李彥賢、賴家玄、蔡佳玲（2017），『以健保資料庫建構頭頸癌併發吸入性肺炎高風險病患之預測模式』，*中華民國資訊管理學報*，第二十四卷，第三期，頁 341-367。

以健保資料庫建構頭頸癌併發吸入性肺炎高風險病患之預測模式

李彥賢*
嘉義大學資訊管理學系

賴家玄
嘉義長庚紀念醫院放射腫瘤科

蔡佳玲
鴻海精密工業股份有限公司

摘要

預防醫學是指以預防疾病的發生，來代替對疾病的治療，其主要目標在於健康的促進以及疾病的預防，藉由讓民眾增加對疾病的認知、改變態度，用預防的概念來管理健康。近年來隨著人口結構與疾病型態的轉變，使得預防醫學逐漸受到重視。根據台灣衛福部 2014 年統計，頭頸癌死亡率在所有癌症中排名第五。頭頸癌的治療方式根據病人狀況通常包含手術、放射治療及化學治療，然而相關治療的後遺症或腫瘤位置的因素，往往引起患者吞嚥的問題而導致嗆咳，嚴重者更會併發吸入性肺炎。根據研究，頭頸癌若併發吸入性肺炎，在 12 個月內的死亡率將近 10%。過去研究雖指出頭頸癌併發吸入性之可能影響因素，但各研究間觀察的變數不同，且研究結果略有差異，而實務上亦仍未建立評估準則可供醫師評估病患。本研究期望能基於健保申報資料，利用資料探勘中分類學習技術，試圖建構預測模式來協助預測頭頸癌併發吸入性肺炎之高風險病患，以期能給予病患適當之衛生教育，預防吸入性肺炎或及早發現相關症狀，以降低患者的死亡風險及相關醫療成本。實驗評估結果顯示，用以建立訓練資料的抽樣方式明顯影響分類器效能，而從整體學習方法的預測效能來看，Boosting 方法在一般資料情況下預測效能優於 Bagging 方法；而 Bagging 方法效能差異，取決於採用的基礎學習演算法，其中以 Decision Tree 方法最佳。儘管如此，本研究評估之五種演算法皆達成相當不錯之預測效能，而以 RBF-Kernel SVM 為基礎學習演算法之 Bagging 方法更是對訓練資料外的非目標類別資料（未併發吸入性肺炎之頭頸癌病患），有相當好的預測效能。

關鍵詞：頭頸癌、吸入性肺炎、國民健康保險資料庫、傾向分數配對、整體學習演算法

* 本文通訊作者。電子郵件信箱：yhlee@mail.ncyu.edu.tw
2016/04/28 投稿；2017/01/25 修訂；2017/03/03 接受

Lee, Y.H., Lai, C.H. and Cai, J.L. (2017), 'A prediction model for head and neck cancer patient complicated with aspiration pneumonia', *Journal of Information Management*, Vol. 24, No. 3, pp. 341-367.

A Prediction Model for Head and Neck Cancer Patient Complicated with Aspiration Pneumonia

Yen-Hsien Lee*
Department of Management Information Systems, National Chiayi University

Chia-Hsuan Lai
Department of Radiation Oncology, Chiayi Chang Gung Memorial Hospital

Jia-Ling Cai
Hon Hai Precision Industry Co., Ltd.

Abstract

Purpose－The treatment-related adverse effects of head and neck cancer and/or the anatomic location of tumors are likely to cause swallowing problems that might lead to the complications such as choking, malnutrition, and aspiration pneumonia. Prior research indicated the 12-month death rate of head and neck cancer patient with the complication of aspiration pneumonia is nearly 10%. The factors that cause the complication of aspiration pneumonia have been observed in prior studies but inconclusive. This study aims to discover Taiwan's National Health Insurance Research Database, the most comprehensive records of medical insurance claim in Taiwan, to construct a prediction model for the head and neck cancer patients who are at risk of aspiration pneumonia.

Design/methodology/approach － We reviewed the literature to identify a collective set of thirteen factors, which are relevant to the head and neck cancer patients with the complication of aspiration pneumonia and whose data values are available in Taiwan's National Health Insurance Research Database, and adopted them as independent variables. We used propensity score matching to create training dataset and implemented bagging-based and boosting-based ensemble learning methods with

* Corresponding author. Email: yhlee@mail.ncyu.edu.tw
2016/04/28 received; 2017/01/25 revised; 2017/03/03 accepted

different learning algorithms to construct prediction models.

Findings ─ The results suggested that the five investigated approaches were effective in predicting the head and neck cancer patients at risk of aspiration pneumonia. The prediction performances achieved by boosting-based ensemble learning methods were better than bagging-based ones. Overall, the proposed approach can be promising to the construction of prediction model for the head and neck cancer patients with higher risk of aspiration pneumonia using Taiwan's National Health Insurance Research Database.

Research limitations/implications ─ This study applies ensemble learning to construct the prediction model for predicting the head and neck cancer patients at risk of aspiration pneumonia. The evaluation results reveal the effectiveness and the practicability of the proposed method, which builds the prediction model based on health insurance database. This study has contributed to the research area of health data mining. Nevertheless, the independent variables used to construct the prediction model are limited to the records of medical insurance claim. Future research is suggested to incorporate other data sets, such as medical records into the construction of prediction models.

Practical implications ─ The proposed method can be developed into a decision support system to support physicians in assessing the head and neck cancer patients who are at risk of aspiration pneumonia. Such patients can be well educated in advance to prevent the occurrence of aspiration pneumonia. The development of such system is feasible because the records of the medical insurance claim required for constructing the prediction model are ready available.

Originality/value ─ This study investigated the factors that may cause the complication of aspiration pneumonia, thereby constructing a prediction model based on the health insurance database to predict the head and neck cancer patients who are at risk. We developed a method for database preprocessing, training dataset creation, and prediction model construction. The evaluation results suggested practicability and effectiveness of the proposed method.

Keywords: head and neck cancer, aspiration pneumonia, National Health Insurance Research Database, propensity score matching, ensemble learning

壹、緒論

台灣社會的高度工業化發展，已促使人口結構、生活型態、與疾病型態的轉變。根據衛生福利部歷年來公布的國人十大死因中可以發現，台灣地區的疾病型態已由傳統的急性疾病與傳染病，轉變為癌症和慢性疾病。癌症或慢性疾病往往需要長期醫療資源投入，然而在現行制度及醫療資源日益緊縮的情況下，勢必改變醫療院所提供醫療服務的方式。近年來預防醫學逐漸受到相當程度的重視，所謂預防醫學是指以預防疾病的發生，來代替對疾病的治療。預防醫學的主要目標在於健康促進以及疾病預防，藉由增加民眾對於疾病的認知並改變其態度，以預防的概念來管理自身健康。事實上許多疾病是可以預防的，然而如何針對適當的對象，提供適當的健康資訊服務，仍是有待處理的問題。

頭頸癌是一種可治癒的局部惡性腫瘤，位於頭部以及頸部，除了腦癌以外的其他頭頸惡性腫瘤。頭頸癌包含口腔癌、鼻咽癌、口咽癌、下咽癌、喉癌、鼻竇癌、唾液腺癌，研究顯示罹患頭頸癌的主要危險因素包含吸煙和酗酒等（Dubray-Vautrin et al. 2015）。根據台灣衛生福利部 2014 年統計顯示惡性腫瘤為國人十大死因之首，而其中口腔癌、口咽癌、及下咽癌在惡性腫瘤中排名第五，僅次於肺癌、肝癌、大腸癌、以及女性乳癌，並有逐年增加的趨勢[1]。美國亦有數據顯示，頭頸癌病患佔了所有癌症病患的 3%，如 2012 年被診斷出頭頸癌的患者超過五萬兩千人，而每年因頭頸癌死亡的人數就有將近八千人（Siegel et al. 2012）。儘管頭頸癌從早期發展到中期平均僅需 3 至 6 個月的時間，然而只要在此時段內接受適當的治療，其痊癒率可達到 80%。頭頸癌的治療方式根據病人狀況通常包含手術、放射治療（Radio-Therapy; RT）、及同步放化療（Concomitant Chemo-Radio-Therapy; CCRT）（Chu et al. 2013）。儘管頭頸癌病患得以痊癒，然而相關治療或腫瘤位置，往往讓病患面臨吞嚥問題。吞嚥問題容易造成頭頸癌患者營養不良，更會造成病患咳嗽、噎到、吞嚥時感覺疼痛、或喉乾等症狀並進而引發嗆咳（Mittal et al. 2003），更甚者則因將食物吸入肺部，導致吸入性肺炎。根據研究指出，許多頭頸癌病患在治療前後都曾有將食物吸入至肺部的情況發生（Rosen et al. 2001）。頭頸癌患者因放化療後而併發吸入性肺炎，在 12 個月內死亡的比率將近 10%（Eisbruch et al. 2002）。吸入性肺炎是一種潛在威脅生命的併發症，部分頭頸癌患者的死亡可能是由併發吸入性肺炎所引起，而非腫瘤本身的因素。

過去研究利用統計分析方式指出性別、年齡、罹患腫瘤的部位、治療方式、醫院等級、共病等因素，可能為影響頭頸癌患者併發吸入性肺炎之相關因素（Chu et al. 2013; Langerman et al. 2007; Mortensen et al. 2013; Xu et al. 2015）。另

1 http://www.mohw.gov.tw/news/531349778

外有研究顯示社經地位與癌症的罹患率及照護方式有關，社經地位與癌症罹患率成反比，與照護方式成正比（Ward et al. 2004）。上述研究說明特定族群的頭頸癌患者，可能具有較高併發吸入性肺炎的風險。儘管如此，由於吸入性肺炎在臨床表現中可能沒有症狀，或者容易造成乾咳、呼吸急促、支氣管痙攣、血性或泡沫痰、呼吸窘迫等（Marik 2001），因而若能提醒高風險之患者在居家照護時注意預防或及早發現相關症狀，將有助於將低病患風險以及醫療成本。然而過去各研究之間所觀察之影響因素多有不同，且所得之結果略有差異。此外目前在臨床上仍無發展出相關指標，可供醫師評估患者是否為併發吸入性肺炎的高風險群，以適時提供適當的衛生教育，來協助病患或家屬於居家照護時，能夠預防吸入性肺炎的發生。

台灣自 1995 年開辦國民健康保險，由於納保率達到 99%以上，且除醫療院所及被保險人基本資料外，已累積 20 餘年的醫療給付的相關資料，使得健保資料在醫藥衛生相關領域的研究中，具有實證資料的代表性。另一方面，資料探勘技術主要用以分析、處理大量資料，並從中發掘隱藏地、有用的資訊，來協助企業或組織進行決策或預測。資料探勘在醫療方面已有許多應用被提出，例如透過對病患資料的分析及處理，來改善像是病患照護、降低費用等醫療問題（Delgado et al. 2001），或是藉由探勘醫療資料來輔助醫療上面的決策（Zorman et al. 2001）。由於過去研究所指出頭頸癌患者可能併發吸入性肺炎之相關因素大多為國民健康保險研究資料庫所涵括，因此本研究認為應能以國民健康保險研究資料庫為基礎，來建立頭頸癌患者併發吸入性肺炎高風險群之預測模式。本研究主要目的則是期望能協助預測經治療後之頭頸癌病患是否為併發吸入性肺炎之高風險群，以利醫療院所能夠及早施與預防性之衛生教育，來降低日後病患併發吸入性肺炎的機率，以及可能引起之相關治療成本。

本研究內容後續編排如下，在第貳節中將會回顧與本研究相關之文獻，包括回顧頭頸癌併發吸入性肺炎相關因素之文獻、介紹本研究使用之國民健康保險研究資料庫、以及說明本研究將採用資料探勘中分類技術。在第參節會說明本研究建立頭頸癌併發吸入性肺炎高風險群預測模式之概念及流程。第肆節則為本研究之資料處理、實驗設計、及評估方法。最後在第伍節中則提出本研究的結論與建議。

貳、文獻回顧

一、頭頸癌併發吸入性肺炎的相關因素

頭頸癌主要是由上呼吸道的黏膜病變而引發的惡性腫瘤，上呼吸道黏膜包含了嘴唇、鼻孔、聲帶、副鼻竇、中耳。由於這些部位皆由鱗狀細胞覆蓋，因此

95%的癌症來自於鱗狀細胞上皮的病變。鱗狀細胞癌主要集中在鼻咽、口腔、口咽、下咽、喉，但仍可能會轉移到頸部淋巴結及遠隔轉移，如肺、骨頭、頭頸部以外的皮膚，而頭頸癌即為這五個部分癌症的統稱。在台灣以口腔癌的比率最高，約佔 60-70%，口咽癌與下咽癌分別約佔 10%，而喉癌則約佔 15%。其中比率最高的口腔癌其致命因子包括有香菸、酒精、檳榔、陽光的曝曬、飲食上維生素 A 與 B 的攝取不足。口腔癌的早期症狀包括形成不會痛的腫塊，或是長期沒有癒合的潰傷，有時喉嚨會有局部性的疼痛感，甚至引發神經痛轉移到耳朵的部位，影響聽力、中耳積水、持續性單側鼻竇炎、鼻塞、流鼻血等。晚期症狀則可能在吞嚥時感到疼痛、吞嚥困難、牙關緊閉、舌頭固定、呼吸道阻塞、視力受損、腦神經病變（複視、眼皮下垂、失明等症狀），另有許多病患是頸部淋巴腺腫大後，才發現罹患頭頸癌（王宏銘等 2009）。

吸入是指口咽或胃裡的內容物被吸入到喉部或下呼吸道（Irwin et al. 1999），吸入性肺炎可能是由於吸入附著在口咽或鼻咽分泌物中的病原體，如流感嗜血桿菌、肺炎鏈球菌等，而引發的感染的過程；或是在沒有病原體的情況下，吸入無菌物質而產生的化學傷害，造成肺部急性的損害（Adnet & Baud 1996）。吸入性肺炎的症狀會視吸入的物質、吸入量、吸入頻率、以及個人對吸入物質的反應，來決定其嚴重程度（Baum et al. 1998）。吸入性肺炎的引發的症狀包含氣道阻塞、肺膿傷、外源類脂性肺炎、慢性間質性纖維化、偶發分枝桿菌肺炎（Irwin et al. 1999）。在臨床表現中，吸入性肺癌可能沒有出現症狀，或者容易造成乾咳、呼吸急促、支氣管痙攣，血性或泡沫痰、呼吸窘迫等（Marik 2001）。吞嚥困難或胃動力的障礙可能使病患發生吸入性肺炎，其中以吞嚥困難為最常見的引發原因，每年就有約有 30 萬至 60 萬美國人，因為吞嚥困難導致吸入性肺炎而死亡（Daniels et al. 1998）。此外經研究發現，因疾病而住院的病患中，亦大約有 10%的病患會在使用藥物過量後併發吸入性肺炎（Roy et al. 1989）。

在文獻中有些許與罹患頭頸癌後併發吸入性肺炎的研究。相關研究較早是由芝加哥大學的 Langerman 等（2007）提出，研究對象為該大學從 1998 年 11 月至 2002 年 8 月所收治的 130 名頭頸癌病患，主要了解病患在接受放射治療後，併發吸入性肺炎的機率。該研究提出幾個可能的因素，來評估頭頸癌併發吸入性肺炎關聯性，像是年齡、性別、腫瘤位置、腫瘤大小、淋巴結狀態、腫瘤階段等，並使用統計方法檢驗相關變數是否顯著。研究結論認為併發吸入性肺炎並不受年齡、腫瘤階段、淋巴結狀態的影響，而腫瘤位置則以喉部和下咽部的影響高於其他部位。其他學者亦有類似的研究，但檢測的影響因素則略有差異。Mortensen 等（2013）檢測的變數包含性別、年齡、位置、腫瘤階段、治療方式（包括放射治療、放化療、術後放射治療、調強放射治療）、管餵與否、治療反應、放射劑量、吞嚥困難程度、共病（中風）、察爾森共病得分，並利用統計方法 X^2-test 及

Kaplan-Meier Curve 進行分析。研究結論認為腫瘤階段、管餵與否、治療反應、吞嚥困難程度，皆顯著影響頭頸癌併發吸入性肺炎。Chu 等（2013）使用的變數包含性別、年紀、醫院等級，治療方式（包括頸淋巴結清除術、胃造口術、同步放化療）、共病（老人痴呆、中風、胃食道逆流、帕金森氏症）、腫瘤位置，並使用 X^2-test、Fisher's exact tests、t-test、logistic regression 等統計方法進行分析。研究結果認為性別、年齡、醫院等級、治療方式（頸淋巴結清除術、胃造口術）、共病（中風、胃食道逆流、帕金森氏症）有顯著影響，另外腫瘤位置則為舌、口腔、口咽、鼻咽、喉有較高影響力。Xu 等（2015）使用變數包含年齡、大都市區與否、教學醫院、美國區域、中位數收入、性別、種族、婚姻狀況、共病得分、診斷年份、腫瘤位置、疾病階段、正電子發射斷層攝影術與否、強調性放射與否、術前放射與否、化療序列、化療類型，並以 Gray's test、Multivariate predictors、Gray regression models 等統計方法進行分析，並得出年齡、性別、教學醫院、腫瘤位置（鼻咽、下咽）對於併發吸入性肺炎有顯著的影響，而區域位置在美國南方則較西方有較低併發吸入性肺炎的機率。另外疾病階段、化療類型、化療順序、強調性放射對於併發吸入性肺炎沒有影響。本研究將相關研究提出可能影響頭頸癌併發吸入性肺炎之因素整理於表 1。

表 1：頭頸癌併發吸入性肺炎影響因素

學者	變數數目	變數名稱
Langerman et al.（2007）	6	年齡、性別、腫瘤位置、腫瘤大小、淋巴結狀態、腫瘤階段
Mortensen et al.（2013）	11	性別、年齡、位置、腫瘤階段、治療方式（放射治療、放化療、術後放射治療、調強放射治療）、管餵與否、治療反應、放射劑量、吞嚥困難程度、共病（中風）、察爾森共病得分
Chu et al.（2013）	6	性別、年紀、醫院等級、治療方式（頸淋巴結清除術、胃造口術、同步放化療）、共病（老人痴呆、中風、胃食道逆流、帕金森氏症）、腫瘤位置
Xu et al.（2015）	17	年齡、大都市區與否、教學醫院、美國區域、中位數收入、性別、種族、婚姻狀況、共病得分、診斷年份、腫瘤位置、疾病階段、正電子發射斷層攝影術與否、強調性放射與否、術前放射與否、化療序列、化療類型

二、國民健康保險研究資料庫

全民健康保險資料庫在學術上的應用備受矚目，主要是由於在全民健保尚未實施之前，由政府主辦的醫療保險業務，包含勞保、公保、農保，其歷年所累積之資料檔案，確實為醫療衛生學術研究方面，提供相當珍貴的研究資源（曾淑芬 1999）。例如 Beasley et al.（1981）便是利用台灣公保歷年的住院醫療檔、死亡給付檔等資料，來進行被保險人追蹤，因而發現 B 型肝炎帶原與罹患肝癌之間有相當大的關聯性。而自 1995 年開辦國民健康保險後，由於納保率達到 99%以上，使得健保資料在做為醫藥衛生相關領域研究的實證資料上，具有相當高程度的代表性，因而以分析健保資料所得出的研究成果，基本上可為醫療衛生政策的參考依據。從現有的文獻可以發現，利用健保研究資料所進行的研究，涵蓋範圍相當廣泛，包含醫療保健上疾病的關係及預防、保險設計、巨量資料的視覺化呈現等，而其中仍以第一項研究居多。此外在醫院管理方面，亦有利用健保資料來提高照護品質及管控成本效益（曾淑芬 1999）。

中央健康保險署每年都會將前一年度可供研究使用的健保資料檔案匯出，將身分欄位加密後，交由國家衛生研究院製作成「國民健康保險研究資料庫」及各類加值資料檔案。健保署提供的資料分為「基本資料檔」與「原始資料檔」兩大類[2]。「基本資料檔」包含醫事機構病床主檔（BED）、醫事機構診療科別明細檔（DETA）、醫事機構基本資料檔（HOSB）、醫事機構副檔資料檔（HOSX）、專科醫師證書主檔（DOC）、醫事人員基本資料檔（PER）、重大傷病證明明細檔（HV）、醫事機構服務項目檔（HOX）、藥品主檔（DRUG）、承保資料檔（ID）、評鑑資料檔（HOSP_GRAD）、醫事機構類別明細檔（HOSTDTL）、執業資料紀錄（LIC）。「原始資料檔」包含住院費用申請總表主檔（DT）、門診費用申請總表主檔（CT）、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）、特約邀局處方醫令檔（GO）、承保資料檔（ID）、物理治療所調劑檔（GDD）、物理治療所醫令檔（GOD）。

三、分類學習技術與整體學習法

資料探勘成為近年來相當熱門的資料分析技術。Frawley、Piatetsky-Shapiro 與 Matheus（1991）提出資料探勘技術（Data Mining），主要目的係從大量資料中，挖掘出其中隱含的、未知的資訊，並轉化成有用的知識以及應用的過程。分類學習技術（Classification）為一種「監督式」的資料探勘技術，從已知類別之

2 http://nhird.nhri.org.tw/date_01.html

資料當中學習或找出其歸類方式、規則，並做為後續推估或預測未知類型資料的基礎。分類學習技術是資料探勘中最普遍應用的技術之一，已有許多相關的分類學習演算法被提出，而較常見的分類學習演算法包含決策樹（Decision Tree，如ID3、C4.5）（Mingers 1989）、簡單貝氏分類器（Native Bayes）（Mitchell 1997）、類神經網路（Neural Network，如倒傳遞類神經網路 Backpropagation Neural Network）（Berry & Linoff 1997）、支援向量機（Support-Vector-Machine, SVM）（Cortes & Vapnik 1995）、K 個最近鄰居法（K-Nearest-Neighbors）（Henley & Hand 1996）等。

整體學習演算法係利用上述基礎學習演算法，針對一個訓練資料集合，建立多個分類器（假說）；在進行新資料類別預測時，則藉由整合各個分類器的投票結果，來決定新資料的最終預測類別（Dietterich 2000）。理論上，整體學習演算法的預測效能基本上優於單一個基礎學習演算法所產生的分類器，其主要理由在於前者能夠解決基礎學習演算法通常會面臨的統計、計算、及代表性問題。統計問題是指基礎學習演算法只能被迫從多個效能皆不錯的假說中，選擇其一做為最終結果，而該假說卻又可能具備某種程度上的偏頗，以致無法準確預測某些新資料；計算問題是指在有限時間內，基礎學習演算法可能無法找到最佳假說；代表性問題則是指若假說空間中不存在好的假說時，基礎學習演算法無法得到具有代表性的假說。Freund 與 Schapire（1996）的研究結果亦指出以決策樹 C4.5 做為基礎學習演算法，在 UCI 的 27 個基準問題中，整體學習演算法在其中的 20 個基準問題上，其預測效能勝過單一分類器。目前來說，最常用來建構分類器的整體學習演算法包括 Bagging 與 Boosting 兩種方法。

為提升弱分類學習演算法建構分類器之效能，Nilsson（1965）提出委員會機器（committee machine）的概念。將同一組訓練資料，利用隨機抽樣方式形成多個訓練資料集合，並用以分別建構分類器組成委員會機器，而在進行類別預測時，則由委員會機器內成員進行共同決策。由於委員會機器中由多個不同面向資料訓練出的分類器所組成，透過整合決策的機制，可以各分類器形成互補之勢，提高整體分類效能。Breiman（1996）修正委員會機器訓練集合的建構方式，改以 Bootstrap 隨機重複抽樣的方式，提出了 Bagging 學習法。從統計角度來看，抽樣後放回的方法較不放回的方式，可建構出多個重疊性小的訓練集合，並訓練出較穩定之委員會機器。

Valiant（1984）在提出的 PAC（Probably Approximately Correct）架構中提到可學習（learnable）的概念，亦即找到一個分類器訓練演算法，使分類器能學習將分類結果逐步達成低錯誤率及高信賴度。Kearns 與 Valiant（1989）定義出弱學習力（weak learnability）的概念，認為挑選建構分類器演算法時，只需選擇能比隨機猜測結果較好一點的演算法，之後則透過學習的方式，來增強弱分類器的分

類效能，並基於此概念提出 Boosting 方法。Freund 與 Schapire（1997）提出 AdaBoost 演算法，其藉由迭代的過程產生多個弱分類器，爾後再藉由委員會的投票機制，產生一個類似強分類器決策，來提升分類準確度。AdaBoost 同樣是以抽取後置回方式產生下一代訓練資料，但是在抽樣過程中加重前代弱分類器分類錯誤之資料，使得分類錯誤資料能有較大的機率被重複抽取，以使下一代弱分類器能夠修正前次資料錯誤分類的問題。在委員會進行決策時，AdaBoost 會依據分類器效能給予不同的決策權重，可降低測試誤差，並避免過適性（Overfitting）問題（Bauer & Kohavi 1999; Dietterich 2000）。

Freund 與 Schapire（1999）研究顯示 AdaBoost 確實能有效將弱學習演算法轉換成為強學習演算法。過去研究顯示，在訓練資料正常的情況下，AdaBoost 的預測效能大抵上優於 Bagging 方法（Bauer & Kohavi 1999; Dietterich 2000; Freund & Schapire 1996）；然而若是訓練資料中包含較多雜訊資料（noise）或是錯標類別（mislabeled）資料時，則會得到相反的結果，亦即 Bagging 方法會優於 Boosting，其主要原因在於 AdaBoost 會將非常高的權重放在雜訊資料上，進而產生一個效能很差的分類器。

總和來說，本研究試圖基於全民健康保險資料庫，並利用資料探勘中分類學習技術來建構頭頸癌併發吸入性肺炎高風險病患之預測模式。本研究綜合過去研究所提出的影響因素，並考量健康保險資料庫可取得的資料，彙整出相關變數資料如表 2 所示，來做為建構分類器的屬性資料。此外過去研究顯示利用整體學習演算法所建構之分類模式能較單一分類器獲得較佳的效能（Freund & Schapire 1996）。整體學習演算法大致可區分為 Bagging 與 Boosting 兩種方法，由於兩種方法的預測效能，在不同特性的資料中互有高低，因此本研究將試圖了解此兩種方法所建構之頭頸癌併發吸入性肺炎高風險病患預測模式的效能差異。

表 2：本研究彙整健保資料庫中可取得之影響因素資料

影響因素	健保資料可取得項目	曾經提出的學者
年齡	年齡	Langerman et al.（2007）、Mortensen et al.（2013）、Chu et al.（2013）、Xu et al.（2015）
性別	性別	Langerman et al.（2007）、Mortensen et al.（2013）、Chu et al.（2013）、Xu et al.（2015）
腫瘤位置	腫瘤位置（口腔、唾液腺、口咽、鼻咽、喉部、鼻腔、下咽）	Langerman et al.（2007）、Mortensen et al.（2013）、Chu et al.（2013）、Xu et al.（2015）

治療方式	管灌膳食與否 放射線診療與否 化學治療與否	Mortensen et al.（2013）、Chu et al.（2013）、Xu et al.（2015）
共病	中風與否 老人癡呆與否 胃食道逆流與否 帕金森病與否	Mortensen et al.（2013）、Chu et al.（2013）、Xu et al.（2015）
醫院等級	醫院等級	Chu et al.（2013）、Xu et al.（2015）
收入狀況（社經地位）	投保金額（依據其收入計算，本研究將投保金額視為反映收入狀況的資料）	Xu et al.（2015）
地區	所在縣市（共 25 個地區）	Xu et al.（2015）

參、頭頸癌併發吸入性肺炎高風險群預測模式建構

本研究主要利用國民健康保險資料庫來建構頭頸癌併發吸入性肺炎高風險群之預測模式。本研究建構分類器的資料為 2002-2012 年健保資料，對象為期間內罹患頭頸癌病患，而在頭頸癌確診後 3 個月內併發吸入性肺炎者，則視為高風險群患者。本研究首先綜合先前研究所提出的相關因素，包括性別、年齡、醫院等級、腫瘤位置、放射治療與否、管餵與否、化學治療與否、共病（老人痴呆、中風、胃食道逆流、帕金森氏症）、所在地區、社經地位（投保金額）。其中因健保資料沒有包含被保險人社經地位資料，但投保金額係依據其收入計算，相當程度上可反應社經地位，本研究因此將投保金額視為社經地位資料。經從健保資料中取出所有研究變數之相關資料後，本研究利用整體學習方法（ensemble learning）來建構高風險病患預測模式。圖 1 為預測模式建構流程圖，其大致分為資料前處理、訓練資料建立、及分類器建構三個階段，以下詳細說明各階段的任務與方法。

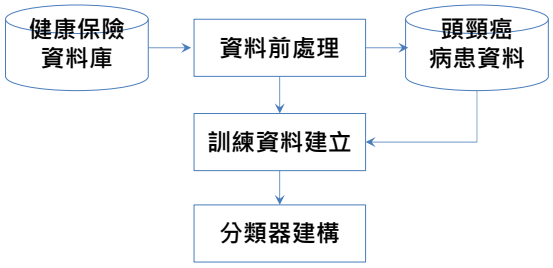

圖 1：預測模式建構流程圖

一、資料前處理

本研究使用 2002-2012 年的國民健康保險資料，由於頭頸癌病患可能會在住院或是門診時確診，因此在確認對象時須使用住院費用清單明細檔（DD）以及門診處方及治療明細檔（CD）進行交叉比對。住院費用清單明細檔主要記錄病患住院的相關資訊，包含醫療機構代碼、病患身份、住院/出院日期、診斷代碼、手術代碼、住院期間的各項醫療費用等等；門診處方及治療明細檔則是記錄病患就診的相關資訊，包含醫療機構代碼、病患身份、就診日期、疾病國際代碼、特殊治療項目等。另外為取得醫療院所及被保險人基本資料，須同時使用醫事機構基本檔（HOSB）及承保資料檔（ID）。醫事機構基本檔包括所有醫療機構的相關資訊，像是醫療機構代碼、機構評鑑等級、機構型態、所在地區、負責醫師代碼、合約相關資料等；承保資料檔包含病患的相關投保資訊，如投保人身份、被投保人身份、投保區域、投保金額、加/退保日期等等。由於健保資料相當繁複，因此本研究將資料前處理過程分成四個步驟，以下分別說明各步驟的目的及處理內容，並簡列於表 3。

表 3：資料前處理步驟

步驟	輸入檔	處理方式	輸出檔/屬性
一	1. 住院費用清單明細檔（DD） 2. 門診處方及治療明細檔（CD） 3. 醫事機構基本檔（HOSB）	1. 分別整合跨 11 年度之住院費用清單明細檔以及門診處方及治療明細檔 2. 透過病患就醫診病日期與出生日期來計算年紀 3. 管餵費用更改為 Y/N 4. 透過醫事機構基本檔取得醫事機構等級 5. 透過 DD 檔中診斷代碼、手術代碼以及 CD 檔中的特殊治療項目取得是否有放射線治療以及化學治療	住院費用清單總檔（DD_all）：身分證統一編號、年紀、性別、入院年月日、診斷代號（五欄）、化學治療與否、放射線診療與否、管灌膳食與否、醫療機構等級 門診處方及治療總檔（CD_all）：身分證統一編號、年紀、性別、就醫日期、國際疾病分類號（三欄）、化學治療與否、放射線診療與否、醫事機構等級
二	1. 住院費用清單總檔（DD_all） 2. 門診處方及治療總檔（CD_all）	1. 利用國際疾病編號分別從 DD_all、CD_all 檔中取得罹患頭頸癌病患資料 2. 利用國際疾病編號分別從 DD_all、CD_all	頭頸癌病患資料總檔（HNC_all）：身分證統一編號、年紀、性別、就醫日期、腫瘤位置、管灌膳食與否、化學治療與否、放射線診療與

		檔中取得罹患吸入性肺炎病患資料 3. 將頭頸癌病患的國際疾病碼轉換成腫瘤罹患位置 4. 分別將頭頸癌與吸入性肺炎病患資料結合並去除重複	否、醫事機構等級 吸入性肺炎病患資料總檔（PN_all）：身分證統一編號、就醫日期
三	1. 頭頸癌病患資料總檔（HNC_all） 2. 吸入性肺炎病患資料總檔（PN_all）	1. 整合頭頸癌病患與吸入性肺炎病患資料總檔，產生頭頸癌病患 90 天內併發吸入性肺炎資料	頭頸癌是否併發吸入性肺炎檔案（HNC_PN_all）：身分證統一編號、年紀、性別、就醫日期、腫瘤位置、管灌膳食與否、化學治療與否、放射線診療與否、醫事機構等級、併發吸入性肺炎與否
四	1. 頭頸癌是否併發吸入性肺炎檔案（HNC_PN_all） 2. 住院費用清單總檔（DD_all） 3. 門診處方及治療總檔（CD_all） 4. 承保資料檔（ID）	1. 利用住院費用清單及門診處方及治療總檔，取得病患併發吸入性肺炎前是否有其他疾病（老人癡呆、中風、胃食道逆流、帕金森病） 2. 利用承保資料檔，取得病患罹患疾病時的投保金額、投保縣市	頭頸癌是否併發吸入性肺炎檔案（HNC_PN_all）：身分證統一編號、年紀、性別、診斷年月日、腫瘤位置、管灌膳食與否、化學治療與否、放射線診療與否、醫事機構等級、併發吸入性肺炎與否、老人癡呆與否、中風與否、胃食道逆流與否、帕金森病與否、投保金額、病患所處地區

第一步驟主要目的在整合個別檔案，將分跨 11 個年度的住院費用清單明細檔（DD）及門診處方及治療明細檔（CD），分別整合成單一檔案，並各自取出本研究所需的欄位資料，最後形成住院醫療費用清單總檔（DD_all），其中包括身分證統一編號、年紀、性別、入院年月日、診斷代號（五欄）、手術代碼（五欄）、管灌膳食與否、醫療機構等級，以及門診處方及治療總檔（CD_all），其中包括身分證統一編號、年紀、性別、就醫日期、國際疾病分類號（三欄）、特殊治療項目（四欄）、醫療機構等級。由於原始資料記錄方式不符合研究需求，因此將某些欄位資料進行調整，例如管餵治療在原始資料中記錄為診療費用，而本研究中只須知道是否有過相關治療，因而將費用欄位資料轉換為有無（Y/N）接

受此項治療。在相關手術方面，需從診斷代碼、手術代碼、與特殊治療項目中，取得是否曾經接受放射線治療與化學治療，其中診斷代碼 V580，手術代碼 922、923 以及特殊治療項目 D1 為放射線治療；診斷代碼 V581，手術代碼 9925 以及特殊治療項目 D2 為化學治療。放射線治療部份，由於放射治療費用與手術代號均有記載，因此只要有放射線治療費用或手術代號記載放射治療記錄者，均視為「有」進行放射線治療。最後，由於 Chu 等（2013）的研究認為醫療機構等級為頭頸癌併發吸入性肺炎的重要影響因素，因此本研究透過明細檔中醫事機構代號與醫事機構基本檔（HOSB）進行交叉比對，來取得病患治療時的醫療機構等級資料。

第二步驟主要目的在取得頭頸癌病患資料，以及吸入性肺炎病患資料，以利第三步驟交叉比對出建構分類器所需之頭頸癌併發吸入性肺炎病患資料。本研究利用國際疾病編號從第一步驟結果中（DD_all 與 CD_all）辨識出頭頸癌病患。頭頸癌在醫學上的國際疾病編號（ICD-9 codes）為 140-149、160、161。由於有些病患會有多次看診資料，例如在多家醫院進行雙重確認或是頭頸癌復發就診，因此本研究依據臨床醫師經驗，將三個月以上的重複就診視為不同個案，反之三個月內則視為同一個案，僅保留最近一次看診記錄。然而由於多數病患在第一次就診時並未進行治療手術，為避免將病患在後續三個月內所接受之相關治療資料篩除，本研究進一步確認並保留病患之前曾進行過之相關手術治療資料。緊接著，本研究依據國際疾病編號比對出病患腫瘤位置，例如口腔部位的編號為 140、141、143、144、145，唾液腺部位的編號為 142，口咽部位的編號為 146、149，鼻咽部位的編號為 147，下咽部位的編號為 148，喉部位的編號為 161，鼻腔部位的編號為 160，並在去除重複病患資料後，形成頭頸癌病患資料總檔（HNC_all），包括身分證統一編號、年紀、性別、就醫日期、腫瘤位置、管灌膳食與否、放射線診療與否、醫事機構等級。同樣的方式被用來取得吸入性肺炎的病患資料，吸入性肺炎在醫學上的國際疾病編號為 480-486、507，最後形成吸入性肺炎患者資料總檔（PN_all），包括身分證統一編號以及就醫日期。

第三步驟主要目的在產生頭頸癌併發吸入性肺炎資料檔。本研究依據第二步驟結果進行交叉比對，來確認頭頸癌病患是否在確診後三個月內又被診斷為吸入性肺炎病患，最後形成頭頸癌病患是否併發吸入性肺炎總檔（HNC_PN_all），包括欄位有身分證統一編號、年紀、性別、就醫日期、腫瘤位置、併發吸入性肺炎與否、管灌膳食與否、放射線診療與否、醫事機構等級。

第四步驟則是持續加入其他研究指出的重要影響因素資料，例如 Chu 等（2013）認為老人癡呆、中風、胃食道逆流、帕金森病等共同病徵會影響頭頸癌併發吸入性肺炎；Ward 等（2004）認為社經地位與癌症的罹患率及照護方式有關，社經地位與癌症罹患率成反比，與照護方式成正比。對此本研究利用第三步

驟產生的頭頸癌病患是否併發吸入性肺炎總檔（HNC_PN_all），與第一階段產生的住院醫療費用清單總檔（DD_all），以及門診處方及治療總檔（CD_all），分別進行交叉比對，來取得頭頸癌病患在併發吸入性肺炎前，是否曾經罹患上述四項會影響併發吸入性肺炎的疾病。根據國際疾病編號，老人癡呆為 290、294，中風為 430-438，胃食道逆流為 530.11、530.81、787.1，帕金森病為 332。最後再與承保資料檔（ID）進行交叉比對，取得病患投保金額以及住所區域代碼或郵遞區號，並將後者依據代碼表轉換成病患居住縣市名稱，形成建構分類器所需的訓練資料集合。

二、訓練資料建立

醫療資料往往存在目標對象與非目標對象數量比例懸殊的情況，若採用全部資料進行分類法則學習時，將使得建構出的分類器偏向預測資料量較大的類別（He & Garcia 2009）。過去研究多採用隨機重覆抽樣（random resampling），包括 Oversampling （Chawla et al. 2002; Lewis & Catlett 1994）以及 Undersampling（Lin et al. 2009; Liu et al. 2009）來平衡訓練資料集合中各類別資料的比例，以避免建構出的分類器會有偏袒預測的情況。儘管如此，隨機重覆抽樣方法仍會面臨樣本過度配適（overfitting）或樣本代表性的問題（Lee et al. 2013）。

在觀察性研究中經常使用受試者配對方式，亦即讓對照組與實驗組具有相似的外來因素，降低受試者選取誤差（selection bias），以方便觀察不同變因對結果的影響。然而並非所有的研究都適用配對方式來形成對照組，因為某些類型的研究並無法將兩組的條件（或稱基準，baseline）控制在相似的情況。例如想觀察中風對頭頸癌併發吸入性肺炎的影響時，並無法控制讓所有受測者均具備中風症狀。因此為降低基準條件（又為干擾因子，confounder）對研究造成干擾，更準確估算變因的影響程度，此類實驗往往會採用傾向分數配對（Propensity ScoreMatching; PSM）方法，抽樣篩選出在某些影響因素下與實驗組最為相近的對照組資料，以降低干擾因子的影響（陳錦華 2014）。傾向分數可以利用邏輯斯迴歸模型（logistic regression model）來估計，將自變項資料（independent variable）放入邏輯斯迴歸模型中，來取得依變項（dependent variable）之預估機率，即為範圍介在 0-1 之間的傾向分數（陳錦華 2014）。計算出各資料之傾向分數後，便可為受試者進行傾向分數配對，Becker and Ichino（2002）提出 PSM 常見的配對方式為最近相鄰配對法（nearest neighbor matching），亦即將每位實驗組受試者配對傾向分數最為相近對照組樣本。根據過去研究指出，利用傾向分數配對可使受試者與其對照樣本之間用以計算傾向分數之共同變數分類相似，據以降低樣本選取誤差（Parsons 2001）。

基於上述，本研究認為若訓練資料中併發與未併發吸入性肺炎之頭頸癌病患描述資料（自變數）能夠相似，將能降低樣本選取誤差，且應有利於分類器歸納出有利區別兩者之顯著特徵樣態。因此，本研究採用傾向分數配對方式，來建立訓練分類器之樣本資料。

三、分類器建構

回顧過去文獻指出整體學習方法的預測效能優於單一分類器（Freund & Schapire 1996），因此本研究決定利用整體學習方法來建構頭頸癌併發吸入性肺癌之高風險病患之預測分類器。基本上，預測整體學習方法為 Bagging 跟 Boosting 兩種方式，過去研究（Bauer & Kohavi 1999; Dietterich 2000; Freund & Schapire 1996）顯示在一般正常資料情況下，Boosting 方法所建構之分類器，其預測效能勝過 Bagging 方法所建構之分類器，然而若資料本身有雜訊的時候，則 Bagging 方法的預測效能反而會優於 Boosting 方法。由於不確定健保資料品質，因此本研究將同時利用兩種方式來建構分類器，並比較兩者的其預測效能的差異。此外 Bagging 的預測效能，過去研究認為 Bagging 的預測效能與其採用之基礎學習演算法有關，不穩定之學習演算法（unstable learning algorithm）其預測效能越佳，亦即當訓練資料有小改變時會大幅改變學習結果之演算法，例如決策樹（decision tree）、類神經網路（neural network）、法則學習演算法（rule learning algorithm）等，其他如線性回歸（linear regression）、最近鄰居法（nearest neighbor）等則屬穩定學習演算法（Dietterich 2000）。據此本研究在 Bagging 方法的基礎學習演算法，將同時比較以機率為主的 Naïve Bayes、決策樹 C4.5、以及非線性核心函數之 RBF-Kernel SVM。Boosting 方法則採用著名的 AdaBoost，以及改良之 LogitBoost，來進行預測效能比較。

肆、實驗評估

本章將說明本研究設計的實證評估方式，包含評估資料描述、實驗設計與評估準則，實驗程序。以下將依序說明。

一、實驗評估資料描述

本研究的實驗評估資料採用國民健康保險 2002-2012 年之 100 萬承保抽樣歸人檔，主要使用的資料檔案有住院費用清單明細檔（DD）、門診處方及治療明細檔（CD）、醫事機構基本檔（HOSB）、承保資料檔（ID）[3]。經本研究資料前處

3 本研究部分資料來源為衛生福利部中央健康保險署提供、財團法人國家衛生研究院管理之「全民健康保

理後共取得罹患頭頸癌病患 17,725 筆，頭頸癌併發吸入性肺炎病患 427 筆，資料中包括性別、年齡、腫瘤位置、管灌膳食與否、放射線診療與否、化學治療與否、醫事機構等級、老人癡呆與否、中風與否、胃食道逆流與否、帕金森病與否、投保金額、病患所處地區等 13 項資料，該資料統計敘述如下，表 4 為實驗評估變項資料分佈。

表 4：實驗評估資料變項分佈(併發/未併發吸入性肺炎)

<table>
<tr><td colspan="10">(a) 二元類別資料</td></tr>
<tr><td colspan="4"></td><td colspan="3">是（男）</td><td colspan="3">否（女）</td></tr>
<tr><td colspan="4">性別</td><td colspan="3">342/13,014</td><td colspan="3">85/4,284</td></tr>
<tr><td colspan="4">管灌膳食</td><td colspan="3">33/207</td><td colspan="3">394/17,091</td></tr>
<tr><td colspan="4">放射線診療</td><td colspan="3">107/1,733</td><td colspan="3">320/15,565</td></tr>
<tr><td colspan="4">化學治療</td><td colspan="3">54/1,135</td><td colspan="3">373/16,163</td></tr>
<tr><td colspan="4">老人癡呆</td><td colspan="3">4/6</td><td colspan="3">423/17,292</td></tr>
<tr><td colspan="4">中風</td><td colspan="3">22/50</td><td colspan="3">405/17,248</td></tr>
<tr><td colspan="4">胃食道逆流</td><td colspan="3">0/0[4]</td><td colspan="3">427/17,298</td></tr>
<tr><td colspan="4">帕金森病</td><td colspan="3">2/1</td><td colspan="3">425/17,297</td></tr>
<tr><td colspan="10">(b) 腫瘤位置資料分佈</td></tr>
<tr><td colspan="2">口腔</td><td colspan="2">唾液腺</td><td colspan="2">口咽</td><td colspan="2">鼻咽</td><td>喉部</td><td>鼻腔</td><td>下咽</td></tr>
<tr><td colspan="2">125/6,410</td><td colspan="2">12/743</td><td colspan="2">45/1,116</td><td colspan="2">153/6,608</td><td>41/1,449</td><td>7/380</td><td>44/589</td></tr>
<tr><td colspan="11">(c) 醫院等級資料分佈</td></tr>
<tr><td>等級</td><td>1</td><td>2</td><td>3</td><td>4</td><td>6</td><td>7</td><td>8</td><td>13</td><td>22</td><td>24</td><td>43</td></tr>
<tr><td>數量</td><td>357/15346</td><td>36/528</td><td>0/5</td><td>0/2</td><td>1/5</td><td>30/1061</td><td>3/205</td><td>0/8</td><td>0/5</td><td>0/129</td><td>0/4</td></tr>
<tr><td colspan="11">(d) 縣市資料分佈</td></tr>
<tr><td colspan="6">最多</td><td colspan="5">最少</td></tr>
<tr><td colspan="6">台北市(46)/台北市(2,845)</td><td colspan="5">金門縣(1)
連江縣(1)/連江縣(13)
澎湖縣(1)</td></tr>
</table>

險研究資料庫」。文中任何闡釋或結論並不代表衛生福利部中央健康保險署、或財團法人國家衛生研究院之立場。

4 有胃食道逆流的頭頸癌病患數量為 0，經與醫師討論可能原因在於健保申報數量有限制，醫師通常選擇申報較嚴重之病徵，而胃食道逆流是相對較不嚴重之疾病，可能因此未被申報。

(e) 數值類別			
	平均值	最大值	最小值
年齡	60/53.94	98/96	2/0
投保金額	4256.20/1532.55	69800/182000	0/0

本研究利用二元邏輯斯迴歸檢定，來了解各個資料變項對於應變數是否有顯著影響。由於胃食道逆流的資料量為 0，因此將檢定資料中胃食道逆流變項去除，在 99%的信賴區間下，可以發現顯著性<0.01 的影響屬性（p 值）包含年齡（0.000）、腫瘤位置（0.000）、管灌膳食（0.000）、放射治療（0.000）、中風與否（0.000）、帕金森病（0.002）、投保金額（0.000）、所在地區（0.000），其中腫瘤位置在唾液腺及喉部較為顯著，而病患所在地區以台中市較為顯著。

二、實驗設計與評估準則

在實驗設計上，本研究從頭頸癌併發吸入性肺炎的 427 筆資料利用傾向分數從未併發吸入性肺炎的 17,298 筆資料配對出 427 筆，使正負個案比例為 1:1，形成共 854 筆實驗資料集合。本研究採用 SPSS 22.0 版本所提供之功能，選擇以「最大化執行效能」方式進行傾向分數配對，亦即最近相鄰配對法，其在選擇配對資料時會優先選取最相近資料，因此配對結果可能會受到資料讀取順序影響。為避免亂數配對影響，本研究先將資料以病患年齡進行順向排列後（小到大），再進行資料配對。本研究以 10 折交叉驗證法（10-fold cross-validation）進行分類器效能評估，亦即將資料集合隨機切割成 10 等份，每次實驗取其中一份為測試資料，其餘為訓練資料，反覆進行 10 次讓每等份資料皆擔任過測試資料，確保所有資料皆被預測過。

如前所述，本研究同時評估 Bagging 跟 Boosting 兩種整體學習法，了解它們所建構之頭頸癌併發吸入性肺炎高風險群預測分類器，在預測效能上的差異。在 Bagging 方法上，本研究觀察並比較三個較具代表性之基礎學習演算法，分別為 Naïve Bayes、Decision Tree、及 Radial Basis Function（RBF）Kernel Support Vector Machine（SVM）。在 Boosting 方法上，則採用較具代表性之 AdaBoost，以及改良之 LogitBoost 方法。本研究之實驗環境為 Weka 3.7.13、SPSS 22.0 安裝於配備 Intel Core 2 Duo CPU 2.83GHz，4GB RAM 以及 Windows 8.1 x64 作業系統之個人電腦上。

分類器效能的評估大多是依據混淆矩陣（Confusion Matrix）來計算相關指標。表 5 為此研究分類問題之混淆矩陣，其中 TP 為實際併發吸入性肺炎患者被預測為有併發吸入性肺炎；FN 為實際併發吸入性肺炎患者被預測為未併發吸入

性肺炎；FP 為實際未併發吸入性肺炎患者被預測為有併發吸入性肺炎；TN 為實際未併發吸入性肺炎患者被預測為沒有併發吸入性肺炎。

表 5：混淆矩陣

		預測結果	
		併發吸入性肺炎	未併發吸入性肺炎
實際結果	併發吸入性肺炎	TP	FN
	未併發吸入性肺炎	FP	TN

本研究採用敏感性（sensitivity）、專一性（specificity）、及準確度（accuracy）做為評估指標，並計算 ROC 曲線下方面積（AUC ROC）。敏感性是指能夠正確預測確實有併發吸入性肺炎病患的比例，其指標定義為 TP/(TP+FN)；專一性是指能夠正確預測確實未併發吸入性肺炎病患的比例，其指標定義為 TN/(TN+FP)；準確度是指能夠正確預測病患為有併發或未併發吸入性肺炎的比例，其指標定義為(TP+TN)/(TP+TN+FP+FN)；ROC 曲線則是由敏感性與（1-專一性）所形成的折衷曲線。

三、實驗評估結果

如前所述，本研究主要想了解 Bagging 與 Boosting 整體學習方法所建構分類器效能上的差異。本研究分別比較以 Naïve Bayes、Decision Tree、及 RBF-Kernel SVM 作為 Bagging 中基礎學習演算法，以及 AdaBoost 與 LogitBoost 等五種分類器的效能。各分類器採用之執行參數以 Weka 內定參數為主，其設定通常參考文獻建議之最佳參數值。

表 6 中實驗評估顯示，除 RBF-Kernel SVM 獲得最低的預測準確度 89.1% 外，其餘方法皆可達 95%以上預測準確度。兩種 Boosting 方法以及 Bagging-Decision Tree 獲得相似的評估結果，而 RBF-Kernel SVM 整體略遜於其他方法，本研究認為其原因可能是 RBF-Kernel SVM 基本上係用來處理線性不可分割資料的問題，對於線性可分割資料問題其預測效能會降低。儘管預測高風險群病患的準確度較差，RBF-Kernel SVM 獲得之 ROC 曲線下方面積仍達 0.9 以上，顯示其預測模式亦具有極佳的鑑別力（Lüdemannet al. 2006; Metz 1978; Obuchowski 2003）。總和來說，本研究實驗結果大致與過去研究結論一致，亦即 Boosting 方法在一般資料情況下預測效能優於 Bagging 方法；而 Bagging 方法效能差異，取決於採用的基礎學習演算法。

表 6：完整屬性資料之實驗評估結果

	Sensitivity	Specificity	Accuracy	AUC ROC
Bagging-Naïve Bayes	0.939	0.967	0.953	0.986
Bagging-Decision Tree	0.974	0.998	0.986	0.983
Bagging-RBF Kernel SVM	0.810	0.972	0.891	0.933
AdaBoost	0.972	0.998	0.985	0.989
LogitBoost	0.972	0.995	0.984	0.990

由於本研究利用傾向分數配對，並以向下抽樣（undersampling）方式，讓正負類別資料數量相同來做為訓練資料，目的在使正負兩類資料自變項分佈相近，讓建構之分類器能較精確預測資料量較小之目標類別，亦即頭頸癌併發吸入性肺炎高風險群。然而從另一角度來看，此建構方式可能會忽略大資料群集中未被配對之資料特性，亦即建構之分類器無法準確預測（排除）非目標類別資料，亦即頭頸癌未併發吸入性肺炎的病患。因此，本研究從未被抽出之非目標類別資料中隨機抽樣 1,000 筆做為測試資料，來評估目前建構之分類器的預測效能。由於只有非目標類別資料，因而本研究僅呈現分類器在 Specificity 方面的效能。實驗結果令人感到特別，如表 7 所示 RBF-Kernel SVM 在此預測上獲得約 0.717 的 Specificity，意謂著其能將 71.7%的未併發吸入性肺炎病患準確地排除為非高風險病患，其他四個方法則沒有意外地，對於未被納入訓練資料內之資料，獲得非常差的預測效能。由於 RBF-Kernel SVM 其特性在於將輸入樣本映射到較高維度特徵空間，以解決部分線性不可分的問題，而此結果暗示著原始資料可能存在線性不可分的情況。若以此角度來看，RBF-Kernel SVM 的整體效能其實應勝過於其他方法。

表 7：未配對之未併發吸入性肺炎病患預測效能

	Specificity
Bagging-Naïve Bayes	0.070
Bagging-Decision Tree	0.026
Bagging-RBF Kernel SVM	0.717
AdaBoost	0.030
LogitBoost	0.041

基於上述，本研究針對 Bagging 策略下之 RBF-Kernel SVM 分類器，進一步分析不同屬性組合的資料中對其預測效能的影響。過去文獻分別提及各種不同因

素會影響頭頸癌病患併發吸入性肺炎，其中 Chu 等（2013）提出的影響因素，包含性別、年紀、醫院等級，治療方式（頸淋巴結清除術、胃造口術、同步放化療）、共病（老人痴呆、中風、胃食道逆流、帕金森氏症）、腫瘤位置，大致為本研究採用之屬性的子集合，本研究將採用此六個屬性的資料集合來建構分類器。此外本研究回顧之文獻中皆共同提到性別、年紀、腫瘤位置三個影響因素，因此本研究會同時比較以此 3 個屬性的資料集合所建構之分類器效能。表 8 為利用 Bagging 策略下 RBF-Kernel SVM 在三種不同數量屬性下建構分類器的預測效能，從表中可以發現三種屬性組合對於配對資料之預測效能近乎相同，以完整屬性所建構之分類器僅在 AUC ROC 略高於其他方法，然而在預測未被配對到的非目標類別資料時，完整屬性建構之分類器，顯著優於其他屬性組合所得之預測效能。

表 8：不同屬性對 Bagging 策略下 RBF-Kernel SVM 分類器效能的影響

	Sensitivity	Specificity	Accuracy	AUC ROC	All-Negatives
完整屬性	0.810	0.972	0.891	0.933	0.717
Chu 提出之 6 個屬性	0.794	0.986	0.890	0.914	0.650
文獻提出之 3 個共同屬性	0.799	0.988	0.893	0.919	0.370

四、隨機抽樣產生資料集合之效能分析評估

由於本研究利用傾向分數配對方法，不同於以往常見的隨機抽樣方法，來產生訓練資料，因此本研究進一步評估以隨機抽樣方式產生之資料集合所建構之分類器的預測效能。同樣地，為形成目標與非目標類別資料數量相等之訓練集合，本研究以隨機抽樣方式從非目標類別資料，亦即頭頸癌未併發吸入性肺炎的一萬多筆資料中，抽取與目標類別資料相對應的數量，形成訓練資料集合，並以 10 折法進行分類器效能評估。本研究重複以上步驟 30 次，並以 30 次所得之平均效能為最終評估結果。如表 9 所示，以隨機抽樣方式產生之訓練資料，基本上降低所有學習演算法的預測效能，從原本九成以上的準確度，降低至六成多；而所得之 ROC 曲線下方面積亦降低到 0.6～0.7 之間。雖然各分類器的效能排名稍有不同，但從實驗結果可以明顯發現，不論是何種整體學習演算法，對於利用隨機抽樣方式所產生之資料集合所建構之分類器，預測效能皆大幅降低。

表 9：隨機抽樣方法之實驗評估結果

	Sensitivity	Specificity	Accuracy	AUC ROC
Bagging-Naïve Bayes	0.801	0.506	0.653	0.698
Bagging-Decision Tree	0.671	0.610	0.641	0.684
Bagging-RBF Kernel SVM	0.851	0.356	0.603	0.640
AdaBoost	0.680	0.577	0.628	0.663
LogitBoost	0.706	0.579	0.643	0.688

伍、研究結論

隨著疾病型態由急性疾病與傳染病，轉變為癌症和慢性疾病，加上健保制度與醫療體制的限制，使得預防重於治療的概念逐漸受到重視。頭頸癌本身是可以治癒的疾病，然而病患治癒後卻可能因後遺症引發吸入性肺炎，進而導致死亡。吸入性肺炎在臨床表現中可能沒有症狀，或者容易造成乾咳、呼吸急促、支氣管痙攣、血性或泡沫痰、呼吸窘迫等。儘管吸入性肺炎是潛在威脅生命的併發症，若能提醒高風險之患者在居家照護時注意預防或及早發現相關症狀，卻是可以預防的。過去研究試圖提出可能造成頭頸癌併發吸入性肺炎的影響因素，其說明特定族群的頭頸癌患者，可能具有較高併發風險。然而過去各研究之間所觀察之影響因素多有不同，且所得之結果略有差異，且實務上仍未發展出相關指標，供醫師評估患者並提供適當衛教，協助病患或家屬於居家照護時預防或察覺吸入性肺炎的發生，以降低感染的發生以及後續衍生的醫療成本。

台灣在 1995 年推行國民健康保險制度後，至今已累積相當年份與數量的醫療保險申報資料。本研究基於該資料庫，並利用資料探勘技術，來分析建構頭頸癌併發吸入性肺炎高風險群之預測模式，期望能協助醫師評估頭頸癌病患是否有併發吸入性肺炎之風險。本研究先藉由回顧過去相關研究，彙整相關可能造成頭頸癌病患併發吸入性肺炎的影響因素，再比對健保研究資料庫中現有資料欄位，最終取得 13 個相關屬性資料，用以建構頭頸癌併發吸入性肺炎高風險群預測模式。本研究採用傾向分數配對法來形成訓練資料集合，並同時利用 Bagging 與 Boosting 兩種不同的整體學習方法來建構預測模式，其中在 Bagging 方法上評估 Naïve Bayes、Decision Tree、以及 RBF-Kernel SVM 三種基礎學習演算法；而 Boosting 方法上則評估 AdaBoost 以及 LogitBoost 演算法。實驗評估結果顯示，此五種方法在 ROC 曲線下方面積最低仍達 0.933，顯示無論利用何種方式所建構之分類器，對於配對資料皆具備有極佳的判別力。在預測準確度方面，除 Bagging 整體學習策略中的 RBF-Kernel SVM 準確度較差，僅達 89%外，其餘方法都能達成 95%以上的預測準確度，且兩種 Boosting 演算法與以 Decision Tree 為基礎的

Bagging 方法獲得最佳的預測效能。儘管如此，RBF-Kernel SVM 對於未被納入訓練資料之非目標群集卻有絕佳的預測能力，因而總和來說 RBF-Kernel SVM 預測效能事實上優於其他方法。

本研究的貢獻可從四個方面來說明。首先，本研究有助於提高醫療院所的服務品質。本研究藉由建構頭頸癌併發吸入性肺炎高風險病患預測模式，來協助醫師評估病患是否屬於較高之併發風險，給予病患適當的衛生教育，協助病患於居家照護時預防或及時察覺吸入性肺炎的發生，避免後續衍生的醫療問題與成本，並相對提高病患對於醫療服務的滿意度。其次，本研究提出一個建置醫療輔助系統的可行方法。本研究建構之預測模式係基於現有的健康保險資料以及現存的資料探勘技術，基本上可以免除建置系統時經常面臨的資料蒐集問題以及技術需求。再者從醫療資料探勘的角度，本研究提出的探勘方式，從實驗的結果來看，確實能建立有效的頭頸癌併發吸入性肺炎高風險病患預測模式。本研究採用傾向分數配對方式來建立訓練資料，其效能優於過去研究常用之隨機抽樣。此外本研究發現，雖然 RBF-Kernel SVM 在訓練資料的預測效能上較差，然而其對於未被納入訓練集合之資料預測效能卻遠高於其他方法。此結果可能暗示著配對所產生之資料集合應為線性可分，但整體原始資料類別卻可能是線性不可分的狀態，這對未來相關研究提供可能的線索。最後，本研究對頭頸癌併發吸入性肺炎的研究亦應有貢獻。本研究回顧過去文獻，並從健保資料庫中彙整出頭頸癌病患併發吸入性肺炎之影響因素資料，而分析結果顯示有 8 個因素對於併發吸入性肺炎有顯著影響。此外本研究發現不同屬性組合資料對於訓練資料的預測效能，大致上沒有影響，然而採用文獻提及之三個共同屬性，亦即性別、年紀、腫瘤位置建構之分類器，對於非目標資料的預測能力會明顯下降。

儘管實驗評估結果顯示本研究所建構之預測模式能達成相當不錯之預測效能，但仍面臨一些研究上的限制，這些限制則建議著未來研究方向。首先，本研究的限制之一來自於研究資料，亦即健保資料庫本身的限制。健保資料庫是屬於申請保險給付資料，僅記錄病患就診最終的結果，亦即診斷的疾病以及採用的診療措施，因此無法取得病況資料，或是與保險無關之被保險人個人資料，因此某些屬性資料僅能以推測方式取得，而非直接資料。因此，本研究希望後續能整合更多面向的資料如：鄉鎮市的國民收入、醫療體系內部的檔案等資料，以期能建立更符合實際情況的預測模式，並促使醫療院所能夠提供更好的醫療服務。其次，本研究雖基於健保資料建構預測模式，且以實驗評估方式驗證其效能；儘管如此，此預測模式仍處於理論階段，缺乏實際驗證的結果，且目前演算法相關參數設定皆採用 Weka 內定，仍未進行實驗探討。因此本研究期望未來能實作資訊系統，並實際評估其在真實環境下的預測效能。最後，本研究提出之探勘方式，亦即利用傾向分數配對方式來建立訓練資料集合，雖能建立有效的預測模式，但

僅限於目前的研究問題以及採用的資料集合。因此未來研究可蒐羅更多相關資料，例如國外的資料集合，或是以其他醫療資料集合，來進一步驗證此探勘方式的有效性。

參考文獻

王宏銘、廖俊達、范綱行、吳樹鏗、詹勝傑、閣紫宸（2009），『頭頸部鱗狀細胞癌治療的新進展』，*腫瘤護理雜誌*，第九卷，第 S 期，頁 51-67。

陳錦華（2014），『傾向分數（propensity score）在估計風險比之使用方法』，*臺北醫學大學生物統計研究中心 eNews*，第二卷。

曾淑芬（1999），『從醫院管理角度論全民健保資料庫』，*中華公共衛生雜誌*，第十八卷，第五期，頁 363-372。

Adnet, F. and Baud, F. (1996), 'Relation between Glasgow Coma Scale and aspiration pneumonia', *Lancet*, Vol. 348, No. 9020, pp. 123-124.

Bauer, E. and Kohavi, R. (1999), 'An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants', *Machine Learning*, Vol. 36, No. 1, pp. 105-139.

Baum, G.L., Crapo, J.D., Celli, B.R. and Karlinsky, J.B. (1998), *Textbook of Pulmonary Diseases*, Lippincott Williams & Wilkins, Philadelphia.

Beasley, R.P., Lin, C.C., Hwang, L.Y. and Chien, C.S. (1981), 'Hepatocellular carcinoma and hepatitis B virus: a prospective study of 22 707 men in Taiwan', *Lancet*, Vol. 318, No. 8256, pp. 1129-1133.

Becker, S.O. and Ichino, A. (2002), 'Estimation of average treatment effects based on propensity scores', *The Stata Journal*, Vol. 2, No. 4, pp. 358-377.

Berry, M.J. and Linoff, G.S. (1997), *Data mining Techniques: For Marketing, Sales, and Customer Support*, Wiley Publishing, Inc., Indianapolis, Indiana.

Breiman, L. (1996), 'Bagging Predictor', *Machine Learning*, Vol. 24, No. 2, pp. 123-140.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002), 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, Vol. 16, No. 1, pp. 321-357.

Chu, C.N., Muo, C.H., Chen, S.W., Lyu, S.Y. and Morisky, D.E. (2013), 'Incidence of pneumonia and risk factors among patients with head and neck cancer undergoing radiotherapy', *BMC Cancer*, Vol. 13, No. 370.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine Learning*, Vol. 20,

No. 3, pp. 273-297.

Daniels, S.K., Brailey, K., Priestly, D.H., Herrington, L.R., Weisberg, L.A. and Foundas, A.L. (1998), 'Aspiration in patients with acute stroke', *Archives of Physical Medicine and Rehabilitation*, Vol. 79, No. 1, pp. 14-19.

Delgado, M., Sánchez, D., Martín-Bautista, M.J. and Vila, M.A. (2001), 'Mining association rules with improved semantics in medical databases', *Artificial Intelligence in Medicine*, Vol. 21, No. 1, pp. 241-245.

Dietterich, T.G. (2000), 'Ensemble methods in machine learning', *Proceedings of the First International Workshop on Multiple Classifier Systems*, Cagliari, Italy, June 21-23, pp. 1-15.

Dubray-Vautrin, A., Ballivet de Régloix, S., Girod, A., Jouffroy, T. and Rodriguez, J. (2015), 'Epidemiology, diagnosis and treatment of head and neck cancers', *Soins*, Vol. 60, No. 798, pp. 32-35.

Eisbruch, A., Lyden, T., Bradford, C.R., Dawson, L.A., Haxer, M.J., Miller, A.E. and Wolf, G.T. (2002), 'Objective assessment of swallowing dysfunction and aspiration after radiation concurrent with chemotherapy for head-and-neck cancer', *International Journal of Radiation Oncology, Biology, Physics*, Vol. 53, No. 1, pp. 23-28.

Frawley, W.J., Piatetsky-Shapiro, G. and Matheus, C.J. (1991), 'Knowledge discovery in databases: An overview', *AI Magazine*, Vol. 13, No. 3, pp. 57-70.

Freund, Y. and Schapire, R.E. (1996), 'Experiments with a new boosting algorithm', *Proceedings of the Thirteenth International Conference on Machine Learning (ICML '96)*, Bari, Italy, July 3-6, pp. 148-156.

Freund, Y. and Schapire, R.E. (1997), 'A decision-theoretic generalization of on-Line learning and an application to boosting', *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119-139.

Freund, Y. and Schapire, R.E. (1999), 'A short introduction to boosting', *Journal of Japanese Society for Artificial Intelligence*, Vol. 14, No. 5, pp. 771-780.

He, H. and Garcia, E.A. (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp. 1263-1284.

Henley, W.E. and Hand, D.J. (1996), 'A k-nearest-neighbour classifier for assessing consumer credit risk', *The Statistician*, Vol. 45, No. 1, pp. 77-95.

Irwin, R.S., Cerra, F.B. and Rippe, J.M. (1999), *Irwin and Rippe's Intensive Care Medicine*, Lippincott Williams & Wilkins, Philadelphia.

Kearns, M. and Valiant, L. (1989), 'Crytographic limitations on learning boolean

formulae and finite automata', *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, Seattle, WA, USA, May 14-17, pp. 433-444.

Lüdemann, L., Grieger, W., Wurm, R., Wust, P. and Zimmer, C. (2006), 'Glioma assessment using quantitative blood volume maps generated by T1-weighted dynamic contrast-enhanced magnetic resonance imaging: A receiver operating characteristic study', *Acta Radiol*, Vol. 47, No. 3, pp. 303-310.

Langerman, A., MacCracken, E., Kasza, K., Haraf, D.J., Vokes, E.E. and Stenson, K.M. (2007), 'Aspiration in chemoradiated patients with head and neck cancer', *Archives of Otolaryngology–Head & Neck Surgery*, Vol. 133, No. 12, pp. 1289-1295.

Lee, Y.H., Hu, P., Cheng, T.H., Huang, T.C. and Chuang, W.Y. (2013), 'A preclustering-based ensemble learning technique for acute appendicitis diagnoses', *Artificial Intelligence in Medicine*, Vol. 58, No. 2, pp. 115-124.

Lewis, D. and Catlett, J. (1994), 'Heterogeneous uncertainty sampling for supervised learning', *Proceedings of the 11th International Conference on Machine Learning*, New Brunswick, NJ, pp. 148-156.

Lin, Z., Hao, Z., Yang, X. and Liu, X. (2009), 'Several SVM ensemble methods integrated with under-sampling for imbalanced data learning', *Proceedings of the Fifth International Conference on Advanced Data Mining and Applications (ADMA'09)*, Beijing, China, August 17-19, pp. 536-544.

Liu, X.Y., Wu, J. and Zhou, Z.H. (2009), 'Exploratory undersampling for class-imbalance learning', *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 39, No. 2, pp. 539-550.

Marik, P.E. (2001), 'Aspiration pneumonitis and aspiration pneumonia', *New England Journal of Medicine*, Vol. 344, No. 9, pp. 665-671.

Metz, C.E. (1978), 'Basic principles of ROC analysis', *Seminars in Nuclear Medicine*, Vol. 8, No. 4, pp. 283-298.

Mingers, J. (1989), 'An empirical comparison of pruning methods for decision tree induction', *Machine Learning*, Vol. 4, No. 2, pp. 227-243.

Mitchell, T.M. (1997), *Machine learning*, McGraw Hill.

Mittal, B.B., Pauloski, B.R., Haraf, D.J., Pelzer, H.J., Argiris, A., Vokes, E.E., Rademaker, A. and Logemann, J.A. (2003), 'Swallowing dysfunction--preventative and rehabilitation strategies in patients with head-and-neck cancers treated with surgery, radiotherapy, and chemotherapy: a critical review', *International Journal of Radiation Oncology, Biology, Physics*, Vol. 57, No. 5, pp. 1219-1230.

Mortensen, H.R., Jensen, K. and Grau, C. (2013), 'Aspiration pneumonia in patients treated with radiotherapy for head and neck cancer', *Acta Oncologica*, Vol. 52, No. 2, pp. 270-276.

Nilsson, N.J. (1965), *Learning Machines*, McGraw-Hill, New York.

Obuchowski, N.A. (2003), 'Receiver operating characteristic curves and their use in radiology', *Radiology*, Vol. 229, No. 1, pp. 3-8.

Parsons, L.S. (2001), 'Reducing bias in a propensity score matched-pair sample using greedy matching techniques', *Proceedings of the Twenty-Sixth Annual SAS® Users Group International Conference*, Long Beach, California, USA, April 22-25, pp. 214-226.

Rosen, A., Rhee, T.H. and Kaufman, R. (2001), 'Prediction of aspiration in patients with newly diagnosed untreated advanced head and neck cancer', *Archives of Otolaryngology-Head & Neck Surgery*, Vol. 127, No. 8, pp. 975-979.

Roy, T.M., Ossorio, M.A., Cipolla, L.M., Fields, C.L., Snider, H.L. and Anderson, W.H. (1989), 'Pulmonary complications after tricyclic antidepressant overdose', *CHEST Journal*, Vol. 96, No. 4, pp. 852-856.

Siegel, R., Naishadham, D. and Jemal, A. (2012), 'Cancer statistics, 2012', *CA: A Cancer Journal for Clinicians*, Vol. 62, No. 1, pp. 10-29.

Valiant, L.G. (1984), 'A theory of learnable', *Communications of the ACM*, Vol. 27, No. 11, pp. 1134-1142.

Ward, E., Jemal, A., Cokkinides, V., Singh, G.K., Cardinez, C., Ghafoor, A. and Thun, M. (2004), 'Cancer disparities by race/ethnicity and socioeconomic status', *CA: A Cancer Journal for Clinicians*, Vol. 52, No. 4, pp. 78-93.

Xu, B., Boero, I.J., Hwang, L., Le, Q.T., Moiseenko, V., Sanghvi, P.R., Cohen, E.E., Mell, L.K. and Murphy, J.D. (2015), 'Aspiration pneumonia after concurrent chemoradiotherapy for head and neck cancer', *Cancer*, Vol. 121, No. 8, pp. 1303-1311.

Zorman, M., Eich, H.P., Kokol, P. and Ohmann, C. (2001), 'Comparison of three databases with a decision tree approach in the medical field of acute appendicitis', *Studies in Health Technology and Informatics*, Vol. 84, No. 2, pp. 1414-1418.