當代醫藥法規
RegMed News

不劣性試驗統計審查重點

吳雅琪*

前言

臨床試驗是探索與確認研發中新藥療效與安全性的重要研究方法。根據其研究目的，可分為兩種類型的試驗，一種為較優性試驗 (superiority trials)，一種為不劣性試驗 (non-inferiority trials)。較優性試驗是以新藥優於對照組（安慰劑或活性對照藥）的方式來證實新藥的療效。不劣性試驗旨在評估新藥與已上市藥物間之相對療效，以新舊藥物間療效差異不小於一個事先選定的量（即不劣性臨界值）來證明其療效。

一般來說，較優性試驗的試驗的設計、執行以及結果分析都相對簡單，但是在有些情況下，不劣性設計是需要的：例如基於倫理的考量，對於治療危及生命之疾病，若市面上已經有確認療效且安全的藥物，安慰劑組的設計則是不能被接受的，例如癌症、心血管疾病或細菌感染疾病等的用藥。或是已上市藥物的療效很好，新藥的療效要超過標準治療藥物的可能性較小，但是新藥又具備其他的特點 (例如更安全、更方便使用或者尊醫囑性更好)，則可以考慮採用不劣性設計。舉例來說，warfarin 之類的藥物用於預防心房顫動患者的中風與血栓栓塞性疾病效果顯著，但是這類藥物需要定期監測凝血功能，以調整使用劑量，使用上並不是很方便。因此，不少藥廠努力研發新型口服抗凝血劑藥物，希望取代 warfarin；因此臨床試驗的目標是設計新藥在預防中風與栓塞性疾病的藥效與 warfarin 相當，但是新藥具有固定劑量、不需定期驗血監測之優點。近年來發表在新英格蘭醫學雜誌(NEJM)，有關 Pradaxa (Dabigatran) (2009)以及 Xarelto (Ribaroxaban) (2011)的臨床試驗，便是具體的例子。這一類型的臨床試驗，我們稱為不劣性試驗。

不劣性試驗的一個重要假設是新藥的療效不次(低)於活性對照藥，這個假設的基礎最重要是對照藥在該臨床試驗中要顯示一定的療效，然而這個試驗設計多不包括安慰劑，因此在藥物臨床試驗中應注意以下幾個關鍵問題。

*財團法人醫藥品查驗中心臨床組統計小組

一、 對照組的選擇

由於不劣性臨床試驗的目的是要確認新藥的療效不劣於以上市之活性對照藥，因此對照藥物的選擇就非常重要。宜針對所擬宣稱之適應症，選取目前廣泛使用的標準治療藥物。除此之外，對照藥的療效是從至少兩個試驗設計良好的臨床試驗所證實，其療效 (treatment effect) 之評估亦應具有設計良好、以安慰劑為對照的試驗中獲得，並且有可靠的資料顯示其結果。

二、 臨界值的選擇

選擇活性對照藥的同時，也要考慮臨界值的大小。臨界值是臨床試驗在統計學上試圖拒絕的新藥與活性對照藥比較的劣效程度。如果新藥與活性對照藥療效間的差異之信賴限區間(常用的是雙尾 95%信賴區間)排除此劣效程度，就可以宣稱新藥不劣於活性對照藥，也就可以確認新藥的療效；如果信賴限區間包含臨界值大小的差值，則不能確認新藥的療效。

臨界值的選取應該要保證新藥的效果顯著優於安慰劑。臨界值的選取除了需要根據統計學推論也需要臨床的判斷，為反映賴以選定的證據之不確定性，宜適當且保守。臨界值的選取多以統合分析方式(meta analysis)推算出對照藥與安慰劑的差值(M_1)，再考慮現今對照組與安慰劑療效差異可能會有減少的情形 (詳見(三))，經臨床判斷後決定臨界值為 M_2 ($M_2 < M_1$)。由於 M_1 的決定是建立在對照藥與安慰劑比較之較優性試驗的基礎上，因此對照組的選擇以及歷史文獻資料的全面收集與分析是必要的。若不能從歷史文獻中得到計算臨界值的資訊，臨床試驗採用不劣性設計宜謹慎。建議不宜在沒有充分資訊下任意選取臨界值，或是以非良好設計(如無對照組設計或是開放性設計等)之試驗結果來決定不劣性試驗之臨界值。

另外，臨界值的選取必須在試驗設計階段確定，並在計畫書或統計分析計畫書中詳細說明其選取之依據，試驗執行前宜與法規單位討論以確定臨界值的選取是否合理可接受。一旦試驗開始執行後，不得再修改臨界值，以免有型一誤差擴增之疑慮。由於所欲偵測之新藥、對照藥療效差異(M_2)是比一般對照藥與安慰劑組療效(M_1)的差異小，一般來說，不劣性試驗所需的樣本數是比較優性試驗更多。

三、 試驗的檢測靈敏度和一致性

不劣性試驗必需具備檢測靈敏度 (assay sensitivity)。由於不劣性試驗旨在評估新藥與已上市對照藥物間之相對療效，且多數不劣性試驗設計是沒有安慰劑組的設計。因此，對照藥組在此不劣性試驗試驗當中必須顯現出療效，因為只有這樣，不會將無效的新藥推論為不劣於有效的對照藥(型一誤差)。當一個試驗有能力區分無效藥物和有效藥物，這樣的試驗我們稱為具有檢測靈敏度。

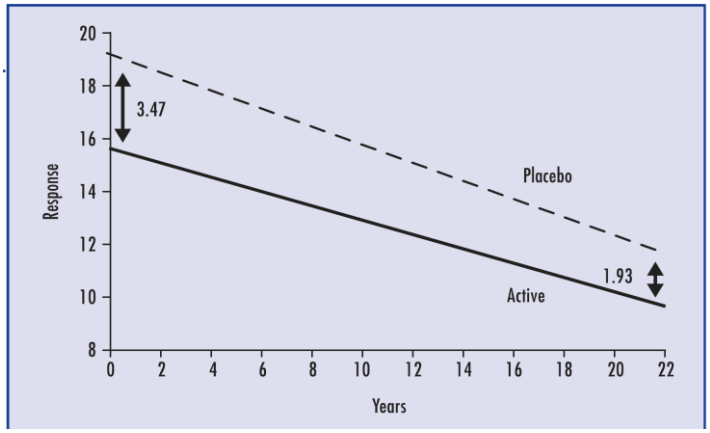
若是不劣性試驗有包含安慰劑組之設計(即新藥、對照藥物與安慰劑三組)，則可經由對照藥物組與安慰劑組之比較判定對照藥物是否具有療效(或是判定試驗是否具有檢測靈敏度)；但是針對兩組的不劣性試驗設計（即只包括新藥與活性對照藥），由於試驗中沒有衡量/保證檢測靈敏度的內部標準(即沒有安慰劑組的設計)，因此試驗是否具有檢測靈敏度很難判定，只有根據以往對照藥歷史資料來判斷，以及從試驗是否被良好執行的證據來加以檢視。

在不劣性試驗中，期望得到的結果是新藥與活性對照藥之間的差異不超過臨界值，而許多臨床試驗的執行上的偏差會縮小新藥與活性對照藥之間的差異，導致增加了將無效藥物判斷成不劣性藥物的可能性。基於此，雖然不劣性試驗是否有檢測靈敏度通常無法驗證，但必須控制不具有這種靈敏度的已知原因。如不劣性的試驗設計和執行必需與歷史的較優性試驗一致，例如主要病人族群(targeted patient population)之納入排除條件、主要評估指標、評估時間點及對照藥用法用量之選取，應與歷史的較優性試驗選取完全相同。值得注意的是，即使兩個試驗主要病人族群之納入排除條件完成相同，但是兩個試驗之人口統計學變數 (如男女比例、年齡分佈、種族等) 或是基值疾病特徵分佈情況是否相似，必須等不劣性試驗完成後方能確定。若兩者病人族群在一些重要且會影響療效的變數很不相同，會導致此不劣性試驗的結果是無法解讀的。

另外，因為不劣性試驗的目的，是希望能在嚴謹的檢測之後，宣稱新藥藥效不比對照藥差，任何可能使得新舊藥差異被低估的風險，都需要特別注意。受試者對計畫書的遵從度(compliance)不佳，即是常見的情形，需要小心維護及評估。遵從度不佳的情形，如：提早退出試驗、中斷服藥、不定時服藥、自行減少服藥劑量或頻率等，可能導致對照藥物組真正的藥效因為未能確實服藥而被低估，使得試驗結果易於達到新藥不劣於對照藥之假象。因此，不劣性試驗執行的品質是需要高度要求的。

由前述所知，不劣性試驗臨界值選取是來自歷史資料推算出來的，因此在試驗中若沒有安慰劑組的設計，則必須要求此試驗之對照藥物的療效效果應與歷史

試驗結果一致，這是所謂一致性假設(consistency assumption)。但是這個假設有時也是很難確認的。例如於 paroxetine 藥物治療憂鬱症之長期研究顯示，於 1981 至 2004 年間所執行的安慰劑對照試驗中，主要評估指標(change in HAMD score)在用藥組(paroxetine)與安慰劑組的反應均會隨時代進展而有改善，惟安慰劑組比試驗藥物組改善更多(詳見圖一)，這種現象叫做 placebo creep。可能的原因也許是安慰劑組所接受的背景治療方式、併用藥物，或是臨床上的處置(medical practice)會隨時代演進而有所改進。這種現象若出現在現今仍有安慰劑組設計的試驗，則不用擔心；但是若是兩組不劣性設計的試驗 (只有新藥與對照藥組)，一致性的條件是否成立，將很難驗證，也就是說，即使不劣性條件達成了，新藥是否能優於現今的安慰劑組仍是有疑慮的。



圖一：placebo creep 現象

由上所述，可知道兩組不劣性試驗 (只有新藥與對照藥組)的缺點就在很難判定試驗是否具有檢測靈敏度或是一致性假設在現今試驗是否適用。若對試驗之檢測靈敏度和一致性的假設有疑慮或難確認，試驗者可考慮以下其他就優性試驗設計：例如採用 add-on 的設計，即新藥+背景治療 vs.安慰劑+背景治療之較優性試驗，或針對使用目前市面上藥物效果都不佳的一群受試者，採用新藥 vs.安慰劑之較優性試驗，或是採用無效及早退出(early-escape)設計或是隨機退出(randomized withdrawal)設計，藉以排除驗證試驗檢測靈敏度和一致性的難題。

另，針對適應症是症狀改善或是所選取之主要療效指標客觀性不夠，如治療憂鬱症、焦慮、失眠、有徵狀的心衰竭、腸躁症徵狀、癡呆、心絞痛等疾病狀態；此類試驗變異性大，活性對照藥的療效不一定總是能夠預期或是重複。因此對於這些適應症的臨床試驗，如果想執行不劣性試驗，建議應採用包括安慰劑的三組的試驗設計（新藥、活性對照藥和安慰劑）。

四、 資料分析群體的選擇

在較優性試驗中，主要分析群體多以「意圖治療」(Intention-to-treat, ITT)群體為主，根據 ICH E9 準則，這是比較保守嚴謹的選擇；但這個選擇在不劣性試驗中的效果並不保守。採用 ITT 群體時，不論試驗對象是否遵從服藥原則、是否中途退出試驗，仍然列入分析，不會排除於分析群體之外。這可能會使得兩組間藥效差異被低估，比較不容易做出新藥與對照藥之間有顯著差異的推論。在較優性試驗中，這樣的保守作法將可減少「將沒有療效的藥物宣稱為有顯著療效」的機率，被視為是嚴謹的分析群體；然而，在不劣性試驗中，若 ITT 分析群體的結果導致新藥與對照藥療效間可能存在的差距被低估，此時反而犯下「將不夠有效的新藥宣稱為與對照藥一樣好」的錯誤。

為了避免上述問題，通常會同時考量「依據計畫書」(per protocol analysis, PP)分析群體，即排除未能遵從試驗計劃書、中途退出者；僅納入符合試驗計劃書執行方式的受試者進行資料分析。但是，這種作法的缺點是破壞了隨機分派的平衡狀態，可能導致選擇性偏差（selection bias）。因此，在較優性試驗中，常以 ITT 為主要的分析群體，但在不劣性試驗中，EMA 準則多建議應同時考慮 PP 與 ITT 分析群體的結果，當兩種分析群體的結果相同時，其研究結果才比較可信。

結語

不劣性試驗旨在評估新藥與已上市藥物間之相對療效，統計部份的審查重點在於臨界值選取的依據是否合理、試驗設計是否與歷史的較優性試驗相同，試驗的執行是否嚴謹、受試者的遵從度是否良好、試驗是否具有檢測靈敏度，ITT、PP 分析群體的療效結果是否一致，以整體評估這個不劣性試驗是否是成功的試驗。

參考資料

1. US FDA "Draft guidance for industry: non-inferiority clinical trials"
2. EMA "Guideline on the choice of the non-inferiority margin"
3. EMA "Points to consider on switching between superiority and non-inferiority"

4. ICH E9 Guideline (Statistical Principles for Clinical Trials) "http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html"
5. Connolly SJ, Ezekowitx MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. New England Journal of Medicine. 2009; 361:1139-51.
6. Manesh R, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. New England Journal of Medicine. 2011; 365:883-891.
7. Julious SA, Wang SJ. How biased are indirect comparisons, particularly when comparisons are made over time in controlled trials? Drug Information Journal, 2008 Vol 42: 625-633.
8. 左曉春. 臨床試驗中採用非劣效設計應該關注的問題. 中國新藥雜誌. 2007, 16(7).
9. 呂瑾立. 淺談不劣性試驗. 台灣腦中風學會會訊. 第十六卷.第三期.
10. 呂瑾立. 再談不劣性試驗. 台灣腦中風學會會訊. 第十六卷.第四期.