醫學文獻評讀概念、方法與等級介紹

Introduction of critical appraisal of medical literature

陳杰峰　王慈蜂*

台北醫學大學 市立萬芳醫院 整形外科

台北醫學大學 市立萬芳醫院 家庭醫學科*

前　言

1992 年實證醫學第一次在學界被廣泛介紹（"Evidence-based medicine. A new approach to teaching the practice of medicine," 1992），實證醫學操作型定義主要有五大步驟，分別爲整理可以回答的問題，搜尋文獻證據，嚴格評讀文獻，應用於病人身上，對過程進行稽核。其中嚴格評讀文獻（critical appraisal）是在認識實證醫學時，被覺得較耗費精力且不容易理解的步驟。2006 年 Paul Glasziou 受邀至台灣演講中提到，實際在臨床應用的實證醫學，評讀迅速找到的正確文獻乃是整個實證醫學運作的關鍵步驟。

評讀文獻之相關因素及其等級

被選來嚴格評讀的主題，通常有下列幾個特色（Evidence-Based Medicine Working Group, 1992）：第一、重要的臨床問題，第二、目前尚無結論且不確定的問題，第三、專家意見無法讓人完全相信。實際評讀文獻的過程，即是審視研究中有可能的偏差（bias）。隨機對照試驗的統合分析，被公認是偏差較少的研究設計，較值得相信。實證醫學中所謂的嚴格評讀文獻，主要是依據其方法學而分等級，早在九〇年代，美國健康照護政策與研究機構（the Agency For Healthcare Research And Quality, AHCPR）及英國牛津大學實證醫學中心（Oxford University , Centre for EBM）都提出證據等級的分類方法，其中治療型文獻的 level 1a 皆爲隨機對照試驗的統合分析。相對的，預後型的 level 1a 文獻是世代研究的系統性回顧。而診斷型的 level 1a 文獻，則是橫跨不同臨床試驗中心所做的世代研究之文獻。

文獻評讀三部曲

在文獻評讀方面最主要有三個主要步驟，即爲 VIP：V（Validity/Reliability）效度/信度、I（Importance/Impact）重要性、P（Practice/Applicability）臨床適用性。審視效度就是問”我們能相信這篇文獻嗎?”。審視重要性就是問”我們相信它，但這個結論重要嗎?”。審視臨床適用性就是

問”如果我們相信它，這個結論可以應用在我們所有的病患嗎?”。這三個問題是在文獻評讀的時候，三個最核心的部份。

Oxford University Question Logbook（http://www.wanfang.gov.tw/ebm/14_tools/index.htm）提到可以用 RAMbo 的方式審核效度。RAMbo 是代表 Randomized、Accounted、Measurement，其中 Randomized 是病人有同等的機會進入臨床試驗，而且檢視病人的治療分派是否爲隨機，隨機的方法是否適當，實驗開始前，實驗組及對照組是否相似、可比較？1920 年代，RA Fisher 是歷史上第一個提出隨機分配研究方法的學者，以作爲實驗設計的基本原則（Fisher, 1935）。而開始應用在醫療照護領域上是 Sir Austin Bradford Hill，於 1948 年所發表的研究（Medical Research Council, 1948）。在 1970 年代，Archie Cochrane 首先提出隨機對照試驗，是評斷醫學研究文獻最重要的研究方法。（http://www.cardiff.ac.uk/insrv/libraries/scolar/archives/cochrane/warexperience.html）Archie Cochrane 在二次世界大戰時，曾經被德軍俘虜到克里特島，在集中營裡大批的囚犯罹患腳氣病，Cochrane 將病人分爲兩組，其中一組在食物中添加酵母菌，另外一組沒有，結果在食物中添加酵母菌的戰俘，腳氣病獲得控制，這個試驗是 Cochrane 最早、最差勁但也最成功的隨機對照試驗（梁繼權, 2003）。第二個概念是『列入』（accounted for all patients），其定義爲病人追蹤是否夠久夠完整，所有病人是否依其原先分派組別做分析。這裡面有一個重要觀念即『治療意圖分析』（intention-to-treat analysis, ITT），隨機試驗的一種分析方法；該法中所有被分配在治療組/對照組的病人，無論是否完成該項治療/安慰劑，都應該被放進治療組/對照組(原分派組別)中作分析。『衡量』（Measurement），則是考量一篇文獻它的測量標準是否依照盲法及客觀這二個要素來操作，也就是病人及醫師是否對治療不知情（blind）？實驗組及對照組是否被同等對待？

第二部曲 importance（我們相信它，但結果是否重要？）。在這個審核的標準中，我們審核這個研究的結果是什麼？研究的結果如何被估計？以及經過多久的時間？結果被估計的方式？我們曾經聽過很多文獻結果的評估方式，如益一需治數（number needed to treat, NNT）、相對風險性降低度（relative risk reduction, RRR）、絕對危險降低度（absolute risk reduction, ARR），其中我們在臨床上比較建議使用益一需治數（NNT），定義是 NNT=1/ARR，就是絕對危險降低度的倒數，即在一段觀察時間內，爲預防一個不良結果或爲使一位病人達到實驗所求之有益結果，所需治療的病人數量（ARR 越大，NNT 越小）。爲什麼 NNT 和病人溝通的時候較容易讓病人理解，是因爲整數比的比例，較容易被人理解。

第三部曲爲臨床適用性的評估：”如果我們相信這個研究，它的結果是否可以應

用在我們的病患身上？”其中要考量的是病患的差異、可運用的資源以及病患的偏好。在這個部份我們可以考量病患的生物因素（biologic issues）即是”同樣的治療應用在不同的病患族群是否有不同的反應？” “我們的病人與研究中的病人是否非常不同，以致無法應用在研究結果？”其他的社會經濟因素（social and economic issues）的考量，是評估這個結果的可行性，也就是”這個治療適用於我們的診療環境嗎？病患的配合度如何？醫療提供者的配合度及能力如何？”另外流行病學因素（epidemiological issues）也是需要列入考慮的，也就是考量”我們的病人是否有其他共病狀況，可能改變治療的結果？影響有多大？病人可能從治療中得到什麼好處或壞處？經由治療而減少的不良後果是否比不治療有明顯的差別？”

如何進行文獻評讀

文獻搜尋及評讀主要有幾個方法，第一、直接應用已經評讀過的文獻資料庫（例如: ACP journal club, Cochrane Library）；第二、自行評讀文獻，也就是用簡單的評讀方式，例如上述提到的 VIP 或是 PICO/RAMbo 等概念；第三、必要時使用評讀工具協助，如：CATmaker 或是 CASP（Critical Appraisal Skills Program），使用評讀工具的好處是協助初學者按部就班完成評讀、協助評讀者較客觀地完成評讀，茲分述如下：

CATmaker（http://www.cebm.net）係由英國倫敦的 National Health Service R&D EBM 中心發展出來的軟體。CATmaker 可以幫助執行重要的臨床統計；同時儲存我們的問題、搜尋策略及評讀結果－草稿儲存爲【kittens】及完稿儲存爲【CATs】。CATmaker 也可以製作成檔案夾，存放於個人資料庫中。進到 CATmaker 首頁先確定所要評讀的文獻爲系統性文獻回顧或是治療、診斷、預後以及病因，再來進行評讀。這是非常適合初學者的自學應用工具，其中有一個 critical appraisal guides，若有問題可以隨時進入這個導讀工具，以利學習。

CASP 則是英國 Learning & Development at the Public Health Resource Unit 的一個計劃，從 1993 年起，該計劃與當地全國及國際性組織共同發展以實證方式從事醫療與社會照護。可從英國公共衛生資源單位（Public Health Resource Unit, PHRU）中免費下載（http://www.phru.nhs.uk）。CASP 也是提出評讀時應考量的三個主要議題－信度、重要性、臨床適用性（VIP），CASP 協助個人發展評讀研究證據的技能，進而將知識化爲行動。CASP 的工作小組與資源分成三個方向，反映出 CASP 的三箭頭標記（three-arrow logo），分別爲搜尋研究證據、評讀研究證據及依研究證據行醫。英國國家醫療衛生服務機構（National Health Services, NHS）包含很多嚴格評讀工具，我們可以從網頁中搜尋到評讀系統性文獻回顧的工具以及其他各種研究文獻報告的

評讀工具。

隨機對照試驗文獻的品質高低，還可以透過量化的工具來評量，其中 Jadad 品質評量表（Jadad Quality Score）較爲常用（Jadad et al., 1996）。該評估表共分成三大面向，以 0 分爲最低分，總分爲 5 分，其三大面向評估原則如下：

第一、研究是否被描述爲隨機？（Was the study described as randomized?）

□基本分 1 分：研究樣本是隨機分派的，但沒有詳細說明如何產生隨機的方式。

□加 1 分（即得 2 分）：有說明隨機分配的產生方式，如以電腦輔助產生的隨機試驗序列。

□減 1 分（即得 0 分）：有說明隨機分配的產生方式，但分配方式不恰當。

第二、研究是否被描述爲雙盲試驗？（Was the study described as double-blind?）

□基本分 1 分：僅提及實驗是採雙盲實驗，但未交待如何進行雙盲實驗。

□加 1 分（即得 2 分）：具體描述如何進行雙盲實驗的方法。

□減 1 分（即得 0 分）：描述如何進行雙盲，但方法不當。

第三、文獻是否有描述病患追蹤的退出、踢除原因？（Was there a description of withdrawals and drop-outs?）

□基本分 1 分：清楚說明失去追蹤（退出及失聯）的原因。

□減 1 分（即得 0 分）：沒有說明失去追蹤的原因。

結　論

不同的機構，往往有不同的文獻評估系統，例如美國預防工作小組的評估系統（US Preventive Service Task Force, USPSTF）即是將證據品質分爲三大類（US Preventive Services Task Force, 1996），和牛津大學的四級分類不同，因此我們在評讀時，需要知道評讀是參考那一個實證品質等級的分類方式。在這些不同的分類方式中，有一個共同的特色就是在治療型的研究中，大型而嚴謹的隨機對照試驗以及其系統性回顧，幾乎是文獻評讀中的金字塔頂端最值得被重視之處，即是 Archie Cochrane 30 幾年前就極力倡導的觀念。

參考文獻

1. Evidence-Based Medicine Working Group. (1992) Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA, 268*(17), 2420-2425.
2. Fisher, R. A. (1935). *The design of experiments*. Edinburgh and London: Oliver and Boyd.
3. Jadad, A. R., Moore, R. A., Carroll, D., et al. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials, 17*(1), 1-12.
4. Medical Research Council. (1948). Streptomycin treatment of pulmonary tuberculosis. *BMJ, 2*(4582), 769-782.

5. Oxford Centre for Evidence-based Medicine. (1998). Levels of Evidence and Grades of Recommendation from http://www.cebm.net/index.aspx?o=1047
6. US Preventive Services Task Force. (1996). *Guide to Clinical Preventive Services* (Baltimore: Williams and Wilkins ed.): 2 edition.
7. 梁繼權. (2003). 臨床醫學模式的典範轉移與實證醫學的發展 *臺灣醫學 7*(4), 531-534.