機器學習在生物醫學的應用：腦瘤患者的分類診斷

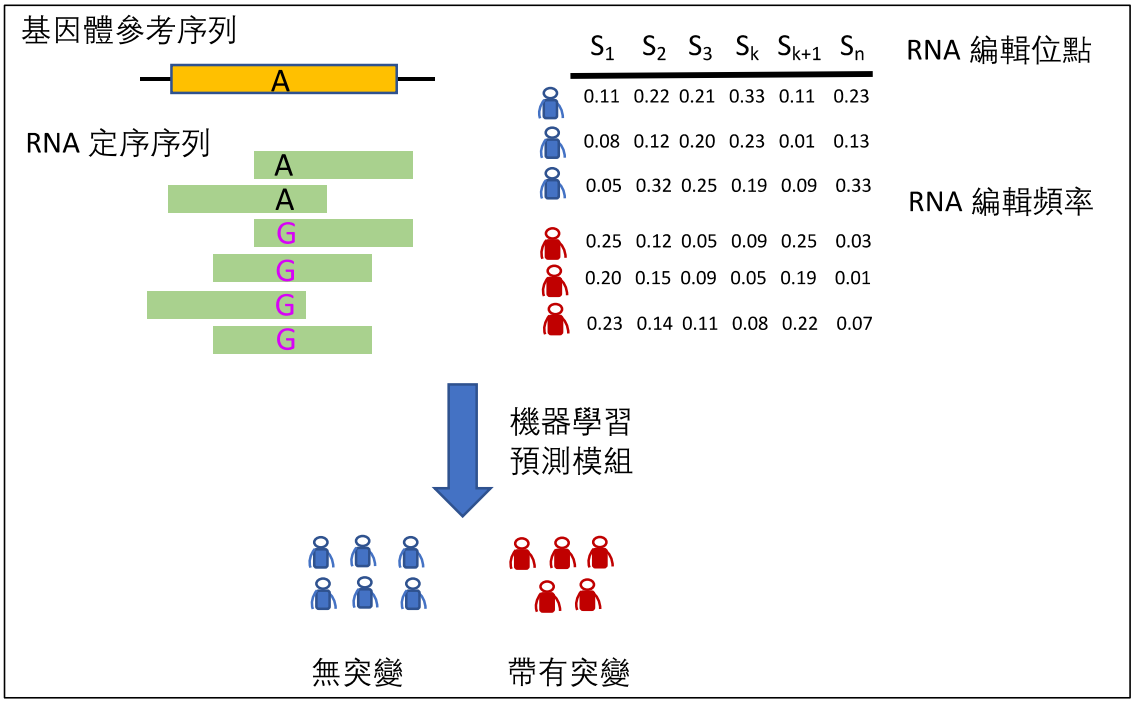
陳俊璋
臺北醫學大學醫學資訊研究所

近年來生命科學技術的快速進展，得以用較低成本與快速的方式產生醫療大數據，再透過機器學習與人工智慧進行分析，可以作為臨床治療的指引，進而實現精準醫療。例如，藉由分析病患的基因序列資料，能快速診斷疾病以及疾病分期，並且協助找出對病患副作用最低的藥物和治療選項。我們接下來介紹用機器學習來分析腦瘤患者的腫瘤 RNA，對患者進行分類。

膠質瘤(glioma)是成人中最常見的原發性腦腫瘤，全球每年發病率接近十萬分之 6，男性發生率略高於女性。世界衛生組織 (World Health Organization) 將膠質瘤依照其侵襲性分為不同級別(grade)：2、3 或 4 級。第 2 級和 3 級患者的預後(prognosis)和存活率通常優於第 4 級的患者。第 4 級的腦腫瘤是最惡性的腦腫瘤，又稱作神經膠質母細胞瘤 (glioblastoma multiforme)，至今仍然無法治癒，即使在接受標準治療 (手術切除、化學治療和放射治療）後，大多數患者在診斷後兩年內死亡。因此，開發對腦腫瘤患者進行分類和存活率預測的生物標誌物(biomarker)，對於疾病的管理、治療的選擇、藥物的研發至關重要。

世界衛生組織在 2016 年採用兩個生物標誌物來對膠質瘤進行分類：(1)異檸檬酸脫氫酶 (isocitrate dehydrogenase，IDH)是否存在突變; (2)染色體 1p/19q 是否有缺失。然而，檢測這兩個生物標誌物的狀態相當費時且耗成本，而且仍然有人工診斷不一致的問題。例如，免疫組織化學染色法 (Immunohistochemistry，IHC) 是常用的檢測 IDH 突變的方法，由於利用抗體來識別突變，IHC 無法偵測不太常見的 IDH 突變。同樣，螢光原位雜合技術 (Fluorescence in situ hybridization，FISH) 在醫院中被廣泛用來檢測 1p/19q 的狀態，但需要有經驗的病理學家確認，但不同病理學家對於同一個檢體仍會有不同的見解。因此，建立標準化、準確、可重複且客觀的判別 IDH 突變和 1p/19q 缺失的方法有其必要性。

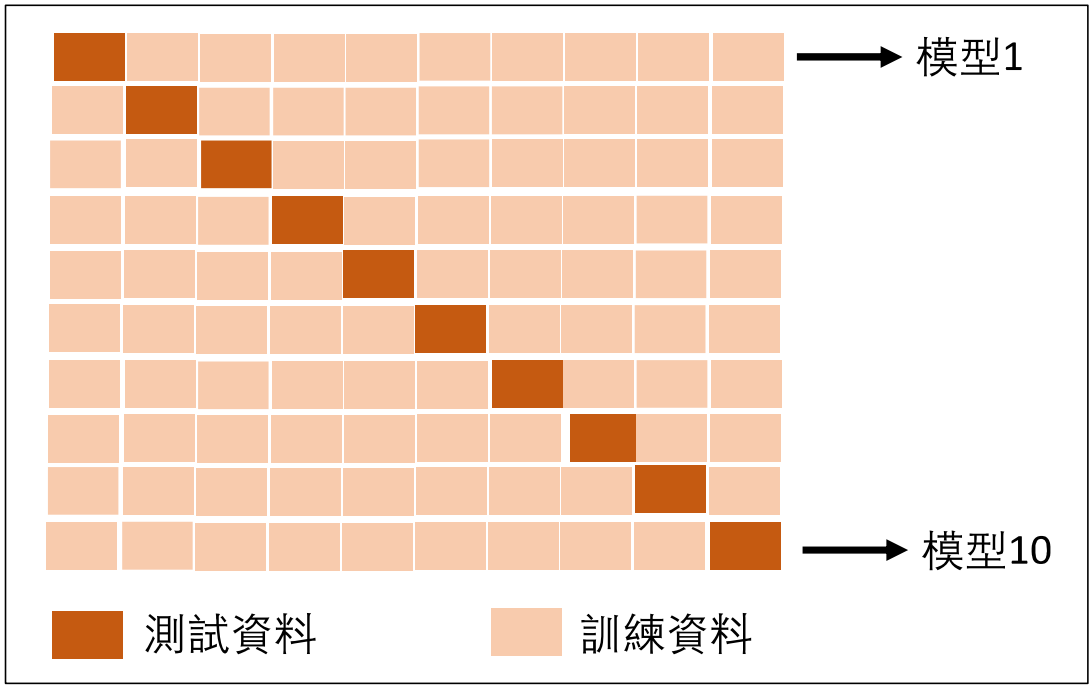
我們使用支持向量機 (Support Vector Machines，SVM)、隨機森林 (Random Forest) 和 AdaBoost (Adaptive Boosting) 三種常見的監督式機器學習算法來分析腦腫瘤的 RNA 編輯事件，以預測 IDH 突變和 1p/19q 缺失的狀態。將 RNA 上的腺苷(adenosine，A)轉化為肌苷(inosine，I) 的 RNA 編輯是一種廣泛存在的後轉錄現象，RNA 編輯引起的核苷酸變化可能會影響蛋白質序列、RNA 的二級結構和受小分子核糖核酸 (microRNA)調控的 mRNA 表達量。透過 RNA 定序的資料，我們找出在腦腫瘤的 A-to-I RNA 編輯事件(定序儀器會把 I 定為 G)，並計算其發生的頻率 (如圖一)，因此產生一個很大的資料表格，透過此表格，再建立機器學習模組來預測是否帶有突變。



【圖一】資料產生和分析流程

因為患者的數目有限，我們採用 10-fold 交叉驗證 (cross-validation)進行模型建立和評估，交叉驗證在樣本數量有限時廣泛使用，可提供準確的預測性能估計。首先將資料隨機平均分成 10 個大小相等的集合 (如圖二)，在每次驗證中，使用 9 個集合做為「訓練資料(Training data)」來訓練模型，沒有用到的那一個集合做為「測試資料(Testing data)」來評估性能，如此重複進行 10 次，直到每一個集合都被當做「測試資料(Testing data)」為止。我們同時還在分類器的

每個 fold 內進行特徵選擇 (feature selection)，以避免過度擬合，並且可能提高預測性能。



【圖二】交叉驗證

我們的結果顯示 RNA 編輯具有至關重要的臨床實用性，我們的模型可以準確的預測 IDH 突變和 1p/19q 缺失。與 IHC 和 FISH 方法相比，我們的模型提供了更客觀的診斷並可以避免人為的標記錯誤。在檢查原始病理報告後，我們發現在預測 1p/19q 缺失時，被判定為模型分類錯誤的六個樣本中，有四個樣本的原始標籤的分類和預測的結果相同，是因為病例填寫過程中的失誤導致被判定錯誤分類，更加的凸顯了我們模型的準確性和臨床實用性。本研究確定了 RNA 編輯為膠質瘤的新型預後生物標誌物，我們的預測模型提供標準化、準確、可重複和客觀的膠質瘤分類。我們的模型不僅對臨床決策有用，而且能夠辨別出在膠質瘤的管理和治療中，有可能當作生物標誌物和治療靶點的編輯事件。

參考資料與文獻：

1. Chen et al. RNA editing-based classification of diffuse gliomas: predicting isocitrate dehydrogenase mutation and chromosome 1p/19q codeletion. BMC Bioinformatics 2019, 20(Suppl 19):659

2. Louis et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. 2016;131(6):803–20.

3. Zou et al. Comparison of immunohistochemistry and DNA sequencing for the detection of IDH1 mutations in gliomas. Neuro-Oncology. 2015;17(3):477–8.

4. Chaturbedi et al. Detection of 1p19q deletion by real-time comparative quantitative PCR. Biomark Insights. 2012;7:9–17.