財團法人醫藥品查驗中心 Center for Drug Evaluation, Taiwan 當代醫藥法規月刊 2018-03-09 89
回 CDE 官網首頁
訂閱/取消電子報

美國 FDA「臨床試驗多重評估指標指引草案」－多重評估指標之釋義與影響

宮玫芬[1]

前言

美國食品和藥物管理局(US FDA)於 2017 年一月發表「臨床試驗多重評估指標指引草案(Draft Guidance for Industry: Multiple Endpoints in Clinical Trials)」。此指引草案在延宕數年之後終於定案。在此之前，歐盟曾於 2002 年頒布「臨床試驗多重檢定問題考量重點(Points to Consider on Multiplicity Issues in Clinical Trials)」，又於 2016 年十二月發表更新版，以「臨床試驗多重檢定問題準則草案(Draft Guideline on Multiplicity Issues in Clinical Trials)」取代 2002 年版之考量重點。歐美兩大法規體系於近年競相發布臨床試驗多重檢定考量之準則或指引，均為了因應各方迫切的需求。

歐盟之多重檢定準則草案，為較具原則性之廣泛論述；美國 FDA 之多重評估指標指引草案，則提供了較具教育性的方法學說明，具體陳述多種解決多重檢定問題的統計調整方法，及各自的優缺點，並佐以案例示範。尤以特別的是，美國 FDA 之多重評估指標指引草案，從各種不同多重檢定問題的來源中，獨挑多重評估指標(multiple endpoints)建立指引，清楚定義主、次要評估指標家族及統計調整之順序性；並釐清共同主要評估指標(co-primary endpoints)及複合式評估指標(composite endpoint)之定義與區別，及其在多重檢定問題上不同的考量。可以預期在此項指引草案發布之後，將有助於業者與法規單位對於臨床試驗設計及分析解讀的溝通，避免評估偏差及錯誤的結論。

本文將摘要介紹美國 FDA「臨床試驗多重評估指標指引草案」內容，包括多重檢定問題、多重評估指標種類、以及針對多個評估指標之多重檢定問題的統計考量原則。

多重檢定問題

依據嚴謹設計之臨床試驗結果，證明研究藥品的療效，為支持新藥上市核准的關鍵條件。惟當臨床試驗設計包含多個主要評估指標或比較，且任一比較檢定成功即認定療效時，一個無效藥被錯誤認定為有效藥(假陽性)的機率，會因多次比較檢定而擴增，此即為臨床試驗之多重檢定問題(multiplicity issues)。多重檢定問題會造成錯誤的結論，使療效證明可信度降低。

[1] 財團法人醫藥品查驗中心新藥科技組

多重檢定問題的來源,除了多重評估指標外,單一評估指標採用多面向的比較分析，包括多個劑量組、多個時間點的比較、或多個次群體的分析，且任一分析比較成功，即認定試驗成功，此多重分析比較亦會增加假陽性的機率。另外，對單一評估指標採取多種分析方法，例如分別執行調整或不調整共變數的統計分析、選擇多種分析群體(ITT、PP、及試驗完成者(completers))、依據不同評估者(臨床醫師評估、或由中央評估委員會評估)的評估結果等等，都會造成多重檢定的問題。此外，臨床試驗期間針對療效執行的期中分析(interim analysis)亦為多重檢定問題的來源之一；惟因期中分析之間及與最終分析之間數據彼此的相依性，已有特定的統計分析方法(例如，群聚循序分析(group sequential analysis))處理此項多重檢定問題，故不列於一般多重檢定問題的考量中。至於其他來源的多重檢定問題，基本上亦可參考此指引草案加以類推，從事必要的統計調整。

臨床試驗包含多重評估指標時，控制試驗型一誤差(Type I error, α)的重要原則為預先定義所有欲從事統計推論之評估指標，及其對應之數據分析方法。統計分析計畫中應敘述如何檢定這些評估指標，包括檢定次序及各個檢定自試驗整體 α 所分配之型一誤差(α*)。並且，統計分析計畫的任何變動，包括新增額外的分析，均應在數據解盲前規劃完成，且應考量多重檢定問題之統計調整。

所以，在美國 FDA 之多重評估指標指引草案中明確陳述，單憑事後分析(post-hoc analyses)結果，是無法證明藥品的療效。亦即，當試驗主要療效分析結果結論為「失敗」時，事後分析特定的評估指標，也許有助於將來可能進行的假說檢定，惟無法認定為試驗之確認性的結果；因為選擇的事後分析，可能會受到「想成功」的慾望所趨，其結果可能會有所偏差。而且，事後分析也有多重檢定的問題，但無法知道到底有多少事後分析須要執行，因此也沒有可靠的統計方法來針對此多重檢定問題進行調整。同樣地，即使是一個成功的臨床試驗，未預先規劃或未納入多重檢定調整的新增評估指標，通常亦不能據以作療效的宣稱。

評估指標家族

療效評估指標(efficacy endpoints)為反應藥物療效的測量觀察變數(variable)，可以為發生的臨床事件(例如，死亡、中風、肺部惡化、靜脈血栓栓塞)、病患症狀(例如，疼痛、喘、憂鬱)、功能測量(例如，行走或活動的能力)、或是上述事件或症狀的替代性評估。在某些情況下，單一評估指標無法充分的證明藥品的療效，因此臨床試驗設計了多個評估指標以檢視驗證藥品的療效。

當臨床試驗中包含多個評估指標時,這許多評估指標通常可歸類至三個家族:主要、

次要、和探討性評估指標家族。評估指標隸屬於那一個評估指標家族，應預先定義。三個評估指標家族，有其本質上的順序性。通常臨床上最重要的評估指標，即指定為主要評估指標(例如，死亡或不可逆之疾病)，但也不盡然如此。不依臨床重要性歸類評估指標的主要理由是，臨床上較重要的評估指標，試驗期間發生的事件可能太少，不足以提供適當的統計檢定力(statistical power)來偵測療效，或是預期臨床上較不重要的評估指標主導了藥品的療效。有時，評估指標是依照證明療效的可能性來歸納入三個評估指標家族。例如，癌症用藥臨床試驗，雖然病患存活時間(survival)一般皆被認定為最重要的評估指標，惟仍常選擇疾病惡化時間(time-to-disease progression)為主要評估指標。主要理由為疾病惡化的療效較容易證明，及疾病惡化較短時間內會發生。相較於存活時間觀察到的療效，可能因疾病惡化後的其他後續治療而被稀釋，以疾病惡化時間為評估指標，通常可看到治療組別間較明顯的療效差距。

一、主要評估指標家族

主要評估指標家族包括用以證明藥品的療效及/或安全性(亦即，證明試驗成功)的評估指標，並據此支持法規審查的上市核准需求。當臨床試驗中僅包括單一之主要評估指標時，除非有多面向的比較，可不須考慮主要評估指標之多重檢定問題。

多個主要評估指標有以下三種情形：1) 須考慮多重檢定問題的多個主要評估指標(multiple primary endpoints)；2) 不須考慮多重檢定問題的共同主要評估指標(co-primary endpoints)；及 3) 不須考慮多重檢定問題之複合式評估指標(composite endpoint)。

(一)多個主要評估指標(Multiple Endpoints)

用以證明藥品療效的主要評估指標，應具有一個或多個疾病的重要特性，且應具有臨床意義。在某些情況下，單一評估指標無法充分符合此項定義。舉例來說，偏頭痛的治療處置，雖然疼痛是最突顯的疾病特性，但偏頭痛的特徵也常以畏光、畏聲、及噁心為表徵，這三樣均為臨床上重要的疾病特性。至於臨床上何者最重要？則由病患來決定。現今對於偏頭痛治療藥品療效的認定，常須證明疼痛加上上述三項病徵最惱人的一項均有改善來認定。另一個例子是治療燒傷的藥品，因不確定此藥品是加速傷口閉合或減少結疤的機率，但此二項均具有臨床重要性，因此試驗可能定義傷口閉合率及疤痕測量為二個主要評估指標。

多個主要評估指標若任一主要評估指標，在法規要求之型一誤差範圍內(雙尾 p 值 <0.05)檢定成功，即可支持療效結論。如此雖增加了試驗成功證明療效的機率，卻使型一誤差擴增，增加假陽性的機率，可能誤導研究藥品之療效結論。因此須考量採用適當

的統計調整方法，以控制整體型一誤差於法規要求的範圍內。

(二)共同主要評估指標(Co-Primary Endpoints)

某些疾病數個疾病特性均為關鍵性，以至於當治療處置未能在這些疾病特性中均證明療效時，藥品即不能被認定為有療效。或是只有一個關鍵的疾病特性，惟僅憑此單一評估指標的療效，不具臨床意義。因此，通常會選擇二個評估指標，一個評估指標為疾病特性的具體呈現，且會直接反應藥品的治療效果，惟在臨床上不易解釋，另一個評估指標則為臨床上容易解釋，但試驗藥品的影響較不具體。以上兩種情況定義的主要評估指標，一般稱為共同主要評估指標。

以治療阿茲海默症(Alzheimer's disease)的藥品為例，此類藥品通常須在兩種疾病病徵上證明療效。一個主要評估指標為認知評量(例如，阿茲海默症評估量表「認知項目」(Alzheimer's Disease Assessment Scale – Cognitive Component))的療效，而另一個主要評估指標則為臨床較易解釋的功能測量，包括臨床醫師的整體評估量表(clinician's global assessment)或日常生活活動評量(Activities of Daily Living Assessment)。

當採用二個或二個以上共同主要評估指標時，藥品療效的決定，必須所有主要評估指標在法規要求之型一誤差範圍內均檢定成功；亦即，只有當所有共同主要評估指標均達統計顯著性時(雙尾 p 值<0.05)，試驗才算成功。在此情況下，並無多重檢定問題，亦即沒有型一誤差擴增的疑慮；但是，卻有型二誤差(Type II error, β)擴增的考量，亦即整體試驗統計檢定力(Power = 1-β)會降低，故宜列入樣本數計算的考量。例如試驗樣本數的估計是依據兩個彼此獨立的共同主要評估指標，各自證明成功的檢定力為 80% (亦即，各自的型二誤差皆為 20%)，依此計算的樣本數，二個評估指標均達顯著亦即試驗成功的檢定力(Study Power)僅為 64% (0.8×0.8)。而當共同評估指標數超過二個時，試驗檢定力則為更低。

(三)複合式評估指標(Composite Endpoint)

某些疾病的臨床試驗，可能會有數個重要的臨床測量，且這些測量值預期都會受到藥品治療處置的影響。與其定義每一個臨床測量為個別的主要評估指標(可能造成多重檢定問題或試驗檢定力降低)，或從中選擇一個臨床測量為主要評估指標，也許更適合的是，將這些臨床測量合併成一個評估指標，此稱為複合式評估指標。此評估指標任何一個組成(component)的發生或達到，即視為此複合式評估指標發生或達到。

當治療處置的目標是預防或延緩發病，或是臨床上重要但不常見的事件時，常採用複合式評估指標。例如，當個別組成皆為臨床事件，且單一事件發生率預期很低時，常將數個事件(例如，心血管疾病造成的死亡、心肌梗塞及中風)合併為一複合式事件評估

指標，而其中任一事件發生即視為複合事件的發生。複合式評估指標若為發生時間測量值，常以最先發生事件的時間為其測量值。但是對於一位病患可能發生不只一個事件的疾病，可能亦須考量所有事件總發生數或發生率的分析。

另外一種具多單項組成(multi-components)的評估指標，係將每一位受試者多個單項的評估，依特定的規則合併成一個整體評分或評階。如此多個單項的合併，可納入疾病多個不同的病徵，但療效評估仍為單一的主要評估指標，並沒有多重檢定的問題。

如果各單項為順序型類別評分或連續型數值評分，合併成一個整體評分，可以是加總或各單項平均分數，假說檢定可比較各組間之平均值的差異。此類範例包括思覺失調症(schizophrenia)研究採用的正負性症狀評量(Positive and Negative Syndrome Scale, PANSS)、評估頸肌張力障礙(cervical dystonia)之多倫多西區痙攣性斜頸量表(Toronto Western Spasmodic Torticollis Rating Scale, TWSTRS)、漢氏憂鬱量表(Hamilton Rating Scale for Depression, HAM-D)、簡明精神症狀量表(Brief Psychiatric Rating Scale)及病患報告結果(Patient-Reported Outcomes, PROs)等。

多單項評估指標也可以是一個二分法(是/否)評估指標，亦即病患達到各單項特定條件與否的判定。此種型式的多單項組成指標，適用於當疾病的數種不同的病徵均很重要，但對於病患的效益又不須所有病徵均具有正向的效果時。例如，類風溼性關節炎(rheumatoid arthritis)的美國風濕病學會(American College of Rheumatology , ACR)評分量表，病患的正向反應可以定義為該量表的一或二個面向的改善，加上至少一個另外指定的疾病病徵的改善。例如定義 ACR20 為治療有反應的條件，即為二項疾病病徵「關節觸痛(tender joints)」及「關節腫脹(swollen joints)」數目有 20%的改善；以及另外五項病徵(疼痛、急性期反應(acute phase reactants)、病患或醫師的整體評估、或失能(disability))中，至少三項有 20%的改善。

複合式評估指標中，對各單項的選取應小心謹慎，因為每一個單項在複合式分析中均同等重要。惟若各單項的臨床重要性並不相同，並且藥品療效主要來自於其中最不重要的單項時，複合式評估指標就不會是適當的主要療效評估指標。另外，亦有可能某一單項組成的重要性很大，惟不具療效，而其他一個或數個重要性較低的單項組成都顯示出正面的療效，使得整體結果仍達到正面療效的統計結論，在此情況下，對於治療處置的臨床價值亦將有所疑慮。此外，為確認治療處置對病人的效益，複合式指標不僅要評估整體的療效，針對各單項的分析亦是相當重要。雖然，將療效的幾個關鍵面向合併成一個複合式主要評估指標，可避免多重檢定的問題，若各單項組成僅為描述性統計，並無療效的宣稱，則不影響主要評估指標顯著性結論；惟若其中的單項欲從事推論性統計分析，並據此宣稱療效，其假說檢定就應納入試驗整體型一誤差的考量，以避免多重檢

定問題。舉例來說，嚴重疾病的某些單項評估指標相當具有臨床重要性，故臨床試驗通常須收集及分析其數據，尤其是死亡或其他嚴重事件(包括中風、骨折及肺部惡化等)，這些評估指標可預期事件發生數相對較少，惟可合併為複合式評估指標；若其單項組成亦設定為主要評估指標，則須考量多重檢定的問題。

二、次要評估指標家族

次要評估指標通常為主要評估指標成功顯示療效後，作為療效進一步的支持，或為已經證明的療效提供特定作用機制之證據。例如，治療骨質疏鬆藥品以骨折為主要評估指標，骨密度的改善為次要評估指標。所以次要評估指標可以是藥效學方面的指標，雖不為可接受的主要評估指標，但卻與臨床療效緊密相關。次要評估指標亦可為與主要評估指標相關之臨床效果，且為其呈現療效之後續臨床結果。例如，心血管疾病用藥之主要評估指標為心臟衰竭相關之住院，次要評估指標可以是存活相關的療效評估。或者，次要評估指標可以是與主要評估指標明顯不同的臨床效益。例如，治療多發性硬化症(multiple sclerosis)的臨床試驗，主要評估指標為復發率，次要評估指標可以是失能性(disability)。而次要評估指標常見的範例，亦包括與主要評估指標具相同的測量變數，惟測量時間點不同；及主要評估指標為病患症狀有否改善，次要評估指標則為症狀分數改善的百分比。

臨床試驗中僅包括單一主要評估指標的比較時，無須考慮多重檢定問題；惟次要評估指標若欲從事療效的宣稱，則須考量多重檢定問題。惟有主要評估指標家族證明了療效，才能解釋次要評估指標的正面療效結果。針對多重檢定的問題，則必須考量主、次要評估指標家族之內及之間型一誤差的控制。另外，若次要評估指標具顯著療效也很重要，則應在試驗設計階段，將其納入樣本數估算的考量。

三、探索性評估指標家族

除了主、次要評估指標之外，其他評估指標則列為探索性評估指標。可能為重要的臨床事件，預期發生數太少，不足以顯示治療的效果；或因某些理由而被認為是不太可能顯示出療效的評估指標，但仍可納入評估，亦有可能產生新的假說，提供未來進一步的驗證。

結語

美國 FDA 於 2017 年一月發表「臨床試驗多重評估指標指引草案(Draft Guidance for Industry: Multiple Endpoints in Clinical Trials)」，自許多多重檢定問題可能來源中，獨挑多重評估指標建立指引，其他來源的多重檢定問題，其統計調整的方法，亦可參考此指引加以類推。在多重評估指標之探討上，指引清楚定義了主、次要評估指標家族及其重要性的順序。另外，亦清楚定義共同主要評估指標(co-primary endpoints)及複合式評估指標(composite endpoint)，及其在多重檢定問題上之考量。預期此指引草案發布之後，對於未來臨床試驗設計將有重大的影響。亦有專家學者認為，日後可能不會在臨床試驗設計中看到「關鍵次要評估指標(key secondary endpoint)」此名詞，臨床試驗之評估指標，將根據指引的定義分別列屬三個評估指標家族，並依循指引所建議之統計檢定策略，控制整體型一誤差於法規要求的範圍內(雙尾 p 值<0.05)。