

## 1. What is the question?

健保資料為健保局從醫療機構定期取得並記錄之資料，其中包含病患個人資訊、症狀以及治療藥物；透過這些資料庫，作者希望能透過文字探勘與比對的技術，配對並建構一個常態的模型，輔助醫療決策的判斷。透過這個模型，醫療單位能夠根據病患的症狀，推斷其需要何種藥物；或根據藥物的服用，診斷其可能的疾病。

## 2. Why should we care about it?

醫療診斷有其實務上的限制，不一定能達到完全的準確，透過以往的經驗、文字探勘的技術建構出來的資料庫以及預測模型，除了能夠協助醫療單位的決策外，同時能偵測與減少誤診的機會、增進整體醫療環境的品質、充實醫學知識的完整度與準確性，對整體醫療產業的裨益相當大。

## 3. What is yours (or author' s) answer?

透過許多不同的文字探勘方法，作者發現 Multiple-Channel Latent Dirichlet Allocation (MCLDA) 這個方法能夠最準確的在藥物與症狀間對應與預測，同時最能夠發現與解釋藥物、症狀、病患資訊間潛在的結構與連結。

## 4. How did you (or author) get there?

文字探勘是一種根據文字間的連結去攫取相關資訊的方法，其中包含了許多層面，像是 Information Extraction 資訊萃取(從一大段文章裡萃取出有用的資訊或重點，如人名、機構、時間等)、Sentiment Analysis 情緒判讀(從一則評論或敘述裡，透過斷詞與文字庫解讀筆者的正負面情緒、批評或讚賞)、Topic Models 主題模型(從一段落分類出不同主題並分析比重)。作者主要透過第三種層面的 Topic Models，使用 MCLDA 的方法根據病患就醫的文字記錄去分類症狀、藥物與病患資訊，反映整合出病患的健康狀態(無法直接觀察)，並定義一個能連結各種可觀察變數與健康狀態的模型，最後發現這個模型能夠比其他方法更準確的吻合實際情形。