臨床試驗之多重檢定

蔡貴鳳[1]　宮玫芬[1]　林敏雄[2]

藥物的療效須藉由臨床試驗之假說檢定來驗證。多重檢定常見於臨床試驗設計，包括多個主要與次要評估指標、多個試驗組別、多個主要評估時間點、多個分析群體以及子群體分析等。爲避免因多重檢定導致試驗將無效藥誤判爲有效的機率擴增，使得試驗對療效結果的宣稱受到質疑，必須選用適當的統計方法控制整體型一誤差。本文目的在討論何種設計下會有多重檢定問題的產生，以及介紹幾種可適當處理多重檢定問題的統計方法，並佐以範例說明供臨床研究者參考。

（台灣家醫誌 *2014; 24: 157-163*） *DOI: 10.3966/168232812014122404001*

關鍵詞：臨床試驗、僞陽性率、多重檢定調整

前　言

藥物的療效須藉由臨床試驗之假說檢定(hypothesis testing)來驗證。臨床試驗因為耗時耗力又花錢，自然會希望儘可能的從一個臨床試驗來回答多個關於藥物療效的問題。當所要回答的問題越多，需要進行的統計假說檢定就越多，伴隨而來就是多重檢定(multiple tests)問題的產生。因此臨床醫師不論是參與臨床試驗或閱讀臨床文獻都應對多重檢定有所瞭解。

一個統計假說檢定是依據試驗樣本(sample)得到的證據，來推論藥物用於目標族群(target population)是否具有療效。依據試驗樣本來推論目標族群的治療效果可能會犯下二種錯誤。第一種錯誤是藥物其實無效，但檢定結果卻判定為有效(false positive)，此種情形稱為型一誤差(type I error)，多以符號"α"表之。此種誤判亦稱為消費者風險(consumer's risk)，是法規單位所特別重視的。第二種錯誤是藥物事實上是有效，但檢定結果卻判定為無效(false negative)，此種情形稱為型二誤差(type II error)，多以符號"β"表之。此種誤判亦稱為生產者風險(producer's risk)。

針對療效確認性試驗，法規單位多要求整個試驗將無效藥誤判為有效的機率控制在雙尾0.05或以下。理想情況下，臨床試驗僅選定一個主要療效指標，從事一次假說檢定來驗證試驗藥物的有效性。若該假說檢定結果犯型一誤差機率設定為0.05，則整個試驗錯誤宣

[1]財團法人醫藥品查驗中心、[2]國泰綜合醫院家庭醫學科

受理日期：103年4月23日　　同意刊登：103年4月25日

通訊作者：林敏雄　　通訊地址：台北市仁愛路四段280號　國泰綜合醫院家庭醫學科

稱藥物為有效的機率仍可以控制在0.05或以下。

假設一個臨床試驗有三個主要療效指標，且從事三次假說檢定來驗證藥物的有效性。若藥物宣稱有效的條件為任何一個指標檢定結果達統計顯著即可，假使每次假說檢定犯型一誤差機率訂為0.05，則整個試驗錯誤宣稱藥物為有效的機率近乎為0.14 (=1-(1-0.05)^3)，比一般法規單位所能忍受的雙尾誤差機率0.05高出甚多。因此，在多重檢定情形下，如何將試驗錯誤判斷藥物為有效的機率（以下稱為整體型一誤差機率；study-wise type I error rate）控制在0.05或以下則是非常重要課題。本文將討論何種情況下會有多重檢定問題的產生，以及介紹幾種常見的多重檢定處理方法，並佐以實例說明。茲分別討論如下。

多重檢定與型一誤差

臨床試驗多重檢定問題主要來自以下幾種設計：多個主要與次要評估指標、多個試驗組別、多個主要評估時間點、不同分析族群以及子群體分析等。

針對某些適應症，試驗需要有多個主要療效指標來驗證藥物療效，此時就需要進行多重檢定。例如，阿茲海默疾病的試驗設計會選用共同主要療效指標(co-primary efficacy endpoints)，以同時評估藥品對知能(cognition)與功能(function)的治療效果來驗證藥品療效。另外，臨床試驗也會選取多個次要療效指標以獲得更多藥物資訊。若某些重要次要療效指標分析結果欲放入仿單宣稱中，則其對應假說檢定應納入型一誤差調整考量，使試驗整體型一誤差仍維持於0.05或以下。同樣的，當試驗組別具兩組以上，或主要評估時間點不只一個時，都會面臨多重檢定的問題。

為評估試驗療效結果的穩健性(robustness)，經常會在不同分析群體或子群體(subgroup)，執行敏感性分析(sensitivity analysis)。若事先定義主要分析群體，而其他分析群體或子群體分析只為提供更多支持性證據，並不作為仿單宣稱依據，則不需考量多重檢定問題。但若欲宣稱某子群體之療效，則必須事先定義相對應之假說檢定，搭配適當之統計分析方法與策略以控制整體型一誤差，始能驗證此子群體之療效。一般而言，若整體試驗群體(overall study population)分析未達療效顯著結果，很難根據子群體的分析結果進行療效的宣稱。

針對上述多重檢定設計與考量，EMEA(European Medicines Agency)所公布之臨床試驗中多重檢定問題指引(Points to consider on multiplicity issues in clinical trials)[1]有更詳盡的說明。此外，在臨床試驗完成前執行期中或多次期間分析(interim analyses)也會產生多重檢定的問題；但此情況較為複雜，已有專門之統計方法處理（例如：群集逐次分析；group sequential analyses），故不在本文討論範疇。

多重檢定處理方法

以下將介紹一些常見多重檢定處理方法。第一部分不需調整個別檢定型一誤差，但須事先定義成功條件；第二部分則是調整個別檢定型一誤差。

個別檢定型一誤差不需調整

假設一個臨床試驗共提出四個主

要療效假說，如果此臨床試驗成功的定義是每個假說檢定的 p 值都必須≤ α ，此種設計稱為all-or-none win設計。試驗成功，則所有假說對應療效指標的療效皆可宣稱，不會造成整體型一誤差擴增；但若只要任一個假說 p 值大於 α ，則無任何療效指標的療效可以宣稱。在此條件下，假設 α =0.05，則試驗宣稱成功的型一誤差機率其實最多不超過0.00000625(=$(0.05)^4$)，遠遠低於可容忍的0.05。但試驗成功的定義似乎過於嚴苛。

順序檢定法(sequential testing procedure)不需調整個別檢定型一誤差，但放寬試驗成功的定義。此方法是依據封閉檢定(closed testing)[2]的概念而來，其最大特點是事先排定檢定順序（如圖1所示），只有在前一個檢定的 p 值小於 α 的情況下，才能執行下一個檢定，藉此設計一個可以進行多重檢定的臨床試驗。

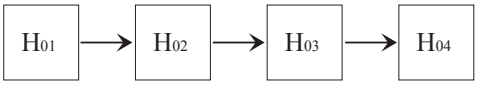

圖1　四個多重檢定之順序檢定程序

如果第一個檢定的 p 值大於 α ，則不須執行後續檢定，試驗分析結束且無法宣稱任何療效。如果第一個檢定的 p 值≤ α ，則試驗成功可宣稱此對應療效指標之療效，並繼續執行下一個檢定；如果所執行的檢定其 p 值大於 α ，則試驗分析中止，並可成功宣稱在此檢定前所有檢定對應療效指標之療效。需注意的是，檢定的順序非常重要，不同順序排列將影響試驗結果的判讀。

個別檢定型一誤差調整

以下介紹三種藉由調整個別檢定型一誤差，控制整體型一誤差的方法，包括Bonferroni校正法、Holm逐步向下分析法[3]及Hochberg逐步向上分析法[4]等。

1.Bonferroni校正法

Bonferroni校正法主要是將整個試驗所允許的型一誤差機率 α 平均分給各個檢定(un-weighted)或依據檢定的重要程度各自給予不同比例的加權(weighted)。以四個多重檢定為例，在 α =0.05以un-weighted Bonferroni 調整下，每個檢定所允許的型一誤差機率不能超過0.05/4 = 0.0125。若任何一個檢定 p 值≤0.0125，則該檢定對應療效指標之療效就可以宣稱。

依據Bonferroni調整，不須事先排定檢定順序，但每個檢定會損失可允許的型一誤差機率（例如，自0.05降至0.0125）。如採用weighted Bonferroni調整，則每個檢定不同比例的加權(weights)必需在臨床試驗計劃書中事先決定好。

2.Holm逐步向下分析法(Holm's step-down procedure)

Holm於1979年提出，結合順序檢定法及Bonferroni 調整方法的優點，且每個檢定不會損失太多型一誤差機率，且又不需事先排定加權比例或檢定順序。主要修正概念是將每個檢定的 p 值依大小排序，並根據排序的大小調整個別檢定可容許的型一誤差機率，並自最小之 p 值開始採逐步檢定。

以四個多重檢定為例，試驗可容許最大型一誤差訂為 α ，檢定策略如表1所示：

-四個檢定的 p 值排序：$p_{(1)} \le p_{(2)} \le p_{(3)} \le p_{(4)}$

-相對應的假說：$H_{0(1)}$、$H_{0(2)}$、$H_{0(3)}$、$H_{0(4)}$

Holm逐步向下分析法自最小之 $p_{(1)}$ 開始，相對應的假說 $H_{0(1)}$所允許的型一誤差機率為四個檢定平分之 $\alpha/4$。

Step 1.如果$p_{(1)}$未小於 $\alpha/4$，則不須執行後續檢定，試驗分析結束且無法宣稱任何療效；如果$p_{(1)}$小於 $\alpha/4$，則否定$H_{0(1)}$，試驗成功宣稱$H_{0(1)}$對應療效指標的療效，並繼續檢定次小的$p_{(2)}$。

Step 2. $p_{(2)}$相對應的假說 $H_{0(2)}$所允許的型一誤差機率為剩下三個檢定平分之 $\alpha/3$。如果$p_{(2)}$ 未小於 $\alpha/3$，則不須執行後續檢定，試驗僅成功否定$H_{0(1)}$；如果$p_{(2)}$小於 $\alpha/3$，則試驗可成功否定$H_{0(1)}$ 及$H_{0(2)}$，再繼續檢定的$p_{(3)}$。

Step 3. $p_{(3)}$相對應的假說 $H_{0(3)}$所允許的型一誤差機率為剩下二個檢定平分之 $\alpha/2$。如果$p_{(3)}$未小於 $\alpha/2$，則不須執行最後檢定，試驗僅成功否定前二個假說 $H_{0(1)}$ 及$H_{0(2)}$；如果$p_{(3)}$小於 $\alpha/2$，則前三個假說對應療效指標之療效均可成功宣稱，且繼續至最後一步。

Step 4.最大 p 值($p_{(4)}$)相對應的假說 $H_{0(4)}$所允許的型一誤差機率為 α 。如果$p_{(4)}$未小於 α ，則試驗僅成功否定前三個假說；如果$p_{(4)}$ 小於 α ，則試驗可成功宣稱所有療效。

3.Hochberg逐步向上分析法(Hochberg's step-up procedure)

Hochberg於1988年提出，其方法與Holm方法相似，但檢定順序相反，自最大之 p 值開始採逐步檢定。以四個多重檢定為例，檢定策略如表2所示：

Hochberg逐步向上分析法則自最大之 p 值($p_{(4)}$)開始，

Step 1.如果 $p_{(4)}$ 小於 α ，則試驗成功，可宣稱所有療效；

Step 2.如果 $p_{(4)}$ 未小於 α ，但 $p_{(3)}$ 小於 $\alpha/2$，則試驗可成功宣稱前三個假說 ($H_{0(1)}$, $H_{0(2)}$, $H_{0(3)}$) 對應療效指標之療效；

Step 3.如果 $p_{(4)}$未小於 α 且 $p_{(3)}$亦未小於 $\alpha/2$，但 $p_{(2)}$小於 $\alpha/3$，則試驗可

表1 Holm逐步向下分析法

檢定策略：

$H_{0(1)}$ is rejected	if $p_{(1)} \le \alpha/4$
$H_{0(2)}$ is rejected	if $p_{(1)} \le \alpha/4$, $p_{(2)} \le \alpha/3$
$H_{0(3)}$ is rejected	if $p_{(1)} \le \alpha/4$, $p_{(2)} \le \alpha/3$, $p_{(3)} \le \alpha/2$
$H_{0(4)}$ is rejected	if $p_{(1)} \le \alpha/4$, $p_{(2)} \le \alpha/3$, $p_{(3)} \le \alpha/2$, $p_{(4)} \le \alpha$

表2 Hochberg逐步向上分析法

檢定策略：

$H_{0(1)}$ is rejected	if $p_{(4)} > \alpha$, $p_{(3)} > \alpha/2$, $p_{(2)} > \alpha/3$, $p_{(1)} \le \alpha/4$
$H_{0(1)}$,$H_{0(2)}$ are rejected	if $p_{(4)} > \alpha$, $p_{(3)} > \alpha/2$, $p_{(2)} \le \alpha/3$
$H_{0(1)}$,$H_{0(2)}$,$H_{0(3)}$ are rejected	if $p_{(4)} > \alpha$, $p_{(3)} \le \alpha/2$
$H_{0(1)}$,$H_{0(2)}$,$H_{0(3)}$,$H_{0(4)}$ are rejected	if $p_{(4)} \le \alpha$

成功宣稱 $H_{0(1)}$ 及 $H_{0(2)}$ 對應療效指標之療效；

Step 4.如果 $p_{(4)}$ 未小於 α 、$p_{(3)}$ 未小於 $\alpha/2$ 且 $p_{(2)}$亦未小於 $\alpha/3$、但 $p_{(1)}$小於 $\alpha/4$，則試驗僅成功宣稱一個假說 $H_{0(1)}$對應療效指標之療效；否則，試驗分析結束且無法宣稱任何療效。

範例說明

以治療高血壓的新藥Drug X為例，執行一個多中心、隨機、雙盲、安慰劑對照試驗，評估Drug X 10 mg與Drug X 20 mg的療效。共同主要療效指標選取治療12週後，相較於基礎值之舒張壓及收縮壓的改變量。四個主要虛無假說為：

(1) H_{01}: 治療12週後，Drug X 10 mg與安慰劑在降舒張壓效果是沒有顯著差異
(2) H_{02}: 治療12週後，Drug X 20 mg與安慰劑在降舒張壓效果是沒有顯著差異
(3) H_{03}: 治療12週後，Drug X 10 mg與安慰劑在降收縮壓效果是沒有顯著差異
(4) H_{04}: 治療12週後，Drug X 20 mg與安慰劑在降收縮壓效果是沒有顯著差異

上述四個假說之檢定 p 值分別為p_{01}=0.081、p_{02}=0.005、p_{03}=0.024、p_{04}=0.020。在all-or-none win設計下，由於p_{01}>0.05試驗失敗，無法宣稱Drug X任何療效。藉由表3可清楚說明順序檢定法在不同次序排列情況下，試驗結論將有所不同。

表4同時比較Bonferroni法、Holm逐步向下分析法及Hochberg逐步向上分析

表3　順序檢定法結果

檢定次序（型一誤差 α =0.05）	試驗結論
$H_{01}\rightarrow H_{02}\rightarrow H_{03}\rightarrow H_{04}$	試驗失敗，無法宣稱任何療效
$H_{02}\rightarrow H_{04}\rightarrow H_{01}\rightarrow H_{03}$	試驗成功，宣稱Drug X 20 mg在舒張壓與收縮壓的療效
$H_{02}\rightarrow H_{01}\rightarrow H_{04}\rightarrow H_{03}$	試驗成功，宣稱Drug X 20 mg在舒張壓的療效

表4　Bonferroni、Holm、Hochberrg方法檢定結果

p 值排序（由小至大）	對應虛無假說	個別檢定之型一誤差		
		Bonferroni	Holm	Hochberg
$p_{(4)}$=0.081	H_{01}	0.0125	0.0500 ↑	0.0500 ↓
$p_{(3)}$=0.024	H_{03}	0.0125	0.0250	0.0250
$p_{(2)}$=0.020	H_{04}	0.0125	0.0167	0.0167
$p_{(1)}$=0.005	H_{02}	0.0125	0.0125	0.0125
試驗宣稱		試驗成功，宣稱Drug X 20 mg在舒張壓的療效。	試驗成功，宣稱Drug X 20 mg在舒張壓的療效。	試驗成功，宣稱Drug X 20 mg在舒張壓與收縮壓的療效；亦可宣稱Drug X 10 mg在收縮壓的療效。

法的檢定結果。

由表4，Bonferroni方法將檢定假說之可容許型一誤差皆調整為α/4=0.0125，則拒絕最小 p 值($p_{(1)}$)的假說。Holm方法由$p_{(1)}$開始檢定，$p_{(1)}$小於可容許型一誤差α/4=0.0125後，再進一步檢定$p_{(2)}$。$p_{(2)}$未小於可容許型一誤差α/3=0.0167，所以分析停止，不須執行後續檢定，試驗僅成功否定$p_{(1)}$對應的假說。而，Hochberg方法由最大的 p 值($p_{(4)}$)開始檢定，$p_{(4)}$未小於0.05，因此繼續檢定$p_{(3)}$。由於$p_{(3)}$小於α/2=0.025，所以試驗可成功宣稱$p_{(1)}$, $p_{(2)}$, $p_{(3)}$對應療效指標之療效。

結　論

臨床試驗常面臨多重檢定設計，為避免因多重檢定導致試驗結論犯型一誤差機率大於可容許機率，使得試驗對療效結果宣稱的可信度受到質疑，必須選用適當統計分析方法與檢定策略來控制整體型一誤差。本文簡介幾種簡單多重檢定策略，如採用逐次檢定法則每個檢定的型一誤差機率不需調整，但檢定的順序必須事先排定，不同的順序排列將影響試驗的結論；Bonferroni校正法將整個試驗所允許的型一誤差機率平均或加權分給各個檢定，不須事先排定檢定次序，但每個檢定會損失可允許的型一誤差機率，較為保守；Holm與Hochberg逐步分析法綜和了逐次檢定法及Bonferroni調整的優點，可依據檢定結果再排序、調整。這幾種方法均可確保試驗整體型一誤差不超過0.05。不論採取何種方法與檢定策略，試驗計劃書或統計分析計劃書必須事先定義並詳細說明，不應任意更改。

參考文獻

1.Points to consider on multiplicity issues in clinical trials. EMEA CPMP/EWP/908/99, 19 September 2002.

2.Marcus R, Peritz E, Gabriel KR: On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976; 63: 655-60.

3.Holm S: A simple sequentially rejective multiple test procedure. Scand J Statist 1979; 6: 65-70.

4.Hochberg Y: A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75: 800-2.

Multiplicity in Clinical Trials

Guei-Feng Tsai[1], Meifen Kung[1] and Min-Shung Lin[2]

A clinical trial is often designed to test multiple hypotheses on the efficacy of a new drug. Multiplicity in clinical trials may be induced by multiple endpoints, multiple treatment arm comparisons, multiple time points, multiple analysis populations, and subgroup analyses. It is well recognized that multiplicity can have a substantial influence on the rate of false positive findings if no appropriate statistical method is used to safeguard against the inflation of false positive findings from multiple tests. The paper accordingly focuses on examining situations with the potential of triggering multiplicity problems and introducing some common multiple testing methods for controlling the false positive rate. An example comparing different statistical methods is provided for the reference of clinical researchers.
(Taiwan J Fam Med 2014; 24: 157-163) *DOI: 10.3966/168232812014122404001*

Key Words: clinical trials, false positive rate, multiplicity adjustment

[1]Center for Drug Evaluation; [2]Department of Family Medication, Cathay General Hospital, Taipei, Taiwan.
Received: April 23, 2014; Accepted April 25, 2014.