醫學研究常用的統計方法與常見的統計缺失

比較 *New England Journal of Medicine* 與 *Nature Medicine*

嚴友君 統計分析師

在 2007 年的 *The American Statistician* 期刊中，有一篇文章回顧並審查 *The New England Journal of Medicine (NEJM)*及 *Nature Medicine (Nat Med)* 兩大醫學期刊中的文章。以這兩本頂尖的醫學期刊為例，整理在醫學研究中常用的統計方法與常見的統計缺失。本期 e-報將以這篇文章為基礎，讓讀者對醫學研究中常用的統計方法與常見的統計缺失有初步的瞭解。

The New England Journal of Medicine (NEJM) 是頂尖的臨床研究期刊，依據 Journal Citation Report 2012，Impact factor 為 51.66，在 "MEDICINE, GENERAL & INTERNAL" 領域 155 本期刊中排名第 1；*Nature Medicine (Nat Med)* 則是頂尖的基礎研究期刊，依據 Journal Citation Report 2012，Impact factor 為 24.30，在"MEDICINE, RESEARCH & EXPERIMENTAL" 領域 121 本期刊中排名第 1。

文章回顧包含 2004 年上半年所有的原創研究論文 (original research articles)，不包含 editorials, letters, case reports 與 review articles。其中 *NEJM* 為 Volume 350, No.1-26，共 91 篇原創研究論文；*Nat Met* 為 Volume 10, No. 1-6，共 34 篇原創研究論文。統計分析使用雙尾費雪精確檢定 (Fisher's exact tests)，顯著水準設在 0.05。精確 95% 信賴區間 (Exact 95% confidence intervals) 以 Clopper and Pearson 方法計算。

常用的統計方法

系統性的記錄這 125 篇文章所使用的統計方法後，將統計方法分為以下幾類: 沒有統計方法、只有描述統計、*t* 檢定 (*t* tests)、基礎列聯表分析 (卡方 (χ^2) 檢定、費雪精確檢定 (Fisher's exact tests))、進階列聯表分析、無母數檢定 (nonparametric tests)、基礎變異數分析 (單因子變異數分析 (one-way ANOA))、進階變異數分析 (advanced analysis of variance)、相關係數 (correlation coefficients)、基礎迴歸 (簡單線性迴歸 (simple-linear regression))、進階迴歸 (advanced regression)、流行病學方法 (epidemiologic methods)、存活分析 (survival analysis)、其他方法、無法判定的方法 (unidentified method/test)。

表一節錄 *NEJM* 與 *Nat Met* 這兩本期刊在 2004 上半年的文章中，最常用的幾種統計方法。

表一 節錄常用的統計方法

	New England Journal of Medicine (總文章數 = 91)		*Nature Medicine* (總文章數 = 34)	
	n	%	n	%
推論統計	86	94.5	28	82.4
t 檢定	32	35.2	14	41.2
卡方檢定	32	35.2	0	0.0
變異數分析	12	13.2	10	29.4
存活分析	39	42.9	4	11.8
無法判定	1	1.1	10	29.4
信賴區間	61	67.0	0	0.0

註：每篇文章可能使用超過 1 種以上的統計方法

在 *NEJM* 的 91 篇文章中，94.5% (95% 信賴區間 87.6–98.2) 的文章內容有含蓋推論統計；而在 *Nat Med* 的 34 篇文章中，則有 82.4% (95% 信賴

區間 65.5– 93.2) 的文章內容有含蓋推論統計。在使用推論統計的百分比上，兩個期刊間沒有統計上的顯著差異 (p 值 = 0.068)。

在 *NEJM* 中的文章，最常使用的推論統計方法為存活分析 (survival analysis: Kaplan-Meier, Mantel-Cox, log-rank test, life table analysis)，佔 42.9%；其次為基礎的 *t* 檢定與卡方檢定，均佔 35.2%。就 *Nat Med* 而言，發表文章中最常見的統計推論方法為 *t* 檢定，佔 41.2%；其次為變異數分析，佔 29.4%。然而在 *Nat Med* 中，卻有 29.4% (10 篇) 被歸類為使用“無法判定的方法”；這個分類在 *NEJM* 中僅有 1.1% (1 篇)。被歸類為使用“無法判定的方法”是因為文章中有呈現 p 值卻沒有說明 p 值是應用了何種統計方法或程序。

在 *NEJM* 中有 67.0% (61 篇)有報告信賴區間，但在 *Nat Med* 的文章均沒有呈現信賴區間 (0.0%)。

另一個關於基礎研究期刊與臨床研究期刊的不同在於量測尺度 (scale of measurements) 上。*Nat Med* 多使用連續 (continuous) 資料。在調查的 *Nat Med* 文章中，沒有發現任何 1 篇使用卡方檢定；而在 *NEJM* 中，幾乎每 2 篇就有 1 篇包含使用卡方檢定 (46.2%)。在基礎科學，*t* 檢定仍然是最常應用的推論統計方法 (41.2%)。然而，由於臨床研究多為觀察性 (observational) 研究，因此較多使用順序(ordinal)或名目(nominal)尺度的資料。

現代醫學研究廣泛的使用到推論統計，比起 1983 年的調查，2004 年的 *NEJM* 文章大量的增加了推論性統計的使用。在 1983 年，只有 42.0%的文章具有分析的特性，而 2004 年則增加到 94.5%，增加的比例超過 2 倍以上。特別是存活分析的使用，在 1983 年幾乎沒有看見，而 2004 年時，存活分析在 *NEJM* 已經是最常使用的方法 (42.5%)。

統計分析的複雜度

將描述性統計以外的統計分析依其應用統計技巧的複雜程度，分為“基礎分析”(Basic Analyses) 或 “進階分析”(Advanced Analyses)。“基礎分析”包含 *t* 檢定、基礎列聯表分析(卡方檢定或費雪精確檢定)、基礎無母數方法、單因子變異數分析、簡單相關及線性迴歸方法。若文章中有使用到任一種其他較複雜的統計技巧，如多變量分析 (multivariate analysis 例如 MANOVA、MANCOVA)、統計模型(statistical modeling)、進階列聯表分析、流行病學統計量或存活分析等，則歸類為“進階分析”。

表二 統計分析的複雜度

	New England Journal of Medicine (總文章數 = 91)		*Nature Medicine* (總文章數 = 34)	
	n	%	n	%
無/只有描述性/無法判定	6	6.6	12	35.5
基礎分析	15	16.5	15	44.1
進階分析	70	76.9	7	20.6

在 *NEJM* 中，只有 16.5%的文章被歸類為 “基礎分析”，是指文章中僅使用基礎的統計技巧，如 *t* 檢定、卡方檢定、費雪精確檢定、基礎無母數檢定、單因子變異數分析或簡單線性迴歸與相關；而有 76.9% (95% 信賴區間 66.9–85.1) 的文章使用 1 種或 1 種以上較為進階的統計方法，被歸類為 “進階分析” 。在 *Nat Med* 中，被歸類為 “進階分析” 的文章佔 20.6% (95% 信賴區間 8.7–37.9)。*NEJM* 在統計分析的複雜度與進階程度顯著的高於 *Nat Med* (p 值 < 0.0001)。

在 *NEJM* 中，統計分析的複雜度有隨著年代的演進而愈趨複雜，但在 *Nat Med* 中卻沒有同樣的發現。這可能是由於在基礎研究和臨床研究的統

計需求有所不同。在 *Nat Med* 中，發表的文章以動物研究為大宗，而 *NEJM* 中的文章主要是人體研究。典型的動物實驗在實驗設計上多屬於高度侵入性、且使用基因上相同的物種、研究時有較少的個體間差異，減少了使用多變量分析方法控制可能干擾效應的需要。然而在臨床研究上，干擾因子是必需特別討論與控制的重要議題。同樣的原因也導致在 *NEJM* 中較常使用的進階統計方法(如存活分析) 很少出現在 *Nat Med* 中。另外，動物研究一般樣本數較少，這也進一步降低使用複雜統計技巧的可能性。

常見的統計缺失與錯誤

在評估常見的統計缺失與錯誤方面，使用標準化的 46 項檢核表來檢核。評估主要針對統計顯著檢定 (statistical significance testing)，其他也包含研究設計、統計分析、使用統計方法的說明記錄、研究發現的呈現及解釋等。這部分的評估只囊括部分的文章，其納入條件為：(1) 有使用到除了描述統計以外的統計推論方法；(2) 在評估主要結果時有使用基本的統計顯著檢定 (例如 t 檢定、卡方檢定、費雪精確檢定、 Mann-Whitney-U-test、Wilcoxon test 等)。若文章僅針對進階統計技巧或統計模型，由於進階統計方法沒有包含在檢核表內，故進一步的統計缺失評估會排除這部分的文章。在 125 篇中，總計納入其中 53 篇 (*NEJM* 31 篇，*Nat Med* 22 篇)，對文章中統計方法的使用做更詳細的品質評估，以瞭解醫學統計上常見的統計缺失與錯誤。只有在沒有疑問、明顯的議題且能明確判定者，才會被歸類為“發生錯誤” (error committed)。

表三 節錄常見的統計缺失與錯誤

	NEMJ (文章數 = 31)		*Nat Med* (文章數 = 22)	
	n	%	n	%
研究設計				
無樣本數/檢定力計算	13	41.9	22	100.0
資料分析				
使用錯誤或較不適合的統計檢定	5	16.1	6	27.3
沒有校正多重比較 (multiple comparison)	11	35.5	6	27.3
方法說明				
沒有清楚且正確的定義或具體說明所有使用的統計檢定方法	20	64.5	20	90.9
結果呈現				
提供標準誤(SE)而不是標準差(SD)來描述資料	8	25.8	16	72.7
呈現 “p<0.05”、 “p>0.05”等，而非實際的 p 值	6	19.4	19	86.4
沒有提供主要效應值 (main effect size measures) 的信賴區間	14	45.2	21	95.5
結果解釋				
當結果為不顯著時，忽略討論型 II 錯誤	5	16.1	5	22.7
有多重顯著檢定時，沒有討論可能的問題	10	32.3	6	27.3

與研究計設相關的缺失中，最常見的是沒有做樣本數估計或計算檢定力。在審查 *Nat Med* 的 22 篇文章中，所有的文章都沒有包含樣本數或檢定力計算的說明 (100.0%)，這可能是由於此期刊的編輯方針沒有這方面的要求。在 *NEJM* 中，有 5 篇 (16.1%) 文章“使用錯誤的或較不適合的統計檢定”，是由於使用的檢定方法與分析的資料不相配、使用不適切的母數方法 (parametric method)、或是所使用的統計方法不適合欲研究的科學假設 (hypothesis)。在 *Nat Med* 中，犯這個錯誤的文章比例稍高，佔 27.3%。其他常見的缺失包括“使用平均數的標準誤 (standard error of the mean, SEM) 來描述研究資料的分散程度”以及“在呈現 p 值時以主觀的臨界值表現 (如 “p < 0.05” 、 “p = ns”) 而不是實際的值”。在解釋研究發現方面相關

的常見統計錯誤包含“當結果為不顯著時，忽略討論型 II 錯誤”及“有多重研究標的 (multiple study endpoints) 時，沒有討論多重檢定的問題”。

由於受審查的文章中有很多沒有對所使用的統計方法提供說明與紀錄，如此，要審查統計分析及研究結果闡述的正確與適合，就顯得非常的困難。尤其是 *Nat Med* 中的文章，22 篇中有 20 篇 (90.9%) 沒有清楚且正確的定義或具體說明所有使用的統計檢定方法，這使得之後的逐項檢核非常的不容易。由於沒有足夠的資訊，對於檢核清單中的“使用錯誤或較不適合的統計檢定”以及“提出的結論沒有研究資料支持”這兩個項目的審查最為困難。除了以上兩個項目外，有 24 篇在初次審查時被歸類為“無法評定/不清楚”。針對這 24 篇，即使再經過第二位統計學家的獨立審查，仍然有近 50%無法評定。

就如同使用適當的統計方法一樣重要，足夠詳細的敘述所使用的統計方法，讓讀者能重新計算重要的研究發現是亦是研究論文發表的重點之一。但是，在 *Nat Med* 卻有超過 90%以上的受審查文章沒有做到這一點，這些文章的作者沒有適當地清楚且正確的具體定義或說明所有使用的統計檢定方法。另外，在 *Nat Med* 中的文章，應更加強調實驗處置差異的強度與數量以及統計估計的方法，而不是僅僅依賴統計顯著檢定。

在 *NEJM* 及 *Nat Med* 發現的統計錯誤與缺失大部分不是那麼嚴重，而且大多與說明不充分有關。但是有特別發現某一個特定的錯誤較常發生在某個特定的期刊的情形。特別是錯誤的使用平均數的標準誤 (SEM) 來描述研究資料的分散程度，似乎是在 *Nat Med* 非常常見的問題，幾乎每 4 篇文章中就有 3 篇 (72.7%) 發現這個統計錯誤。雖然一般而言這些錯誤並不一定必然導致不正確的結論，但是由於信度、效度與研究結果可驗證度的降低，會為一個研究專案的科學影響帶來負面的效果。只要稍微加以注意，

不論是基礎研究者或臨床研究者都可以很容易地避免以上所提的錯誤。

統計學家的角色及參與程度在基礎研究與臨床研究似乎也有很大的不同。臨床研究者傾向經常且及時的統計諮詢，然而基礎研究者在計畫及進行研究的時候通常不會尋求專業的統計協助。特別是在基礎科學，如果在期刊編輯及審查時，統計學家有較多的參與，應該能明顯增加科學研究的品質。統計學家應該在更早的階段就參與研究，不論在研究設計、資料蒐集及資料管理都應承擔重要的責任，則應可有效的避免以上發現的統計誤用及缺失。

除了少部分高影響力的臨床研究期刊，大部分的醫學研究論文都沒有接受統計審查，而且也只有少部分的醫學期刊有公佈給作者的統計指引。*Nat Med* 的編輯在 2005 年 (Vol. 11, No. 1, p. 1) 特別討論了他們刊出的研究論文在統計上過於草率的問題，之後 *Nature* 系列期刊也提供了給作者的統計檢核表(Nature's Statistical Checklist for Authors)。以上統計品質的深入評估結果建議我們，在發表前期準備手稿時，各期刊應有更清楚的統計方針、給作者更明確的說明以及在編輯上更密切的注意統計方法，以提升醫學論文發表的品質。

結語

最後，同樣引用 American Statistical Association 的 Ethical Guidelines for Statistical Practice (Approved by the Board of Directors, August 7, 1999)的一段，給所有參與醫學研究發表的研究人員： *[...] The use of statistics in medical diagnoses and biomedical research may affect whether individuals live or die, whether their health is protected or jeopardized, and whether medical science advances or gets sidetracked. [...] Because society depends on sound statistical practice, all practitioners of statistics, whatever their training and*

occupation, have social obligations to perform their work in a professional, competent, and ethical manner.

參考資料與延申閱讀

Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P., Ulmer, H. (2007), "The Use of Statistics in Medical Research," The American Statistician, 61, 47–55. doi: 10.1198/000313007x170242

Emerson, J. D., and Colditz, G. A. (1983), "Use of Statistical Analysis in the New England Journal of Medicine," New England Journal of Medicine, 309, 709–713.

Editorial. (2005), "Statistically Significance," Nature Medicine, 11, 1.

Nature's Statistical Checklist for Authors
http://www.nature.com/nature/authors/gta/Statistical_checklist.doc

American Statistical Association. (1999), "Ethical Guidelines for Statistical Practice," http://www.amstat.org/about/ethicalguidelines.cfm

Journal Citation Report 2012, Institute of Scientific Information, Thomson Corp.