾病特性，解決合併症指標因地區性疾病差異影響預測結果的問題，推測這也是預測表現優於合併症指標的原因之一。

此外，疾病是不斷的演化的，近期的合併症指標也加入了近代文明病，包括肥胖、憂鬱症等，透過機器學習方式找出影響未來健康的特徵值，可盡快得知近期影響未來健康風險的疾病。因此，使用地區性資料並透過機器學習方式尋找預測特徵值，可當作地區性合併症指標調整參考或預測特徵值，並加速了解疾病與未來健康風險的關聯性。

5.2 未來研究方向

本研究證明了兩點，第一、結合診斷及藥物的合併症指標能提升預測能力，第二、機器學習方式找出的合併症指標能用於預測未來住院機率。因此未來的研究，可以朝以下方向進行。

在合併症指標的部分，近期診斷基礎的合併症指標有一些新版本，例如：Elixhauser，此合併症指標包含的疾病類型較為廣泛，文獻中也說明修訂後的合併症指標預測結果比常用的 Deyo/CCI 來的好，因此可嘗試結合較新版本的以診斷基礎合併症指標，觀察模型的預測表現。

在機器學習的方面，主要使用的特徵值為疾病診斷碼及藥物代碼，目前以二元方式表達：是否有得病、是否有用藥，在 LDA 的前處理也只串聯患者被診斷出的疾病及使用的藥物。因此，未來可嘗試加入診斷碼出現的次數及藥物的用量，透過更完整的資料提升模型的預測表現。此外，除了預測未來是否住院之外，也可以嘗試預測未來住院天數或其他健康風險，提供患者更多的資訊參考。

參考文獻

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. the Journal of machine Learning research, 3, 993-1022.

[2] Chu, Y. T., Ng, Y. Y., & Wu, S. C. (2010). Comparison of Different Comorbidity Measures for Use with Administrative Data in Predicting Short-and Long-term Mortality. BMC health services research, 10(1), 140.

[3] Dong, Y. H., Chang, C. H., Shau, W. Y., Kuo, R. N., Lai, M. S., & Chan, K. A. (2013). Development and Validation of a Pharmacy‐Based Comorbidity Measure in a Population‐Based Automated Health Care Database. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 33 (2), 126-136.

[4] Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity Measures for Use with Administrative Data. Medical care, 36 (1), 8-27.

[5] Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., & Szolovits, P. (2014, August). Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining , 75-84, ACM.

[6] Lemke, K. W., Weiner, J. P., & Clark, J. M. (2012). Development and Validation of a Model for Predicting Inpatient Hospitalization. Medical care, 50 (2), 131-139.

[7] Mao, Y., Chen, W., Chen, Y., Lu, C., Kollef, M., & Bailey, T. (2012, August). An Integrated Data Mining Approach to Real-time Clinical Monitoring and Deterioration Warning. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 1140-1148. ACM.

[8] Parker, J. P., McCombs, J. S., & Graddy, E. A. (2003). Can Pharmacy Data Improve Prediction of Hospital Outcomes?: Comparisons with a Diagnosis-based Comorbidity Measure. Medical care, 41 (3), 407-419.

[9] Putnam, K. G., Buist, D. S., Fishman, P., Andrade, S. E., Boles, M., Chase, G. A., ...& Chan, K. A. (2002). Chronic Disease score as a Predictor of Hospitalization. Epidemiology, 13 (3), 340-346.

[10] Saria, S., McElvain, G., Rajani, A. K., Penn, A. A., &Koller, D. L. (2010). Combining Structured and Free-text Data for Automatic Coding of Patient Outcomes. In AMIA Annual Symposium Proceedings, Vol. 2010, 712. American Medical Informatics Association.

[11] Von Korff, M., Wagner, E. H., & Saunders, K. (1992). A Chronic DiseaseScore from Automated Pharmacy Data. Journal of clinical epidemiology, 45 (2), 197-203.

[12] 朱育增, &吳肖琪. (2010). 回顧與探討次級資料適用之共病測量方法. 臺灣公共衛生雜誌, 29 (1), 8-21.

附錄 A

表 A.1 預測兩年後住院診斷碼特徵值

	2005		2006	
	icd9	症狀	icd9	症狀
1	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)
2	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)
3	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)
4	780.4	Dizziness and giddiness(眩暈)	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)
5	564.0	Constipation(便祕)	564.0	Constipation(便祕)
6	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)
7	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)	780.4	Dizziness and giddiness(眩暈)
8	414.9	Chronic ischemic heart disease, unspecified(缺血性心臟病)	414.9	Chronic ischemic heart disease, unspecified(缺血性心臟病)
9	366.10	Senile cataract, unspecified(老年性白內障)	599.0	Urinary tract infection, site not specified(泌尿道感染)
10	780.52	Insomnia, unspecified(失眠)	585	Chronic kidney disease(慢性腎衰竭)
11	533.90	Peptic ulcer of unspecified site, unspecified as acute or chronic, without mention of hemorrhage or perforation, without mention of obstruction(消化性潰瘍)	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)
12	599.0	Urinary tract infection, site not specified(泌尿道感染)	600.0	Hypertrophy (benign) of prostate(良性攝護腺肥大前列腺增生)
13	585	Chronic kidney disease(慢性腎衰竭)	414.01	Coronary atherosclerosis of native coronary artery(冠狀動脈粥樣硬化)
14	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)	486	Pneumonia, organism unspecified(肺炎)
15	486	Pneumonia, organism unspecified(肺炎)	438.9	Unspecified late effects of cerebrovascular disease (中風後遺症;腦血管疾病後遺症;陳舊性腦血管意外)

表 A.2 預測兩年後長期住院診斷碼特徵值

	2005		2006	
	icd9	症狀	icd9	症狀
1	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)	401.9	Essential hypertension, unspecified(本態性高血壓;自發性高血壓)
2	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)	250.00	Diabetes mellitus without mention of complication(糖尿病;第二型糖尿病)
3	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)	V04.8	Need for prophylactic vaccination and inoculation against other viral diseases(流行性感冒疫苗接種)
4	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)	496	Chronic airways obstruction, not elsewhere classified(慢性阻塞性肺疾病)
5	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)	564.0	Constipation(便祕)
6	564.0	Constipation(便祕)	486	Pneumonia, organism unspecified(肺炎)
7	491.21	Obstructive chronic bronchitis with (acute) exacerbation(阻塞性慢性支氣管炎，併急性發作)	438.9	Unspecified late effects of cerebrovascular disease (中風後遺症;腦血管疾病後遺症;陳舊性腦血管意外)
8	486	Pneumonia, organism unspecified(肺炎)	402.90	Unspecified hypertensive heart disease without conge(高血壓性心臟病)
9	438.9	Unspecified late effects of cerebrovascular disease (中風後遺症;腦血管疾病後遺症;陳舊性腦血管意外)	491.21	Obstructive chronic bronchitis with (acute) exacerbation(阻塞性慢性支氣管炎，併急性發作)
10	366.10	Senile cataract, unspecified(老年性白內障)	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)
11	780.52	Insomnia, unspecified(失眠)	599.0	Urinary tract infection, site not specified(泌尿道感染)
12	780.4	Dizziness and giddiness(眩暈)	414.9	Chronic ischemic heart disease, unspecified(缺血性心臟病)
13	414.9	Chronic ischemic heart disease, unspecified(缺血性心臟病)	782.3	Edema(水腫)
14	428.0	Congestive heart failure, unspecified(鬱血性;充血性心臟衰竭)	585	Chronic kidney disease(慢性腎衰竭)
15	295.90	Unspecified schizophrenia, unspecified(精神分裂症)	780.4	Dizziness and giddiness(眩暈)

表 A.3　2006 年預測次年住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
2	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
3	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
4	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等;可預防心臟病及腦中風之發生。
5	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
6	A03FA01	METOCLOPRAMIDE HCL MONOHYDRATE(胃斯妥持效膠囊)	消化器機能異常(噁心、嘔吐、腹部膨滿感)
7	N02BE01	ACETAMINOPHEN (普除痛錠)	解熱、止痛(關節痛、肌肉痛、風濕痛、神經痛、月經痛、牙痛、頭痛之舒緩)。
8	A02AX	STACAINE(胃必康)	急慢性胃炎、胃痛、胃灼熱、胃酸過多、胃部不適感
9	A03AX13	DIMETHYLPOLYSILOXANE(瓦斯康錠)	適合空氣嚥下症及腹部具有膨滿感、鼓腸等。
10	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
11	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
12	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
13	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
14	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
15	N07CA91	DIPHENIDOL HCL(敵芬尼朵糖衣錠)	止吐
16	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓

表 A.4　2007 年預測次年住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
2	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
3	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
4	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
5	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
6	C08CA01	AMLODIPINE (脈優錠)	高血壓、心絞痛。
7	A03FA01	METOCLOPRAMIDE HCL MONOHYDRATE(胃斯妥持效膠囊)	消化器機能異常(噁心、嘔吐、腹部膨滿感)
8	N02BE01	ACETAMINOPHEN (普除痛錠)	解熱、止痛(關節痛、肌肉痛、風濕痛、神經痛、月經痛、牙痛、頭痛之舒緩)。
9	A02AX	STACAINE(胃必康)	急慢性胃炎、胃痛、胃灼熱、胃酸過多、胃部不適感
10	A03AX13	DIMETHYLPOLYSILOXANE(瓦斯康錠)	適合空氣嚥下症及腹部具有膨滿感、鼓腸等。
11	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
12	N07CA91	DIPHENIDOL HCL(敵芬尼朵糖衣錠)	止吐
13	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
14	M01AB15	KETOROLAC TROMETHAMINE(克多炎注射液)	疼痛之短期療法
15	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠

表 A.5　2006 年預測次年長期住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
2	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
3	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
4	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
5	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
6	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
7	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠
8	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
9	N03AE01	CLONAZEPAM(癲可錠)	癲癇
10	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
11	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
12	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
13	N04AA01	TRIHEXYPHENIDYL HCL(顫立靜錠)	帕金森氏症
14	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
15	M01AC06	MELOXICAM(美樂錠)	類風濕性關節炎、骨關節炎及僵直性脊椎炎之症狀治療

表 A.6　2007 年預測次年長期住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
2	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
3	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
4	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
5	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
6	C08CA01	AMLODIPINE (脈優錠)	高血壓、心絞痛
7	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
8	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
9	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠
10	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
11	N03AE01	CLONAZEPAM(癲可錠)	癲癇
12	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
13	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
14	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
15	C01DA14	ISOSORBIDE 5-MONONITRTATE(寬心持續性藥效錠)	預防狹心症之發作

表 A.7　2005 年預測兩年後住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
2	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
3	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
4	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
5	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
6	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
7	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
8	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
9	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
10	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
11	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
12	A03FA01	METOCLOPRAMIDE HCL MONOHYDRATE(胃斯妥持效膠囊　)	消化器機能異常(噁心、嘔吐、腹部膨滿感)
13	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠
14	N07CA91	DIPHENIDOL HCL(敵芬尼朵糖衣錠)	止吐
15	C01DA14	ISOSORBIDE 5-MONONITRTATE(寬心持續性藥效錠)	預防狹心症之發作。

表 A.8　2006 年預測兩年後住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
2	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
3	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
4	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
5	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
6	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
7	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
8	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
9	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
10	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
11	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
12	A03FA01	METOCLOPRAMIDE HCL MONOHYDRATE(胃斯妥持效膠囊　)	消化器機能異常(噁心、嘔吐、腹部膨滿感)
13	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠
14	N07CA91	DIPHENIDOL HCL(敵芬尼朵糖衣錠)	止吐
15	C01DA14	ISOSORBIDE 5-MONONITRTATE(寬心持續性藥效錠)	預防狹心症之發作。

表 A.9　2005 年預測兩年後長期住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。
2	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
3	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
4	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
5	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生。
6	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
7	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
8	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠
9	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
10	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
11	N03AE01	CLONAZEPAM(癲可錠)	癲癇
12	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
13	N04AA01	TRIHEXYPHENIDYL HCL(顫立靜錠)	帕金森氏症
14	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓
15	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充。

表 A.10　2006 年預測兩年後長期住院藥物特徵值

	ATC Code	藥物學名與中文商品名	適應症
1	A06AB06	SENNOSIDEs(便通樂膜衣錠)	便秘
2	B05XA03	SODIUM CHLORIDE(生理食鹽水注射液)	水分及電解質之補充
3	C03CA01	FUROSEMIDE(樂泄靜脈注射液)	利尿、高血壓、急性肺水腫
4	A02AA02	MAGNESIUM OXIDE(氧化鎂錠)	緩解胃部不適或灼熱感、或經診斷為胃及十二指腸潰瘍、胃炎、食道炎所伴隨之胃酸過多
5	B01AC06	ASPIRIN(暢血脈腸溶微粒膠囊)	因感冒引起的頭痛發燒、肌肉痛、關節痛以及各種疼痛等；可預防心臟病及腦中風之發生
6	N05CF02	ZOLPIDEM HEMITARTRATE(舒眠諾思膜衣錠)	失眠
7	N05BA06	LORAZEPAM(悠然錠)	焦慮症狀
8	A06AB02	BISACODYL(秘瀉樂腸溶糖衣錠)	便秘
9	N03AE01	CLONAZEPAM(癲可錠)	癲癇
10	N05CD04	ESTAZOLAM(艾斯樂錠)	失眠
11	A10BA02	METFORMIN HCL(克醣錠)	第二型糖尿病
12	B01AC07	DIPYRIDAMOLE(保心丁注射液)	狹心症
13	N04AA01	TRIHEXYPHENIDYL HCL(顫立靜錠)	帕金森氏症
14	N05AH04	QUETIAPINE FUMARATE(安保思樂錠)	精神分裂症、雙極性疾患之躁症發作
15	C08CA05	NIFEDIPINE(壓達能軟膠囊)	高血壓