測量誤差與資料解釋

謝清麟教授 Sep. 2019
臺大醫學院職能治療學系

Guiding questions:

1. 評估之於臨床 OT 之意義為何？
2. 評估誤差之來源為何？
3. 目前臨床執行評估之瓶頸為何？
4. 未來如何突破目前臨床評估之瓶頸？
5. 你覺得評估對於「臨床效能」或「臨床專長之培養」重要嗎？

測量或評估是臨床治療、掌握病情的第一步，也是臨床決策與推動實證醫學的根基。就如「對症下藥」之臨床原理，理論上，欲「對症」就須執行完整、精準的評估，以掌握全面、零誤差的評估結果。然而舉凡測量皆有誤差，臨床評估亦然。嚴重的測量誤差，將深切影響臨床評估所得資料的解釋，也干擾醫療人員對於病情的掌握與臨床決策。

以下就「評估工具」及「以全人為觀點--全面/個別化評估」之評估誤差，分別闡述其如何影響資料解釋。

評估工具皆有系統誤差與隨機誤差

就個別評估工具而言，誤差的來源有二: 系統誤差與隨機誤差。使用者必須掌握所使用評估工具的誤差大小，否則難以解釋評估所得分數，究竟誤差多少。

系統誤差源自特定的因素，有系統地影響測量結果。這些因素主要包含評估工具之設計問題、執行問題等。以 ADL/IADL 概念及評估為例，一般 OT 於臨床上，鮮少使用標準化評估工具，而使用機構自行設計之評估工具。然而自行設計之評估工具的理論基礎以及心理計量特性(如信度、效度等)往往付諸闕如，造成評估結果因評估工具之設計問題，產生系統性偏誤，且偏誤程度亦因缺乏驗證或不易驗證，而難以估計。也就是評估結果究竟為何，難以確認，亦無法校正。因此使用非標準化或未經嚴謹驗證之評估工具，其測量誤差難以掌握，遑論資料解釋。

在 ADL/IADL 的概念上，至少有二種：「實際執行 ADL 能力(capability)」與「日常 ADL 表現(performance)」，需要明確區隔。定義上，「實際執行 ADL 能力」代表個案親身從事 ADL 各項目之能力高低。治療師通常請個案於治療室實際從事每項 ADL 活動，治療師再從旁觀察個案執行每項 ADL 活動之過程，藉以評估個案各項 ADL 之能力。「日常 ADL 表現」代表個案平常在家或病房實際

從事 ADL 之情形或依賴程度。治療師因無法實地觀察，通常以訪談方式得知個案 ADL 之平常表現。以上二種概念，理論上，不難區辨。但實務上，必須各別評估、分別紀錄，以免混淆不清。尤其二種概念所代表之意義（資料解釋）不同，評估「實際執行 ADL 能力」以掌握個案執行 ADL 之困難，有助於設計治療活動或給予輔具。但「日常 ADL 表現」代表個案之依賴程度，也是常用之療效指標。許多 OT 皆遇過一些個案能力足以於治療室從事某項 ADL，但在病房或家中，卻依賴看護或家屬。這樣的現象，也代表二種概念不一。如果未能區隔 ADL 概念之差異，必然導致系統誤差，嚴重影響評估結果以及資料解釋。

執行 ADL 評估過程亦可能造成的系統誤差，包含如訓練不足，有些施測者評分標準較嚴苛，導致分數一致地偏低。或是評估工具的重複使用，造成學習/練習效應，導致再測分數變高。一般的認知評估工具，皆易受到學習效應影響，除了造成系統誤差，也嚴重干擾評估結果之解釋，使用者難以區辨個案之變化，是因為學習效應，還是真實的認知功能變化。

隨機誤差則是由隨機、偶然的原因所造成。例如評估情境的吵雜或改變、評估人員評估個案時的心情好壞、照護人員的干擾施測等，這些不可避免的偶然因素造成評估結果的波動，引起隨機誤差，有時大，有時小，有時正，有時負。因此隨機誤差的大小，影響資料解釋甚鉅。

隨機誤差雖不易掌控，但可估計。相關指標如測量標準誤 (standard error of measurement)、最小可偵測之變化值 (minimal detectable change)等。臨床各種評估工具之隨機誤差估計，也是近期國內外研究潮流之一，詳第 4-6 頁（評估誤差之深入介紹）。

<u>以全人為觀點的評估，執行關鍵為評估範疇之個別化及完整性</u>

以全人為觀點的評估，符合以個案為中心、個別化醫療的當今醫學主流。評估誤差的來源為：評估範疇之個別化程度及評估範疇之完整性。施測者必須掌握個案之特性（包含病情、個人特質/經驗、期待等）並與個案/家屬充分溝通後，共同設定治療目標，也同時決定療效指標之評估範疇。

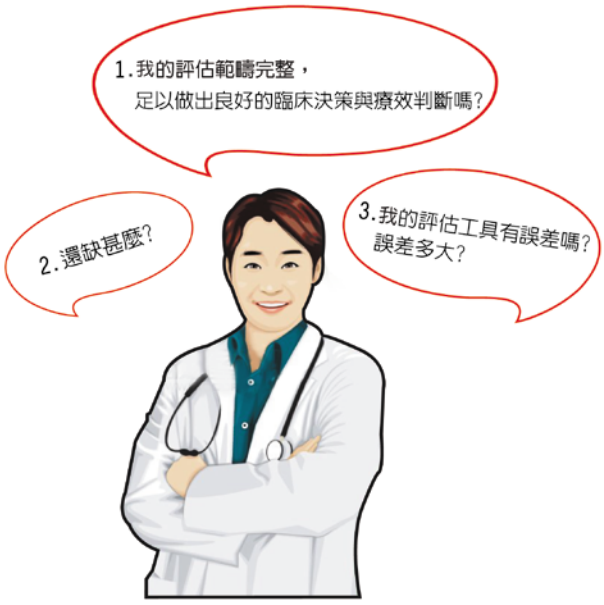
由於全人觀點的評估，牽連廣大，過程較為繁複，也可說是理想化的評估與治療模式。必須結合相關專業人員，以個案為中心、群力合作，醫病間充分溝通且務實考量個案病情及期待，始能達成以個案為中心、個別化之最佳醫療。反言之，我們若無法做到以個案為中心，顧及個案之個別特性與需求，醫病間同床異夢，醫療人員所做的評估，很可能不是病人所需或問題所在。除了浪費醫療資源外，個案也難以理解/認同評估內容與結果，造成個案認為評估非必要，進而影響醫病關係與臨床效能。

綜合以上，臨床評估之誤差實屬必然，無法避免。臨床人員唯有掌握評估誤差之大小，始能做出精確的資料解釋與臨床決策。以個別評估工具而言，唯有評估分數的改變超過評估誤差時，始能代表所評估的特質產生真正的改變，也才能宣稱患者進步或退步，而後續的臨床決策與治療計畫才得以進行。以全人為觀點而言，個別化且全面的評估，符合個案之需求，亦能提升醫療效率、醫病關係以及醫療滿意度。

目前國內多數臨床人員仍以「臨床經驗」為掌握病情、臨床推理及判斷療效之主要依據，並未有系統/全面地使用標準化評估工具，因此更難掌握評估之誤差。作者個人認為造成此現象的主因，除了臨床過於忙碌、時間有限、評估工具不夠全面完整、費時評估、評估標準不一、且誤差過大，甚至超過臨床人員之經驗判斷。可能因為上述因素，造成一些臨床人員全面或部分棄用標準化評估工具。作者認為目前國內 OT 各領域的評估工具，難以克服上述問題。也就是棄用標準化評估工具之現象仍將繼續存在。然而「臨床經驗」過於主觀，難以估計誤差、也造成專業人員間之溝通困難，這些現象勢將影響治療師之能力養成，亦影響治療成效之提升。

最後如前所言，如果我們無法發展適合臨床所需之評估工具，藉以全面掌握個案特質、甚至降低評估誤差，將侵蝕掌握病情與臨床決策之根基，也難以推動實證醫學。所以個人認為 國內 OT 各領域的評估工具之範疇與誤差問題 是目前 OT 發展的關鍵瓶頸之一，即使國內外之實證證據再多，我們未能掌握個案之病情、對症治療，OT 焉能有效。因此我們必須掌握或發展最適合臨床使用與低誤差之各種評估工具，以突破瓶頸、造福個案。

另外，臨床人員可用於評估之時間，通常相當有限。因此電腦適性測驗 (computerized adaptive testing) 或可解決此問題，將於此文最後簡述之。



OT於使用評估工具與資料解釋時之考量要點

評 估 誤 差

在臨床上，大致每週或隔週必須對於接受治療的患者進行再評估，依據再評估的結果判斷治療進展，修改治療計畫。然而，所有評估所得之數據皆包含評估誤差。評估誤差根據來源不同，可分為系統誤差與隨機誤差。[1] 系統誤差是由某種固定的原因造成的，使評估結果系統性地偏高或偏低，當重複進行評估時，它會重複出現，例如評估人員評估時所使用的工具不準確，系統地導致評估結果總是偏高或偏低。隨機誤差則是由隨機的偶然的原因造成的，例如評估時情境的微小改變、評估人員評估每位個案時的微小差別等，這些不可避免的原因造成評估結果在一定範圍內波動，引起隨機誤差，有時大，有時小，有時正，有時負。系統誤差影響的是評估工具之效度，隨機誤差影響的是評估工具之信度，故在此「評估誤差」章節中所提之評估誤差指的是隨機誤差，而非系統誤差。

掌握每一評估工具的評估誤差，有助於評估結果的解釋與臨床決策的制定，[2] 唯有評估結果的改變超過評估誤差時，才能代表所評估的特質產生真正的改變，也才能宣稱患者進步或退步，而後續的臨床決策與治療計畫才得以進行。以下將介紹說明各種評估誤差之估計。

1. 評估標準誤(standard error of measurement, SEM)

SEM 為信度指標其中一種，SEM 代表個別評估結果之不穩定程度或隨機評估誤差大小。SEM 數值即用來解釋個別分數的評估誤差的大小。[3-5] 由圖一可知，當我們對個案評估一次巴氏量表，即為此分佈中的一點，由於評估工具或被評特質的不穩定性，評估多次之後即形成了圖一的常態分佈。因此，SEM 的估計公式如下：

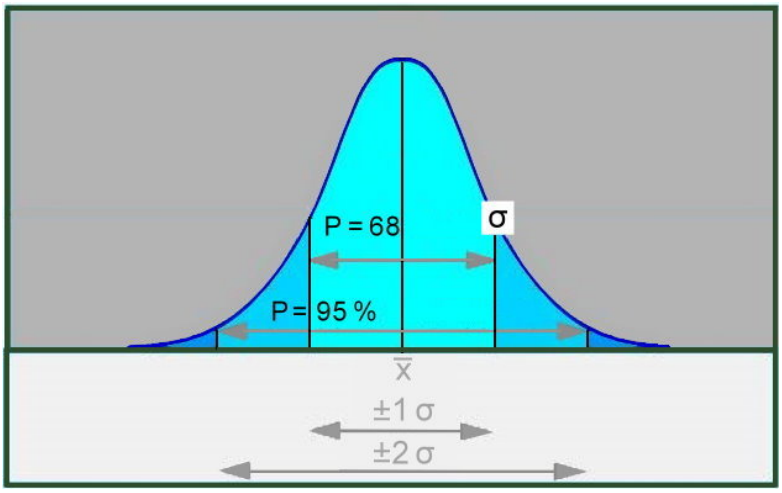
$$SEM = SD_{baseline} \times \sqrt{1-R}$$

其中 R 為 ICC 值，即由再測信度的數值推估個案評估多次之誤差帶(error band)，即其 68%信賴區間或 95%信賴區間。[3, 6]

所謂的信賴區間是一種估計數值的方法，估計數線上的一個區間，而非一個點。例如以信賴區間的概念而言，欲表示群體之評估分數，例如這群患者的巴氏量表平均分數落在 8 至 12 分之間，同時更近一步地說明母群體個案的巴氏量表平均分數落在在 8 至 12 之間的機率為 95%。因此信賴區間是一個區間，而非一個點。

以圖一說明信賴區間的概念，如果我們抽取樣本多次，求得許多個樣本平均數($\bar{x}$)，則這些所構成的次數分配圖將如圖一所示為常態分配。因此將有 68%的$\bar{x}$的值在 $\mu \pm 1\sigma_x$ (z = 1)(μ：母群體平均數；σ_x：母群體平均數之標準差，即標準誤)之間；95%的$\bar{x}$的值在 $\mu \pm 2\sigma_x$ 之間(z = 2)；68%的$\bar{x}$的值在 $\mu \pm 1\sigma_x$ 之間(z = 1)。依據常態分配，可再進一步求出，$\bar{x}$的值在 $\mu \pm 1.96\sigma_x$ 之間佔 95% (z = 1.96)；$\bar{x}$的值在 $\mu \pm 2.58\sigma_x$ 之間佔 99% (z = 2.58)。換句話說，如果有 95%的$\bar{x}$落在 μ 的左右 1.96 個標準誤之內，則在任何一個$\bar{x}$減

去或加上 1.96σ_x，則會得到一個($\bar{x}$-1.96σ_x)至($\bar{x}$+1.96σ_x)的區間，有 95%的機會(即 95%的信心水準)μ 落在此區間中。



圖一、常態分佈圖

2. **最小可偵測之變化值**(minimal detectable change, MDC)

最小可偵測之變化值可用以判斷個案改變的分數是否超過隨機評估誤差。[6] 因為個案至少被評估 2 次，若要具有 95%的信心水準宣稱個案的變化超過評估誤差，則需要加上 2 次評估所增加之誤差，SEM × 1.96 × √2，因此 MDC 的估計方式為：

$$MDC = SEM \times 1.96 \times \sqrt{2}$$ [6]

在 MDC 解釋上，若評估工具的 MDC 相對於評估工具總分低於 10%，可視為良好可接受的評估誤差。[7] 目前最小真正改變量之詞彙尚未一致，有許多同義詞彙皆代表可偵測之變化值，如 minimum detectable change, minimal detectable difference, smallest real difference, smallest detectable difference 等。

MDC 的意義代表臨床上單一個案的前後兩次評估分數改變值須超過 MDC 值的改變才有 95%的信心水準宣稱：超過評估誤差。因此 MDC 可提供臨床與研究人員接受治療後是否具有改變的參考標準。例如，中風患者於伯格氏平衡量表（Berg Balance Scale, BBS）與中風病人姿勢控制量表（Postural Assessment Scale for Stroke patients，PASS）之 MDC 分別為 6.7 與 3.2，[8] 其意義代表臨床上中風患者在接受治療後，於 BBS 及 PASS 分數的改變必須超過 6.7 與 3.2，才能宣稱患者於平衡功能之進步超過評估誤差，因此患者已產生真實的平衡能力變化；反言之，若中風患者於 BBS 及 PASS 分數的改變未超過 6.7 與 3.2 分，則此改變可能僅是評估誤差造成，而非患者平衡功能的改變。

根據不同的信心水準，評估工具有不同的 MDC 值，以上述中風患

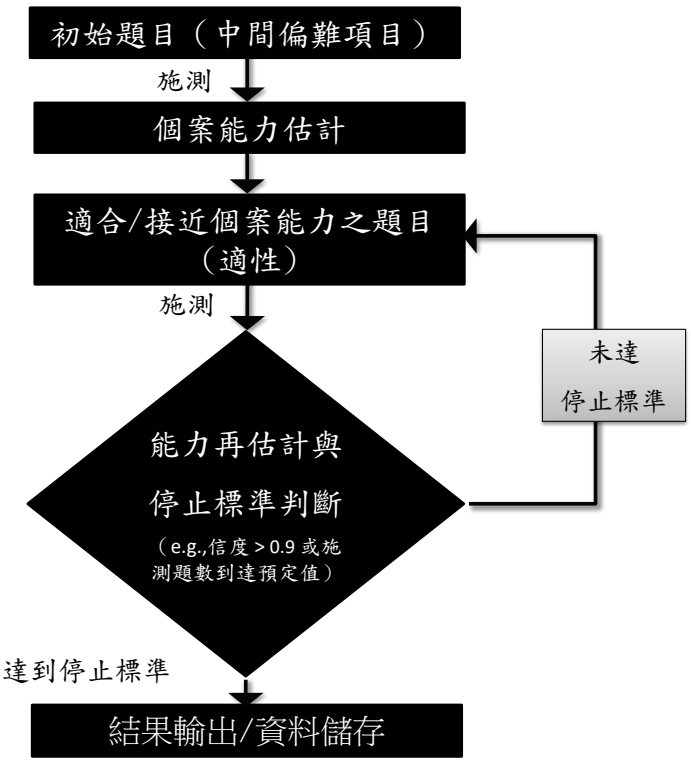
者 BBS 及 PASS 之 MDC 值為例，如表五。因此，中風患者前後測的 BBS 分數差異若為 6.7 分，PASS 分數差異若為 3.2 分，則有 95%的信心水準聲稱個案進步；前後測 BBS 分數差異若為 2.9 分，PASS 分數差異若為 1.4 分，則有 60%的信心水準聲稱個案進步。

表五、不同信心水準之不同MDC值

z值	信心水準	BBS之MDC	PASS之MDC
2.58	99%	8.8	4.2
1.96	95%	6.7	3.2
1.64	90%	5.6	2.7
1.28	80%	4.4	2.1
1.04	70%	3.5	1.7
0.84	60%	2.9	1.4

電腦適性測驗

電腦適性測驗 (computerized adaptive testing, CAT) 具備三大特性，可達成快速且精準之評量：(1)精準有效率：CAT 評量系統所需評量之項目精簡，大量縮短臨床評估時間，可大幅降低施測者及個案的負擔，提昇評量之效率。多向度測驗內容（如評量平衡、ADL 與動作功能等）也可藉由各向度之關聯性快速推估個案之各種向度之功能狀態，進而縮減評量項目，快速完成施測。[9, 10] 各向度間的關聯程度愈高，則可簡化的項目愈多，評量效率愈高，但又不喪失評量精準度。當評量項目的內容與個案狀況越接近時，評量精準度就越高，施測人員可以先決定評量精準度（信度）標準，評量過程中一旦達到此信度時，即終止評量（圖二）。因此 CAT 能以少量項目，得到精確的評量結果，故 CAT 評量系統可提供快速暨完整評量中風患者認知功能之理想施測技術。(2)提供個別測量誤差估計值：CAT 奠基於項目測驗理論 (item response theory)，故於個案完成施測後，可同時估計個案分數之 SEM 及信度，可提供傳統測驗缺乏的測量資訊，協助臨床人員掌握個別患者的認知功能是否發生實質改變 （非因測量誤差產生之分數改變），提供醫療決策與判定療效之依據。(3) 提升資料處理之效率：CAT 可立即呈現評量結果並且數位化儲存。



圖二：電腦適性測驗之施測流程

CAT 技術早期運用於教育與心理測驗中，如：托福測驗 (Test of English as a Foreign Language, TOEFL) 及 GRE (Graduate Record Examination) 等。近期於復健或健康相關領域亦逐漸採用，例如測量中風病患日常生活活動的 ADL CAT，[11] 小兒的 Pediatric Evaluation of Disability Inventory，[12] 或 Activity Measure for Post-Acute Care、Participation Measure for Post-Acute Care 等。[13, 14] 國內研究人員亦已發展出平衡功能 CAT 以及日常生活功能 CAT。[11, 15] 有興趣者可至 http://13.114.225.208/cat/ 試用。

References:

1. Portney LG, Watkins MP. *Foundations of clinical research: Applications to practice*. NY: Pratice Hall Health; 2000.
2. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. *Am J Phys Med Rehabil*. 2005;84:719-723.
3. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26:217-238.
4. de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. *Int J Technol Assess Health Care*. 2001;17:479-487.

5. Safrit MJ, Wood TM. *Measurement concepts in physical education and exercise science*. Champaign (IL): Human Kinetics; 1989.
6. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res*. 2001;10:571-578.
7. Flansbjer UB, Holmback AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med*. 2005;37:75-82.
8. Liaw LJ, Hsieh CL, Lo SK, Chen HM, Lee S, Lin JH. The relative and absolute reliability of two balance performance measures in chronic stroke patients. *Disabil Rehabil*. 2007:1-6.
9. Hsiao YY, Shih CL, Yu WH, Hsieh CH, Hsieh CL. Examining unidimensionality and improving reliability for the eight subscales of the sf-36 in opioid-dependent patients using rasch analysis. *Qual Life Res*. 2015;24:279-285.
10. Chou CY, Chien CW, Hsueh IP, Sheu CF, Wang CH, Hsieh CL. Developing a short form of the berg balance scale for people with stroke. *Phys Ther*. 2006;86:195-204.
11. Hsueh IP, Chen JH, Wang CH, Hou WH, Hsieh CL. Development of a computerized adaptive test for assessing activities of daily living in outpatients with stroke. *Phys Ther*. 2013;93:681-693.
12. Haley SM, Ni P, Ludlow LH, Fragala-Pinkham MA. Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Arch Phys Med Rehabil*. 2006;87:1223-1229.
13. Haley SM, Gandek B, Siebens H, Black-Schaffer RM, Sinclair SJ, Tao W, et al. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: Ii. Participation outcomes. *Arch Phys Med Rehabil*. 2008;89:275-283.
14. Haley SM, Siebens H, Coster WJ, Tao W, Black-Schaffer RM, Gandek B, et al. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: I. Activity outcomes. *Arch Phys Med Rehabil*. 2006;87:1033-1042.
15. Hsueh IP, Chen JH, Wang CH, Chen CT, Sheu CF, Wang WC, et al. Development of a computerized adaptive test for assessing balance function in patients with stroke. *Phys Ther*. 2010;90:1336-1344.