文獻評讀的重要性

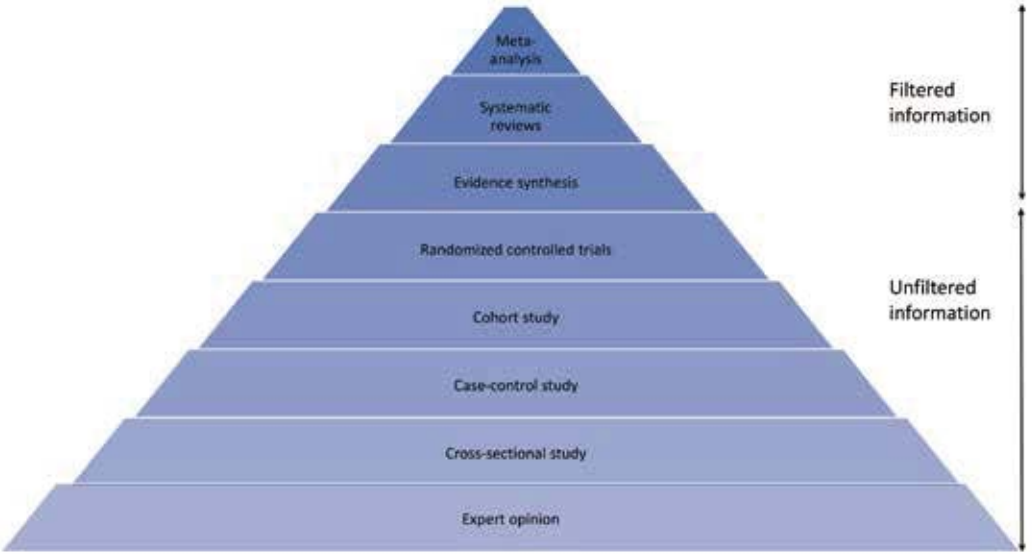
國立臺灣大學醫學院附設醫院 內科部 賴台軒

前言

現在是個資訊爆炸的年代，各種醫療資訊充斥讓人目不暇給。臨床工作者閱讀醫學文獻無非是要吸收醫學新知，以提供最佳的醫療給最需要的病患。然而近年來掠奪性期刊興起，造假論文風波不斷，到底我們閱讀到的醫學文獻是否真確，就成了一項重要的問題，這關係到病人的福祉，不能不慎重。試問：我們可否只閱讀所謂的高影響係數（impact factor）期刊？答案可能是否定的，一方面許多重要的論文都發表在專業特定的雜誌，其影響係數不見得高；另一方面高影響係數的期刊即使相對嚴謹與值得信賴，也不乏文章內容出現造假或是錯誤、最後遭受被撤回的命運。歷史上最有名的例子為重量級期刊刺胳針（Lancet）在1998年刊登了英國醫師Andrew Wakefield的研究，文中指出接種麻疹、腮腺炎、德國麻疹三合一疫苗（MMR）可能引起自閉症[1]。此文一出引起全世界家長的恐慌，導致疫苗接種率逐年下降，即使後來的研究均否定了這個相關性，仍然阻止不了反疫苗接種的風潮，最後導致歐美兩次大規模的麻疹爆發；雖然在2010年刺胳針撤銷了這篇論文，但是傷害已經造成[2]。因此嚴格的文獻評讀確有其必要性，而文獻評讀的技巧更是所有醫學相關人員應該具備的基本訓練。

證據等級

以實證醫學的精神而言，證據等級如同金字塔一般，越高代表證據等級越高。金字塔頂端為過濾過的資訊（filtered information），是有系統、專業地整合現有資訊而呈現的證據，包括證據整合（evidence synthesis）、統合分析（meta-analysis）、和系統性回顧（systematic reviews）等(圖一)。其次是未整理過的資訊（unfiltered information），



圖一 實證醫學證據等級金字塔

包括目前期刊上大多數的臨床研究均屬於此類。依其證據力由高到低分別如隨機試驗（randomized controlled trial）、世代研究（cohort study）、病例對照研究（case control study）、橫斷研究（cross sectional study）、和案例系列報告（case series report）。最底層則為專家意見（expert opinion）(圖一)。隨機試驗是以隨機分配的方式減少偏誤，將受試者分為實驗組與對照組，來看某種治療效果與副作用的實驗設計。臨床試驗據說於17世紀就已經存在，而隨機臨床試驗的實驗設計是於19世紀初由鼎鼎大名的Ronald A. Fisher所提出的；隨機試驗目前已經成為藥物或醫療技術上市前的基本要求、以及標準的研究模式。然而隨機試驗並不是完全沒有缺點的，樣本數足不足夠、隨機有無正確執行、有無進行雙盲或三盲等都會影響到試驗結果，因此讀者仍須做好文獻評讀才能正確解讀試驗結果。另外，隨機試驗不是萬靈丹，譬如一些危險的暴露就不能進行臨床試驗；舉例來說，抽菸對肺癌的危害就必須用世代研究的方法進行證實。世代研究又稱追蹤性研究或是縱貫性研究，是流行病學領域很重要的研究方法，顧名思義是在一群未患某特定疾病的群體中，看有某特定危險因子或暴露的族群，與沒有暴露的族群在一定時間的追蹤下的發病率；大規模的、嚴謹的、長期追蹤的世代研究其證據力其實不亞於臨床試驗。另一方面，若所探討的疾病發生率很低，則適合利用病例對照研究進行探討。專家意見在以往是重要的科學依據，但一方之言並無嚴謹的科學方法佐證，因此證據力是最低的。

文章的結構

在做文獻評讀前，首先我們要先了解文章的結構。標題（title）常常是看到文章的第一印象；標題不一定能明確告訴我們文章的內容，更不知文章的好壞，但是一個好的標題往往能吸引讀者的注意。接著是摘要（abstract），一個好的摘要會包含研究問題的產出，主要的研究方法與結果，和這篇文章的結論；摘要可能不見得能窺得研究全貌，但是很適合用來篩選文章。文章的主體則包括引言（introduction）、資料與方法（materials and methods）、結果（results）、和討論（discussion）數個部分。引言常依序是背景知識的介紹，研究的問題是甚麼，目前的科學上的缺口（scientific gap）在哪，最後提出問題的假說與預計進行何種研究。資料與方法部分以臨床試驗為例，應該要交代如何篩選受試者，包括納入與排除條件（inclusion and exclusion criteria），如何進行隨機分派，與實驗設計；如果是觀察性研究，則應告訴讀者母群體為何，如何從母群體取樣找到受試者，暴露、結果與各種變項的界定，以及採用的統計分析方法。結果應該會從描述性統計開始，讓讀者知道整個研究群體的分布與概況，而後才是研究分析的主要與次要結果。討論則是作者與讀者間的對話。好的討論會告訴我們這篇文章主要的發現，與之前的研究結果相同在哪裡，相異在何處，可能的機轉和未來的方向。

文獻評讀的方法

評讀文章要問三個問題，有一個口訣就是VIP。V是validity：這篇文章有效嗎？ I是importance：這篇文章重要嗎？P是population或是patient：這篇文章可以運用在病人嗎？以下分別針對各層面進行說明：

1. 何謂效度（validity）？效度指的是概念的定義（conceptual definition）及操作型定義（operational definition）間是否吻合。效度的測量很難達到，因為概念是抽象的，因此我們判斷研究是否具效度並不直接量測，而是設立指標來做判斷。效度還可分為外在效度（external validity）和內在效度（internal validity）；外在效度指的是研究的結果可否外推到母群體的效度，而內在效度則是研究結果在研究族群內的效度。舉例來說：母群體是末期腎病需透析的患者，而我們的臨床試驗是看65歲以下血液透析患者使用乙型阻斷劑對於心血管疾病的預後，本試驗在研究族群內各年齡層或有無糖尿病等是否有效為內在效度。而本試驗外推到65歲以上或是腹膜透析患者則是外在效度。

以臨床試驗研究為例，我們主要看幾個面向來評估研究是否有效力。第一是隨機分派（randomization）是否真確。隨機分派，顧名思義是將研究對象隨機分組，對不同組別實施不同的干預，看所觀測結果的差異。好的隨機分派可使兩組間無論是已知或是未知的變項趨近平衡，降低組與組之間比較的干擾，並確保後續統計檢定之效力。隨機的方法應該要明確寫在methods內容中，是用隨機表、電腦亂數、分層隨機法（stratification）還是區塊隨機分派（block randomization）？其中，區塊隨機分派是目前很常使用的方法，先決定區塊大小而後在區塊內進行隨機分派，如表一。這樣的好處是可使兩組人數相當，且相對於簡單隨機法而言更可靠。隨機分派如何執行、也是非常重要的一項環節，參與試驗人員和受試者都無法得知被分派到的治療方式，也就是分配隱藏（allocation concealment）必須受到確認。利用一個試驗中心執行電腦的分配隱藏是較完整的作法，也有人使用信封（envelope）等方法，但是否能確實做到隱藏會有疑慮。再來是盲

表一　區塊隨機分派的示意圖

治療方式的數目	2	A and B
Block的長度	4 (需是治療數的倍數)	
可能的排列組合	4!/2*2! = 6	AABB, ABAB, BAAB, BABA, BBAA, ABBA
產生一個隨機排列數字	1-6	253146
最後隨機分派的結果		ABAB BBAA BAAB AABB BABA ABBA

化設計（blinded），根據盲化的程度可以分為單盲（single blinded）(受試者不知)；雙盲（double blinded）(受試者與研究人員不知)；以及三盲(雙盲再加上資料監測者也不知)。盲化的設計主要是可以降低執行性偏差（performance bias），避免受試者或研究人員因知道接受的治療方式而產生對於結果可能的影響。

2. 第二個問題是這篇文章重要嗎？這個治療或是介入的強度如何，效應有多大？估計治療效果的估計是否精確？就統計意義而言，我們習慣上會看p值的大小或是信賴區間(confidence interval)的寬窄來看結果的可信度。但p值的錯誤解讀近年來一直被統計學家所提醒[3]。p值並非是虛無假設(null hypothesis)為真的機率，而0.05這個分界也只是個慣例，並無法二分為有意義或沒有意義；更重要的是p值大小並不代表這個效應有多大或多重要，這與效應樣本(effect size)有很大的關係，我們還需要看的是效果的估計值，包括相對的估計值：如相對危險性（relative risk）、相對危險降低度（relative risk reduction），以及絕對的估計值：如絕對危險降低度（absolute risk reduction）或是益一需要數（NNT, number need to treat），如此方能代表臨床上的重要性，如表二。舉個例子，使用某個藥物每30人就可以預防一次中風，相較於另一藥物每300人才可以預防一次中風，前者的臨床重要性更高。

3. 第三個問題是這個研究能應用在病人身上嗎？我們的病人與研究所收治的病人屬於同樣群體嗎？這個結果可以應用在我的病人嗎? 這個治療我們能夠實施嗎？我們的病人接受這樣的治療可能的好處或是壞處有甚麼？對於研究的結果是我們的病人期待或是想要的嗎? 這值得我們在文獻評讀時好好的思考。

文獻評讀的工具

由於文獻評讀要注意的事情相當多且繁瑣，因此國內外許多的組織單位都發展出各自建議文獻評讀的工具，提供給大家方便使用。

表二 評估重要性的指標

Abbreviation	Name	Meaning
CER	Control event rate	控制組某種病況的發生率
ERR	Experimental event rate	實驗組某種病況的發生率
ARR	Absolute risk reduction	|CER-ERR|
ARI	Absolute risk increase	|CER-ERR|
NNT	Number need to treat	1/ARR 避免一個病患發生某種病況所需治療的病人數
NNH	Number need to harm	1/ARI 給予治療多少病人數會發生一位病患受副作用所害

譬如CASP (critical appraisal skills programme)，這是一個英國的組織，提供各式研究方法的checklist可免費下載[4]。此外牛津大學的CEBM (center of evidence-based medicine)提供critical appraisal tools；針對臨床試驗的CONSORT statement (consolidated standards of reporting trials)等都是很好的文獻評讀工具[5, 6]。讀者可以一面閱讀文章，一面用評讀工具來做檢測，對於文獻的整理可收到事半功倍之效。

結語

Glasziou教授在一篇登在BMJ的文章“Evidence-based medicine and the medical curriculum”中提到，21世紀的醫師若不會文獻評讀，其臨床重要性就好像不會量血壓一樣[7]。在這個資訊倍增的時代，醫病雙方都會接觸到大量資訊，我們更應該做好文獻評讀的工作，找出嚴謹且有意義的研究，為我們的病人做出最好的決定。

參考文獻

1. Wakefield AJ, Murch SH, Anthony A, et al: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 1998; 351(9103): 637-41.
2. Retraction--Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 2010; 375(9713): 445.
3. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999; 130(12): 995-1004.
4. https://casp-uk.net/casp-tools-checklists/.
5. http://www.consort-statement.org/.
6. https://www.cebm.net/2014/06/critical-appraisal/.
7. Glasziou P, Burls A, Gilbert R: Evidence based medicine and the medical curriculum. BMJ 2008; 337: a1253.