

ICH E9: 臨床試驗統計指導原則
Statistical Principles for Clinical Trials

衛生福利部食品藥物管理署
中華民國 115 年 2 月

前言

為提供國內臨床試驗申請者與研究者臨床試驗統計原則之技術指導，因此，參考國際醫藥法規協會(International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, ICH)於 1998 年所頒定之 E9 (Statistical Principles for Clinical Trials)指引，制定「臨床試驗統計原則與考量」，闡述臨床試驗設計以及數據分析之統計學考量，著重於如何減少試驗偏差(bias)並提升精確度(precision)，滿足當前我國臨床試驗現實需求。

目錄

1. 緒論.....	1
1.1 背景與目的.....	1
1.2 範圍與指南.....	2
2. 臨床試驗整體考量	3
2.1 試驗內容.....	3
2.1.1 發展計畫	3
2.1.2 確認性試驗	4
2.1.3 探索性試驗	4
2.2 試驗範圍.....	4
2.2.1 受試者族群	5
2.2.2 主要指標與次要指標	5
2.2.3 複合指標	6
2.2.4 整體評估指標	6
2.2.5 多個主要指標	7
2.2.6 替代指標	7
2.2.7 類別型指標	8
2.3 避免偏差之設計技巧.....	8
2.3.1 盲性	8
2.3.2 隨機分派	10
3. 試驗設計之考量	11
3.1 設計類型.....	11
3.1.1 平行設計	11
3.1.2 交叉設計	11
3.1.3 因子設計 (Factorial Designs)	12
3.2 多中心試驗.....	13
3.3 比較的類型.....	15
3.3.1 較優性試驗	15
3.3.2 相等性或不劣性試驗	15
3.3.3 劑量反應關係試驗	17
3.4 群組序貫設計 (Group Sequential Designs).....	17
3.5 樣本數.....	18
3.6 數據收集與處理.....	19

4. 試驗執行之考量	19
4.1 試驗監測與期中分析	19
4.2 變更納入與排除條件	20
4.3 受試者收納速度	20
4.4 調整樣本數	20
4.5 期中分析與試驗提早終止	21
4.6 獨立數據監測委員會(IDMC)的角色	22
5. 資料分析的考量	23
5.1 預先設定分析方法	23
5.2 分析群體	23
5.2.1 全分析群體	24
5.2.2 依計畫書群體	25
5.2.3 不同分析群體的角色	26
5.3 缺失值與離群值	26
5.4 數據轉換	27
5.5 參數估計、信賴區間及假說檢定	27
5.6 顯著水準與信賴水準的調整	28
5.7 次群體、交互作用與共變數	28
5.8 數據的完整性與電腦軟體的效度	29
6. 安全性與耐受性評估	29
6.1 評估的範圍	29
6.2 指標的選擇與數據的收集	29
6.3 分析群體與數據的呈現	30
6.4 統計評估	31
6.5 綜合總結	31
7. 報告撰寫	32
7.1 評估與報告	32
7.2 臨床試驗資料的總結	33
7.2.1 療效資料	34
7.2.2 安全性資料	34
詞彙表	35

ICH E9: 臨床試驗統計指導原則

1. 緒論

1.1. 背景與目的

醫藥產品的有效性和安全性須由臨床試驗來驗證，所採用的臨床試驗須遵循「藥品優良臨床試驗準則」來執行。該規範認定統計學在臨床試驗設計與分析扮演重要的角色。隨著統計學研究在臨床試驗領域不斷地發展，加上臨床研究在新藥核准過程與健康照護上的關鍵角色，制定一份簡明的臨床試驗統計相關文件是必要的。本指導原則旨在協調於歐洲、日本和美國申請藥品上市所附臨床試驗中統計方法學的原則。

首先，本指導原則採用了歐盟專利醫藥品委員會(CPMP)於1994年12月所發布之「醫藥品上市申請之臨床試驗統計方法」指引的意見，並參照日本衛生福利部於1992年3月所發布之「臨床試驗中之統計分析指引」，以及美國食品藥物管理局於1988年7月發布之「新藥查驗登記之臨床與統計文件的格式與內容基準」。其他國際醫藥法規協和會(ICH)的規範中亦包含一些統計原則與方法學的相關議題，特別是下面列出的指導原則：

- E1A：參與臨床安全性評估之總體範圍
- E2A：臨床安全性資料管理：快速通報的定義與標準
- E2B：臨床安全性資料管理：個案安全性報告通報要點
- E2C：臨床安全性資料管理：上市藥品定期安全性報告(PSUR)
- E3：藥品臨床試驗報告之格式及內容基準
- E4：支持藥品查驗登記之藥物劑量反應資訊
- E5：接受國外臨床資料之族群因素考量
- E6：藥品優良臨床試驗規範
- E7：特殊族群研究：老年醫學
- E8：臨床試驗之一般考量
- E10：臨床試驗中對照組的選擇
- M1：制訂規章用的醫學術語標準化
- M3：藥物臨床試驗前所需之非臨床安全性研究

本指導原則旨在提供廠商於藥物整體研發階段之臨床試驗設計、執行、分析和評估的方向，亦可協助負責準備申請文件或是評估主要來自晚期臨床試驗有效性和安全性證據之科學專家。

1.2.範圍與指南

本指引聚焦於統計原則，而不探討特定統計流程或是方法的使用。透過特定流程來確保統計原則有被正確實行是試驗廠商的責任。所有臨床試驗數據的整合有被討論，但並非本指引的主要焦點。與數據管理或是臨床試驗監督活動相關的選定原則與流程會在其他 ICH 指引說明而在此不探討。

本指導原則對於各學科背景之專業人士都會有幫助的。然而，正如「藥品優良臨床試驗規範」所述，所有臨床試驗相關的統計工作，皆應由具有適當資格及足夠經驗的統計人員負責。試驗統計人員(詳見詞彙表)在與其他臨床試驗專業人員合作時，其角色與責任是確保支持藥物研發的臨床試驗中適當地應用統計學原則。因此，試驗統計人員須有足夠的專業訓練與教育，並有足夠的經驗貫徹本指導原則。

所有用於上市申請而執行的臨床試驗，其設計、執行與採用的統計分析方法等重要細節，應於試驗開始前於計畫書中清楚說明。計畫書中的步驟被遵循的程度以及主要分析被事先規畫的程度，都將決定試驗最終結果和結論的可信度。試驗計畫書及後續變更，必須獲得包括試驗統計專業人員在內的負責人員同意。試驗統計專業人員必須確保計畫書及後續任何的變更中，都能清楚準確地涵蓋所有相關的統計問題，並使用適當的術語描述。

本指導原則所闡述的原則主要是針對晚期臨床試驗，其中很多是療效確認性試驗。除了療效之外，確認性試驗也可以把安全性指標(如：不良事件、臨床檢驗數據、心電圖資料等)、藥效學、藥動學指標(例如在生物相等性的確認性試驗)作為主要評估指標。此外，有些療效確認性試驗結果可能來自於不同試驗之間整合的數據，本指導原則中有些原則可適用於此種情況。最後，儘管藥物研發的早期階段本質上主要是探索性的臨床試驗，本規範的統計學原則仍可應用到這類試驗。因此，本指導原則的內容應盡可能應用到臨床試驗的各個階段。

本指導原則中所闡述的原則多是以降低偏差(詳見詞彙表)至最低、提升精準度(precision)至最高為目標。「偏差」一詞在本指導原則是指與臨床試驗設計、執行、分析與結果解讀相關的任何因素，使得治療效果(詳見詞彙表)估計值偏離其真實值的系統性傾向。因偏差的存在將嚴重危及臨床試驗得到正確結論的能力，應盡可能地找出導致偏差所有的潛在來源，才能減少這些偏差的產生。

有些偏差源自於試驗設計，例如將病情較輕的受試者系統性地分配到某一組。其他偏差源

自於臨床試驗的執行與分析，例如：違反計畫書標準以及基於對受試者結果的瞭解將其排除於統計分析之外，即為偏差的可能來源，這可能會影響到治療效果的準確估計。

偏差常在不知不覺中發生，且難以直接測量，因此，評估試驗結果和主要結論的穩健性是很重要的。穩健性的概念是指整體結論對於數據的各種限制、假設和數據分析方法的敏感性。穩健性意味著當基於不同的假設或統計方法進行分析時，試驗治療效果和主要結論皆沒有受到實質性的影響。在解讀治療效果與治療組間比較的不確定性之統計測量時，應探討偏差對 p 值、信賴區間或統計推論的潛在影響。

由於臨床試驗設計和分析的主要方法是基於頻率學派統計方法(詳見詞彙表)，因此，本指導原則在討論假說檢定和信賴區間時，主要是使用頻率學派方法。這並不意味著其他方法不適當，若使用的理由充分且所得結論是足夠穩健的，則貝氏方法(詳見詞彙表)及其他方法的使用亦可納入考量。

2. 臨床試驗整體考量

2.1. 試驗內容

2.1.1. 發展計畫

新藥臨床試驗的主要目標是尋找是否存在可以證明該藥物既安全又有效的劑量範圍和用法用量，在此範圍內之風險利益比是可接受的，同時還需要確定可能受益於該藥物的特定族群及其使用的具體適應症。

為滿足此一目標，通常需要一系列循序漸進的臨床試驗，每一個試驗有它自己特定的目標(見 ICH E8)。這應該在一個或一系列臨床計畫中闡明，這些計畫應具有適當的決策點和隨著知識累積而進行變更的彈性。一個新藥上市申請應清楚地描述各試驗的主要內容，以及每個試驗的貢獻。解讀和評估整個藥物發展計畫的證據需整合各個試驗所提供的證據(見第 7.2 章節)。透過確保試驗的多項特徵都採用共同標準，如醫學術語、主要測量的定義與時程、處理違反試驗計畫書的方法等，將有助於實現這一目標。當醫學問題涉及一個以上的試驗時，統計學上的總結、概述或統合分析(見詞彙表)可能會提供資訊。這應盡可能在計畫中預先考量，以便預先清楚地確定相關的試驗，並規定這些試驗在設計上所必要的共同特徵。若有其他主要統計議題會影響共同計畫中的試驗，亦須在該計畫中闡述。

2.1.2. 確認性試驗

確認性試驗(confirmatory trial)是一種事先提出假說並進行檢驗且具有適當對照組的試驗。通常，需要通過確認性試驗提供療效或安全性的確切證據。在這類試驗中，依據試驗的主要目標事先定義主要假說，並在試驗完成後進行檢定。在確認性試驗中，準確估計療效大小和將這些療效結果和臨床意義聯繫起來是同等重要的。

確認性試驗旨在針對所欲提出的宣稱提供確切證據，因此，確實依照試驗計畫書及標準操作步驟進行是格外重要的，對於一些不可避免的改變須提出解釋及記錄，並檢驗其所產生的影響。每個此類試驗設計的合理性，以及其他重要統計部分，如預計分析(planned analyses)的主要特點等，都應在計畫書中陳述。每個試驗應僅探討有限數量的問題。

要支持任何宣稱所需要的確切證據，須由確認性試驗證實研發中的藥物具有臨床效益。因此，確認性試驗應足以清楚明確地回答與療效或安全性宣稱相關的每個關鍵臨床問題。此外，能夠理解和解釋將試驗結果推論到目標病人群體的基礎是很重要的；這也會影響到所需試驗中心或所需的試驗數目和類型(例如專科或一般醫師)。確認性試驗的結果應當穩健。在某些情況下，來自單一確認性試驗所提供的證據可能就足夠了。

2.1.3. 探索性試驗

確認性試驗的理論基礎及設計幾乎總是基於一系列早期探索性臨床試驗。跟所有臨床試驗一樣，這些探索性試驗也應有清楚和明確的目標。但與確認性試驗相比，探索性試驗的目標並不限於對預先設定的假說進行簡單的檢定。此外，探索性試驗有時需要採用更為彈性的方法進行設計，以因應逐漸積累的結果對試驗設計進行修改。它們的分析可能僅限於數據探索；可能也會執行假說檢定，但此假說的選擇可能視數據而定(data dependent)。儘管這類試驗可能對整體療效證據有貢獻，但不能作為證明有效性的正式依據。

任何單一臨床試驗可能同時具有確認和探索兩個面向，例如，在大多數確認性試驗中，也會對資料進行探索性分析，作為解讀或支持試驗結果的基礎，並為後續研究提出進一步的假設。試驗計畫書應明確區分試驗的哪些方面將用於確認性證據，哪些方面將為探索性分析提供數據。

2.2. 試驗範圍

2.2.1. 受試者族群

在藥物開發的早期階段，臨床試驗中受試者的選擇可能會受到希望能最大化觀察到特定臨床效果的影響，因此，受試者可能僅是最終適應症核准之整體病人族群中一個非常局限的次群體。但在確認性試驗階段，試驗受試者應更能反映目標族群。因此在這些試驗中，在能保持足夠的同質性以便精確地估計治療效果的同時，在目標族群中盡可能地放寬納入和排除標準通常是有幫助的。由於地理位置、研究時間、特定研究者和醫療院所的醫療處置等因素影響，沒有一個單一臨床試驗可以期望完全代表將來的使用者。儘管如此，我們仍應盡可能減少這些因素的影響，並在闡述試驗結果時，對上述影響加以討論。

2.2.2. 主要指標與次要指標

主要指標(又稱目標變數、主要終點)應該是能夠提供與試驗的主要目標直接相關的最具臨床相關性和具令人信服證據的指標。通常主要指標只有一個。因為大部分確認性試驗的主要目標是提供與療效相關的強有力的科學證據，所以主要指標通常是療效指標。安全性或耐受性有時也可能是主要指標，並始終是重要的考量。有關生活品質和健康經濟 (health economics)的測量也是進一步可能考慮的主要指標。主要指標的選擇應反映相關研究領域中已被接受的準則和標準。建議採用先前試驗或已發表文獻中已有使用經驗之可信(reliable)且有效(validated)的指標。在納入和排除條件所描述的病人族群中，應該有充分的證據說明主要指標能夠對一些臨床相關和重要治療效果提供有效且可信的測量。主要指標通常是用於樣本數估計的指標(見 3.5 節)。

在很多情況下，評估受試者結果的方法可能並不直接，故須仔細的定義。例如將死亡率選為主要指標而無進一步說明或嚴格的定義就是不適當的；死亡率的評估可以是比較某些固定時間點的存活比例，也可以是比較在特定時段內的整體存活曲線分布。另一個常見的例子是復發事件，療效指標可以是簡單的二分法指標(在特定時段內的任何復發)、到首次復發的時間、或復發率(單位時間內觀察到的事件數)等。在慢性疾病的療效研究中，隨時間而改變的功能性狀態對主要指標的選擇增添了其他的挑戰。這有很多可能的評估方式，例如比較觀察時段內開始和結束時所做的評估，比較由觀察時段內所有觀察值中計算得出的斜率，比較超過或低於特定閾值的受試者比例，或基於重複量測數據方法的比較。為避免因事後定義而產生的多重檢定問題，在計畫書中提供主要指標精確的定義是很重要的，因為在統計分析中將使用它。另外，所選定的主要指標其臨床意義和相關量測過程的有效性(validity)，通常需要在計畫書中加以闡述。

主要指標及其選擇的理由應在計畫書中詳細說明。解盲後重新定義主要指標通常是不可接

受的，因為由此所產生的偏差難以評估。當主要目標所定義的臨床療效需要以多種方式測量時，計畫書中應根據臨床相關性、重要性、客觀性或其他相關特性，選定一種實際可行的測量方法作為主要指標。

次要指標是與主要目標相關的支持性指標，或與次要目標相關的療效指標。在計畫書中預先定義次要指標，並說明它們的相對重要性以及在解讀試驗結果時的作用，也都是很重要的。次要指標的數目應有所限制，且應與試驗中所要回答的有限數量的問題相關。

2.2.3. 複合型指標

如果與主要目標相關的多種測量方法中，難以確定單一的主要指標時，另一種方法是按預先確定的演算法將多個指標組合成一個單一或複合的指標。事實上，主要指標有時是組合多個臨床指標而成(如在關節炎、精神疾病或其他疾病中的評分尺度)。這種方法雖涉及多重檢定的問題，但不需對型一誤差進行調整。將多個指標組合的方法應在計畫書中詳細說明，且應就複合指標尺度所對應之臨床療效大小進行說明。當複合指標被選為主要指標時，可以對複合指標中具有臨床意義的單一指標進行單獨分析。此外，當評分量表被選為主要指標時，對內容效度(見詞彙表)、評估者間和評估者內的信度(見詞彙表)及疾病嚴重程度變化的鑑別度等因素進行處理是格外重要的。

2.2.4. 整體評估指標

在某些情況下，可使用整體評估指標(見詞彙表)來評估藥物的整體安全性、療效和有用性。這種指標整合了客觀指標與研究者對受試者的狀況或狀況改變的總印象，它常常是一個順序類別評分量表。在某些治療領域如神經科和精神科，已經建立整體評估指標來評估療效。

整體評估指標通常帶有主觀成分。使用整體評估指標作為主要或次要指標時，需要在計畫書中對量表的以下方面進行詳細說明：

- (1) 該量表與試驗主要目標的相關性；
- (2) 該量表之效度和信度的基礎；
- (3) 如何根據所收集的單一受試者數據，將他歸類到量表中的特定類別；
- (4) 如何將有缺失資料的受試者歸類至整體評估指標的特定類別，或有其他處理方法。

當研究者選取的整體評估指標中包含客觀指標，則這些客觀指標應作為額外的主要指標或至少作為重要的次要指標。

有用性的整體評估整合了利益與風險兩個因素，也反映了醫師的決策過程，在做出使用藥品的決策時，醫師須權衡使用這些藥物的利弊。使用整體有效性評估指標的一個問題是，在某些情況下可能將兩種不同利益風險的藥物判為相同，例如，將一種藥物的整體有用性等同或優於另一種藥物時，可能掩蓋了其效果甚微或無效而同時不良反應較少的事實。因此，不建議採用整體評估指標作為主要指標。如果整體評估指標被選作主要指標時，則將特定的療效和安全性結果分別作為附加的主要指標的考慮是非常重要的。

2.2.5. 多個主要指標

有時，我們會需要使用多個主要指標，每一個(或其中一些)都足以支持藥物某部分的療效。使用這類指標的證據須清楚載明。例如，應當清楚說明是否對任何一個指標，或最少幾個指標，或全部指標的影響，被認為是達到試驗目標所必須的。針對已定義的主要指標應清楚地說明主要假說或感興趣的假說與參數(如平均數、百分比、分布)，並清楚地描述統計推論方法。而多重檢定問題可能造成的型一誤差影響也應解釋(見第 5.6 章節)，亦應在計畫書中訂定控制型一誤差的方法。評估型一誤差影響時，所提出之主要指標之間的相關程度可能也可以考慮。若試驗的目標是證實所有指定的主要指標的效果，則無須調整型一誤差，但是必須仔細考慮對型二誤差和樣本數的影響。

2.2.6. 替代指標

如果不能通過觀測實際臨床效果來直接評估受試者的臨床效果，可以考慮採用間接指標(替代指標，見詞彙表)。有些被普遍接受的替代指標已用於一些適應症，在這些適應症中，替代指標被認為是臨床療效上可靠的預測指標。在選用替代指標時，主要考量有兩點，第一，它可能不是感興趣的臨床結果之實際預測值，例如，它可以測量與一個特定的藥理學機轉相關的療效，無論是有反應還是沒反應，它可能無法提供藥物作用的範圍與最終療效的完整資訊。已經有很多例子指出，替代指標顯示有高度正效應的藥物，最終卻對受試者的臨床結果有害；相反地，也有一些例子指出有些治療具有臨床效益，但所選用的替代指標卻沒有可測量到的影響。第二，選用的替代指標可能不會產生可直接與藥物不良反應相權衡的定量臨床效果測量。儘管驗證替代指標的統計標準已經具備，但如何運用這些標準之經驗還相當有限。事實上，替代指標所提供的證據強度取決於：(i)替代關係在生物學上的合理性；(ii)替代指標對臨床結果預後判斷之流行病學研究證據；(iii)從臨床試驗中獲得的替代指標治療效果與臨床結果相關的證據。一種產品的臨床指標和替代指標之間的關係，並不一定能適用於治療同一種疾病但具有不同作用模式的另一種產品。

2.2.7. 類別型指標

有時可能需要將連續型或順序型指標轉化為二分法或其他分類。「成功」和「有反應」的準則就是這種二分法指標的常見例子，這些分類標準需要明確定義，例如，在連續型指標中(相對於基值)最小百分比的改善，或者按照等於或超過順序型指標中某個尺度閾值(如「好」)劃分的等級。又如，舒張壓下降到低於 90 mmHg 亦是一個常見的二分法案例。對有明顯臨床相關性的變量進行分類時，是最有用的。惟由已知的試驗結果很容易對選擇分類標準的選擇產生偏差，因此在計畫書中應事先定義和詳細說明分類標準。由於採用分類指標通常意味著會喪失部分資訊，由此導致分析的檢定力降低，應在樣本數估算時加以考量。

2.3. 避免偏差的設計技巧

在臨床試驗中，用於避免偏差產生最重要的設計技巧是盲性和隨機分派，這些應是申請上市中所包含的大多數有對照組之臨床試驗的一般特徵。大多數這樣的試驗採用雙盲設計，依據適宜的隨機分派計畫，事先將藥物進行包裝並提供給試驗中心，只標明受試者編號和服用時間，因此參與試驗的任何人都不知道分配給任何特定受試者的具體治療藥物，甚至不知道編碼字母。這種方法將在 2.3.1 章節和 2.3.2 章節的大部分內容中進行介紹，例外情況將在最後討論。

透過計畫書特定的程序可以在設計階段減少偏差，這個目標是盡可能減少試驗過程中任何可預期的違規事件發生而影響分析，包括各種情況的違反試驗計畫書、提早退出試驗和缺失值(missing values)。計畫書應考量如何減少這些狀況發生的頻率及發生在資料分析時該如何處理。

2.3.1. 盲性

盲性是為了控制臨床試驗的執行過程中以及結果解讀時產生的有意或無意的偏差(bias)。這些偏差來源於知道所採用的治療後對受試者的納入和隨機分派、受試者的後續照顧、受試者對治療的態度、評估指標的量測、對提早退出試驗者的處理、從分析中排除數據等所產生的影響。盲性的根本目的是防止辨認出受試者所接受的治療，以控制產生偏差的所有機會。

在雙盲試驗中，所有受試者及所有參與受試者的治療或臨床評估的研究人員及廠商人員都不知道受試者所接受的治療。這包括確定受試者資格者、量測評估指標或評估計畫書順從性的人均須保持盲性。這種程度的盲性要持續到整個試驗完成，只有當數據被清理到可接

受的品質時，才可對適當的人員解盲。如果需要對不參與受試者的治療或臨床評估的廠商人員解盲治療組別編碼(如生物分析學家、試驗監測人員、參與嚴重不良事件通報者等)，廠商應該制定適當的標準操作程序，以防止治療組別編碼不當地傳出。在單盲試驗中，研究者或其成員知道受試者接受何種治療，但受試者本身不知道，亦或相反。在開放性試驗中，所有人都可能知道治療組別訊息。雙盲設計是最佳的方法，雙盲設計要求在試驗中所採用的治療在用藥前或用藥期間均無法從外觀、味道等識別出來，且在整個試驗期間均適當地保持盲性。

但要做到雙盲，會遇到很多困難，比如：兩種治療型態可能完全不同，如手術和藥物治療；或兩種藥物可能是兩種不同的劑型，雖然用膠囊技術可使兩者無法分辨，但改變劑型可能會改變藥物動力學或藥效學的特性，因此需要建立不同劑型的生體相等性；又或兩種藥物每日用法可能不同。在這種情況下，採用「雙虛擬」(見詞彙表)技術是實現雙盲條件的一種方法。這種技術有時會強制實施一種不尋常的用藥方案，對受試者的參與動機和順從性產生負面影響。倫理上的困難也會干擾它的應用，例如，當涉及到假手術的操作(dummy operative procedures)時。無論如何，應當努力克服這些困難。

某些臨床試驗的雙盲性質可能由於明顯的治療誘導效應(treatment induced effects)而遭到部分破壞。在這種情況下，通過使研究者和廠商相關人員對某些檢驗結果(如所選擇的臨床實驗室測量)保持盲性，有助於盲性的維持。在獨特或特定的治療效果可能導致個別受試者被解盲的試驗中，應考慮採取類似於下文中的方法，以減少開放性試驗產生的偏差。

如果雙盲設計不可行，則應考慮採用單盲設計。在有些情況下，只有開放性試驗才可行或符合倫理。單盲和開放性試驗提供了額外的靈活性，但特別重要的是，研究者雖然知道受試者下一步將接受哪種治療，但不應影響該位受試者進入試驗的決策；這個決定應該在知道隨機分派治療之前。針對這些試驗，應考慮採用中央化的隨機(centralized randomization)方法，如用電話通知指定之隨機方法。除此之外，應該由不參與治療受試者並對治療組別保持盲性的醫療人員進行臨床評估。在單盲或開放性試驗中，應盡最大努力使各種已知的偏差來源降到最低，並且主要評估指標應盡可能客觀。應該在計畫書中說明所採用的盲性程度的原因，以及通過其他方法使偏差最小化的步驟。例如，廠商應當有適當的標準作業流程以確保在資料庫清理以供分析之前，適當限制對治療組別編碼的獲取。

只有當受試者的醫師認為知道治療組別對受試者的照護是必須的，才應考慮針對單個受試者解盲。任何有意或無意地破壞盲性，不管是什麼理由，在試驗結束時都應被記錄並給予解釋。解盲的過程及時間點亦須記錄。

在本指導原則中，數據的盲性評估(見詞彙表)是指在試驗完成(對最後一位受試者的最後一次觀察)到解盲之間的這段時間內對數據的查核。

2.3.2. 隨機分派

在臨床試驗中，隨機分派是將機率元素引入到受試者的治療分派中。在後續的試驗數據分析中，它為治療效果相關證據的定量評估提供了良好的統計基礎。它傾向於使各治療組間，已知和未知預後因子分布相似。若能與盲性設計合用，在受試者的選擇和分配時，隨機分派有助於避免因治療組別的可預期而可能出現的偏差。

臨床試驗的隨機分派表記錄了隨機分派給受試者的治療。在最簡單的情況下，它是記錄治療組別的序列表(或者在交叉試驗中的治療順序)，或者是按受試者編號對應的編碼。有些試驗的後勤工作(logistics)，如具有篩選期的試驗，可能會使問題變得更加複雜，但對受試者唯一事先規畫的治療或治療順序之分派應當明確。不同的試驗設計將需要不同的程序來生成隨機碼，而隨機分派表應當有再現性(如果需要)。

儘管無限制條件的隨機分派是可接受的方法，但以區塊(block)隨機分派受試者一般可獲得某些優勢。這有助於增加治療組間的可比性，特別是當受試者的特徵可能隨時間變化時，例如，由於招募受試者的策略的改變。它還能更好地保證各組的樣本數幾乎相等。在交叉試驗中，它提供了獲得具有更高效率和更易於解釋的平衡設計的方法。在選擇區塊長度時須注意，每個區塊長度要盡可能的短，以防止組間人數不平均，但又要足夠長，以避免對區塊內末端治療組別的可預測性。研究者及其他相關人員應對區塊長度保持盲性；使用兩種或以上的區塊長度，或對每個區塊的隨機選擇長度，可達到同樣目的。(理論上，在雙盲試驗中，可預測性是不用擔心的，但藥物的藥理學作用可能提供了猜測的機會。)

在多中心試驗(見詞彙表)中，應採中央化的隨機分派過程。建議每一個中心建立一個單獨的隨機分派表，也就是依中心分層，或為每個中心分配數個完整的區塊。更常見的是，依基礎值資料中的重要預後因子(如疾病的嚴重程度、年齡、性別等)進行分層，有時可能對促進各層內各組人數達到平衡很有幫忙；這在小型臨床試驗中有更大的潛在助益。一般很少需要多於兩個或三個的分層因子，因為很難達到平衡且很麻煩。使用機動隨機分派(dynamic allocation procedure，見後述)可能有助於同時在多個分層因子之間達到平衡，只要可以調整其餘試驗流程以適應這種類型的方法。應當在後續的分析對分層隨機分派的因素加以考量。

下一個進入試驗的受試者接受的是隨機分派表上所對應的治療組別(如果是分層隨機分派，則對應到分層內的隨機分派表)。只有當已經確認下一個受試者進入到試驗的隨機分派後，才應該為該受試者分派適當的編碼及相對應的治療。會使人容易預測的隨機分派細節(如區塊的長度等)不應揭露於試驗計畫書中。隨機分派表須由廠商或獨立的單位安全存檔，以一種能夠確保在整個試驗過程正確地維持盲性的方式。在試驗期間獲得隨機分派表，應考慮緊急情況下可能必須為任何受試者需要解盲的可能性。而解盲所應遵循的步驟、必要的文件以及受試者後續的治療和評估均須在計畫書中載明。

機動隨機分派是另一種隨機分派方法，在該程序中，受試者的治療分派受到當前已分派治療的平衡情況所影響，並且在分層試驗中，受到受試者所屬的層以及該層內平衡情況的影響。應當避免確定性動態隨機分派(deterministic dynamic allocation)，應當為每個治療分派導入適當的隨機化成分。應盡一切努力維持試驗的雙盲。例如，治療組別編碼資訊僅限於執行機動隨機分派的中心試驗辦公室相關人員才知道，該辦公室一般是透過電話聯繫來控制動態分派。如此可對納入條件額外確認再進入試驗，這些特點在某些類型的多中心試驗中可能有價值。一般雙盲試驗事先將藥物包裝並編號的系統仍可採用，但藥物分派順序不再是依序的。最好是使用一個適當的電腦演算法，使中心試驗辦公室的人員對治療組別編碼保持盲性。當考慮採用機動隨機分派時，應仔細評估操作上的複雜性以及對分析的潛在影響。

3. 試驗設計之考量

3.1. 試驗類型

3.1.1. 平行設計

對於確認性試驗，最常採用的試驗設計是平行設計，在該設計中受試者被隨機分派到兩組或多組中的其中一組，每組被分派到不同的治療。這些治療組包括藥品的單一或多個劑量，以及一個或多個對照組，如安慰劑或活性對照組。該設計的假設並不比大多數其他設計複雜。然而，與其他設計一樣，可能會有使分析和解讀試驗結果複雜化的額外試驗特性(例如共變數、重複量測、交互作用、違反試驗計畫書、退出試驗及失訪等)。

3.1.2. 交叉設計

在交叉設計中，每位受試者被隨機分派至一個包含兩種或多種治療的序列中，這是一種以自我為對照組的設計方法。這種設計受歡迎主要是因為可以減少受試者人數，以及通常減

少了達到特定檢定力所需要的評估數量，有時減少的幅度會很大。在最簡單的 2×2 交叉設計中，每位受試者在相繼的兩個治療期分別接受兩種治療。治療的順序是隨機指派，而兩個治療期之間常有一個廓清期(washout period)。最常見設計可擴展到在 n 個治療期比較 $n(>2)$ 種不同治療，每個受試者都接受 n 種治療。這類設計存在有多種不同類型，如每位受試者接受 $n(>2)$ 種治療的一部分(子集)，或受試者接受重複同種治療的設計。

交叉設計有很多問題足以使結論無效，最主要的原因是殘餘效應(carryover)，即治療的殘餘效果存在於後續的治療期。在一個加成模型(additive model)中，不相等的殘餘效應將使治療組的直接比較產生偏差。

在 2×2 設計中，不能從統計學上將殘餘效應從治療組別與期間的交互作用中區分出來，並且因為相應的比較是受試者間的，故這兩種比較都缺乏檢定力。這一問題在高階設計(higher order designs)中不嚴重，但不能完全排除。

因此，在使用交叉設計時，重要的是要避免殘餘效應。最好的方法是在充分瞭解疾病領域和新藥的基礎上有選擇地和謹慎地使用本設計。所研究的疾病應當是慢性的，且病情穩定。治療期間應充分發揮藥物的相關效應。廓清期應該足夠長，以使藥物的作用完全消退。應該在試驗前利用已有資訊及數據確定是否可能滿足這些條件。

在交叉試驗中，有一些額外的問題需要密切注意。其中最值得注意的是，受試者失訪導致分析和解讀都變得很複雜。另外，殘餘的潛在作用導致在後續的治療周期中發生的不良事件，很難判斷是何種治療所導致的。這些問題和其他的問題，在 ICH E4 已有論述。交叉設計一般應限於在試驗中預期僅有少數受試者失訪的情形方可使用。

2×2 交叉設計的一個最常且通常令人滿意的應用是，用以驗證相同藥物的不同劑型的生物相等性。應用於健康受試者時，如果廓清期足夠長，最不可能發生對相關的藥物動力學指標的殘餘效應。然而，由所得到的資料來驗證此一假設仍然是很重要的，例如，確認在每一個治療期開始時並無藥物可被檢出。

3.1.3. 因子設計 (Factorial Designs)

在因子設計，透過的不同治療組合，可以同時評估兩個或多個治療組。最簡單的例子是 2×2 因子設計，在該設計中受試者被隨機分配到兩種治療(如 A 和 B)的四種可能組合之一。四種可能組合為單獨 A，單獨 B，既有 A 又有 B，既無 A 又無 B。在很多情況下，該設計主

要用於檢定 A 和 B 的交互作用。但如果樣本數是基於檢定主效用(main effect)來計算的，則用於檢定交互作用的檢定力將會不足。當該設計被用於檢驗 A 和 B 的聯合效應時，特別是如果兩者可能被一起使用，這種考量是很重要的。

因子設計的另一個重要用處是，建立同時使用治療 C 和 D 時的劑量-反應特徵(dose-response characteristics)，特別是在先前試驗中每種單一療法的有效性已經在某個劑量上被確定時。選擇治療 C 的 m 個劑量(通常包括 0 劑量，即安慰劑)，和相似的治療 D 的劑量數 n。整個設計由 $m \times n$ 個治療組構成，每個治療組接受一種不同的治療 C 和 D 的劑量組合。反應曲面(response surface)的結果估計將有助於在臨床上決定治療 C 和 D 最適合的劑量組合(見 ICH E4)。

在某些情況下， 2×2 設計評估兩種治療的效果所需的受試者人數與單獨評估任一治療組別的有效性所需的受試者人數相同，因此，可以很有效率的使用臨床試驗受試者。對於死亡率非常高的試驗，這種策略已經被證實是非常有價值的。這方法的效率及有用性是基於假設治療 A 和 B 沒有交互作用。因此，治療 A 和 B 在主要療效指標是遵循加成模型，無論有無加上治療 B 的效果，治療 A 的療效實際上都是一樣的。如同交叉設計，應當在試驗之前根據先前的資訊和數據，確立很可能會滿足這種條件的證據。

3.2. 多中心試驗

採用多中心試驗主要有兩個理由。首先，多中心試驗是被大家所接受，能更有效地評估新藥的方法；在某些情況下，它可能是在合理的時間範圍內獲得足夠受試者以滿足試驗目標的唯一可行的方法。原則上，具此性質的多中心試驗可用在臨床研發的各個階段。它可能有幾個試驗中心，每個中心都有大量的受試者；或者，就罕見疾病來說，它們可能有很多試驗中心，而每個中心只有少數幾位受試者。

其次，將一個試驗設計成多中心(和多個研究者)試驗，主要是為其研究結果的後續外推性(見詞彙表)提供更好的基礎。這源於從更廣泛的人群中納入受試者，並在更廣泛的臨床條件中施用藥物的可能性，從而呈現出更具代表性的未來使用實驗情境。在這種情況下，許多研究人員的參與，也可提供關於藥物價值的更廣泛的臨床判斷的潛力。這樣的試驗在藥物研發的晚期將成為確認性試驗，且可能涉及大量的研究者和試驗中心。為了更進一步增進外推性(見詞彙表)，有時會在不同國家執行。

如果要對多中心試驗進行有意義的解讀和外推，在各中心執行該計畫書的方式應該是明確

的和相似的。此外，樣本數與檢定力的計算均假設各中心的治療效果是相同的且是無偏估計的。制訂一個共同的試驗計畫書並以此執行試驗是很重要的。應該盡可能完全地對程序進行標準化。透過召開研究者會議、試驗前的人員培訓、以及試驗期間的嚴密監查，可以減少評估標準和評估方式的差異。良好的設計一般應旨在使得各個中心內各治療組別的受試者分布相同，而且良好的管理應維持這一設計的目標。倘若之後發現需要考量各中心治療效果的異質性，則避免各中心樣本數差異過大的試驗以及避免有些中心樣本數太少的試驗更具優勢，因為這樣減少了在估計治療效果時，不同加權估計值之間的差異(這一點並不適用於所有中心的樣本數都很小的試驗，以及在分析中不考慮中心效應的試驗)。如果事先沒注意這一點，再加上對結果同質性的懷疑，在嚴重的情況下，可能將多中心試驗的價值降低到不能為廠商宣稱的適應症提供令人信服的證據的程度。

在最簡單的多中心試驗中，每一個研究者將負責在一家醫院收納受試者，所以中心指的是研究者或醫院。然而，在很多試驗中，情況會更加複雜。一個研究者可能從幾個醫院收納受試者；一個研究者可能代表一個臨床醫生團隊(協同研究者)，他們都從各自的或幾個相關的醫院收納受試者。一旦對統計模型中關於中心的定義有疑問，計畫書中的統計章節(見第 5.1 章節)都應當在特定試驗的背景下定義該術語(例如，依研究者、場所或地區)。在大多數情況下，可以用研究者來定義中心，在這方面 ICHE6 提供了相關規範。在有疑問的情況下，則目標應該是清楚地定義中心，以便使對影響主要指標測量值及治療效果的重要因子達到同質性。任何將中心合併起來進行分析的規則均須在試驗計畫書中得到合理的解釋並事先訂定，無論何時，應用此方法做決定都應該始終對治療組別保持盲性，例如在對資料作盲性評估時。

在計畫書中應該闡述用於估計和檢定治療效果的統計模型。主要治療效果可首先使用允許有中心之間差異的模型來進行研究，但模型不包含中心與治療組別의 交互作用項。若不同中心之間的治療效果是同質性的，則在模型中常規地納入交互作用項將降低對主要治療效果檢定的效率。當不同中心的治療效果存在真正異質性時，對治療效果的闡述是有爭議的。

在一些試驗中，例如一些大型死亡率試驗，每個中心只有少數幾個受試者，不預期中心會對主要或次要指標有任何影響，因為它們不可能代表具有臨床重要性的影響。在其他試驗中，可能一開始就意識到，每個中心有限數量的受試者將使統計模型中包含中心效應變得不切實際。這些情況下，模型包含中心項是不合適的，而且，這種情況下也沒有必要按中心分層隨機分派。

如果在一個試驗中發現藥物的治療結果顯示是有效的，而每個中心有相當數量的受試者，

則通常應該探索不同中心之間治療效果的異質性，因為這可能影響結論的外推性。透過對單個中心結果的圖示，或通過分析方法，如中心與治療組別間交互作用項的統計檢定，可檢測異質性。當運用統計檢定時，必須注意到在為了檢定主效果而設計的試驗中，檢定交互作用項的檢定力通常是很低的。

如果發現治療效果存在異質性，則應當謹慎地加以解釋，並應力圖從試驗的管理或受試者的特徵等其他方面尋找原因。這樣的原因通常會建議進一步適當的分析和闡述。在缺乏原因的情況下，治療效果的異質性被證實，例如，通過顯著的定量交互作用(見詞彙表)，意味著可能需要其他的估計方法評估治療效果，採用不同中心予以不同權重的估計方法，使治療效果的估計具有穩健性。尤為重要的是，須瞭解任何以顯著的定性交互作用(見詞彙表)為特徵的異質性的基礎，若未能找到合理的解釋，要想可靠地預測治療效果，可能需要進一步執行更多的臨床試驗。

到目前所討論的多中心試驗都是基於固定效應模型(fixed effect model)。而混合模型(mixed model)也可用來探索治療效果的異質性。混合模型是將中心效應和中心與治療組別間的交互作用(treatment-by-center)視為隨機效應，且當中心數量特別多的時候，該模型是適當的。

3.3. 比較的類型

3.3.1. 較優性試驗

從科學上來講，通過安慰劑對照試驗中顯示優於安慰劑、通過顯示優於活性對照組、或通過顯示劑量反應關係，所證實的療效證據是最可信的。這種類型的試驗被稱為較優性試驗(見詞彙表)。除非有明確說明，一般情況下，本指導原則中都假定為較優性試驗。

對於嚴重疾病來說，當存在由較優性試驗證實有效的治療時，採用安慰劑對照試驗就可能被認為是有違倫理的，此時考慮採用活性對照試驗是合乎科學性的。安慰劑對照和活性對照相比的適用性應當根據各個試驗的情況考慮。

3.3.2. 相等性或不劣性試驗

在某些情況下，將研究用藥與參考治療相比的目標不是為了證實較優性。這種類型的試驗依據研究目標可分為兩大類：第一類是相等性試驗(見詞彙表)，第二類是不劣性試驗(見詞彙表)。

生物相等性試驗屬於前一類。在某些情況，進行臨床相等性試驗是為了其他法規上的理由，

例如，學名藥產品與市售產品相比，當化合物不被吸收並因此不存在於血液中時，則需要執行臨床相等性試驗來驗證其相等性。

很多活性對照試驗被設計為顯示試驗用藥的療性不比活性對照組差，這屬於第二類。另一種可能是在試驗中將試驗用藥的多個劑量與建議劑量或與標準藥品的多個劑量相比，這種試驗目的是同時顯示試驗用藥的劑量反應關係，並將試驗用藥與活性對照組進行比較。

活性對照組的相等性或不劣性試驗也可納入安慰劑組，從而在一個試驗中尋求多個目標：例如，它們可以建立與安慰劑相比的較優性，從而驗證試驗設計，並同時評估相對於活性對照藥的有效性與安全性的相似程度。在活性對照的相等性(或不劣性)試驗中，有很多眾所皆知的困難使其無法納入安慰劑組或新藥的多個劑量組，導致缺乏內在效度的測量(與較優性試驗相比)，因此外在效度的評估是必要的。相等性(或不劣性)試驗本質上是不保守的，因此試驗設計或執行中的許多缺失，可能使之偏向得出相等性的結論。基於這些理由，設計這種試驗需要特別注意，執行的過程也須特別小心。例如，特別重要的是儘量減少違反納入條件、順從性不良、退出、失訪、資料缺失及其他違反計畫書事件的發生率，並儘量減少它們對後續分析的影響。

活性對照藥物應謹慎選擇。一個適當的活性對照的例子將會是一個廣泛使用的治療藥物，它在相關適應症中的療效已經在設計良好、紀錄完備的較優性試驗中所確認並量化，以及在周密計畫的活性對照試驗中，它能可靠地預期顯示出相似的療效。為此，新的試驗設計必須與前述較優性試驗具有相同的設計重點(如主要評估指標、活性對照藥的劑量、納入排除條件等)，因活性對照藥物已在前述較優性試驗中清楚地顯示其療效，因此可將臨床與統計實務上進一步的考量納入新試驗中。

於驗證相等性或不劣性的試驗計畫書中，清楚地表述該試驗的明確意圖是極其重要的。計畫書必須訂定相等性臨界值；這個臨界值是臨床上所能接受的最大差異，並且應當小於在活性對照藥於較優性試驗中所觀察到的差異。對於活性對照的相等性試驗，須指定相等性臨界值的上界和下界，但對於活性對照的不劣性試驗，則只需要定臨界值的下界。臨界值的選擇須考量其臨床上之合理性。

統計分析通常是採用信賴區間方法(見第 5.5 章節)。於相等性試驗，應當使用雙尾信賴區間。當信賴區間完全落在相等性臨界值之內時，則可推斷具有相等性。在操作上，這等於同時進行兩個單尾檢定，檢定的虛無假說為治療差異落在相等性臨界值之外，對立假說為治療差異落在相等性臨界值之內。因為這兩個虛無假說之間是無交集的，所以型一誤差可被適

當地控制。不劣性試驗則是使用單尾信賴區間，這種信賴區間方法為一個單尾檢定，所檢定的虛無假說是治療差異(研究用藥減對照組)等於或小於不劣性臨界值，而對立假說是治療差異大於不劣性臨界值。型一誤差大小的選擇應當與使用單尾或雙尾檢定分開來考量。樣本數的計算也應依據這些方法(見第 3.5 章節)。

須特別注意的是，當虛無假說為研究用藥與活性對照藥沒有差異時，基於觀察到兩者無顯著差異的檢定結果，而做出相等性或不劣性的結論是不適當的。

在選擇分析群體時也存在一些特別的議題。從治療組或是對照組退出或失訪的受試者，往往反應不佳，因此，使用全分析群體(見詞彙表)的結果將可能產生趨向顯示相等性的偏差(見第 5.2.3 章節)。

3.3.3. 劑量反應關係試驗

新的研究產品的劑量反應關係可在研發的所有階段通過各種方法得到答案(參見 ICH E4)。劑量反應試驗有多重目標，其中特別重要的有：確認療效、建立劑量反應(dose-response)曲線的形狀和位置、估計適當的起始劑量、決定個別劑量調整的最佳決策、確定最大劑量，超過該劑量不太可能會有更多的好處。為了達到這些目標，應該收集研究用藥不同劑量組的資料，如果可以的話，也包括安慰劑(零劑量)組的資料。為此，估計劑量與反應關係的方法，如信賴區間的建構以及使用圖形法等，都與使用統計檢定同樣重要。而所使用的假說檢定方法可能需要根據劑量的自然順序，或關於劑量反應曲線形狀(如單調性)的特定問題進行調整。擬選用的統計程序的細節須在計畫書中詳細說明。

3.4. 群組序貫設計 (Group Sequential Designs)

群組序貫設計是可用於期中分析的方法(見 4.5 節與詞彙表)。儘管群組序貫設計不是唯一可用於期中分析的設計，但它們是最常被使用的，因為它在試驗期間定期對受試者的資料進行評估，比在每個受試者的資料都收集到時的連續評估更可行。在獲得治療結果和受試者所分派的組別(如解盲，見第 4.5 章節)的訊息之前，應在計畫書中充分說明統計方法。獨立數據監測委員會(見詞彙表)可對源自於群組序貫設計的數據進行審查或者執行期中分析(見第 4.6 章節)。儘管群組序貫設計已被廣泛且成功地應用於觀察死亡率或非致命結果的療效指標之大型、長期的試驗，它在其他情況下的應用也在增加。特別的是，由於已認識到在所有試驗中都必須監視安全性，故因安全性理由而提早終止試驗的正式程序也都需考慮到。

3.5.樣本數

臨床試驗中的受試者數目必須足夠多，以確保能對所提出的問題給予可靠的答案。樣本數大小通常依試驗的主要目標來決定。若是以其他根據來決定，則須清楚說明並合理解釋。例如，一個基於安全性問題或需求，或者基於重要的次要目標來決定樣本數的試驗，可能會比基於主要療效問題來決定樣本數的試驗需要更多的受試者(見 ICH E1A)。

用以決定適當樣本數的一般的方法，需要確定以下各項：主要指標、統計檢定量、虛無假說、所選劑量下的對立假說(呈現出對所選受試者族群和所選的劑量下的治療差異被檢測出或被拒絕的考量)、錯誤地拒絕虛無假說之機率(型一誤差)、錯誤地不拒絕虛無假說之機率(型二誤差)，以及對失訪、違反計畫書的處理方法。在某些情況下，事件率是估計檢定力的主要參數，因此須做出一些假設，再以所需的事件數推算出試驗的最終樣本數。

計算樣本數的方法及計算時所需的統計估計值(如變異數、平均數、反應率、事件率、預期的治療差異值)須明定於試驗計畫書中，亦須提供這些估計值的依據。探討樣本數對現有假設之各種偏離程度的敏感性是很重要的，這可以藉由假定一個合理的偏離區間而得到一個樣本數的範圍。在確認性試驗中，樣本數計算的假設一般應依據公開發表的文獻資料或早期試驗的結果。預期的治療差異值可基於臨床治療這類病人中，可接受的最小的療效或者對新治療方法預期效果的判斷，兩者相比之下後者之值較大。按照慣例，型一誤差機率設在 5% 或更小，或者由多重比較考慮所需要的任何調整來決定，精確的選擇可能會受到所檢定的假設之先驗合理性及這些結果的預期效果之影響。型二誤差機率傳統上都定在 10% 到 20%，廠商會希望在可行的範圍內越低越好，特別是當試驗很難或不可能再次執行時。相較於傳統型一與型二誤差數值的設定，在某些情況下，其他替代數值或許也可接受，甚至更為合適。

樣本數的計算須參考主要分析所需的受試者人數。如果是「全分析群體」，則估計的療效大小會比「依計畫書群體」(見詞彙表)還小，這是因為納入失訪或順從性差的受試者的數據所產生的治療效果被稀釋的狀況。因此，關於變異性的假設可能也需要修改。

相等性或不劣性試驗(見 3.3.2 節)的樣本數估算，通常應基於達到治療差異的信賴區間的目標，該信賴區間表明治療組間的差異至多為臨床上可接受的差異。當在真實的差異為 0 的條件下計算相等性試驗的檢定力時，如果真實差異不為 0，則欲達到此檢定力所需的樣本數將會被低估。當在真實的差異為 0 的條件下計算不劣性試驗的檢定力時，如果研究藥品的療效低於活性對照組時，則欲達到此檢定力所需的樣本數將會被低估。選擇一個「臨床

上可接受的」差異須說明它對將來病人的意義，且它可能小於上述旨在確定存在差異的較優性試驗中所提到的「臨床上有關的」差異。

在群組序貫試驗中不能預先確定確切的樣本數，因為其值依賴於機會作用(chance)、所選擇的試驗停止條款以及真實的治療差異之綜合考量。試驗停止條款的設計須考量後續樣本數的範圍，通常這會落在期望樣本數和最大樣本數之間。

當事件發生率低於預期或變異性大於預期，在不解盲數據或不進行治療組間比較的情形下，可藉由適當的統計方法重新估算樣本數(見第 4.4 章節)。

3.6.數據收集與處理

收集數據並從研究者向廠商傳送數據可透過各種媒介進行，包括紙本病例記錄表、遠端監視系統、醫學電腦系統和電子傳輸器。無論採用何種數據收集工具，所收集數據的形式和內容都應完全符合計畫書的要求，並應在試驗執行前確定。它應當著重於收集預計執行分析所需的數據，包括確認計畫書順從性或確定重要計畫書偏差所需要的前後相關資訊(如服藥的時間)。缺失值應該與「0 值」或不存在(characteristic absent)相區別。

從數據收集到資料庫最終確定的過程應該按照藥品優良臨床試驗規範(ICH E6，第 5 章節)進行。具體來說，需要有及時可靠的數據記錄流程以及更正錯誤和增補遺漏，此為建立高品質資料庫並透過預計執行的分析達到試驗目標所必需的。

4. 試驗執行之考量

4.1.試驗監測與期中分析

按照試驗計畫書謹慎地執行臨床試驗，對試驗結果的可信度有著重要的影響(見 ICH E6)。謹慎地監督試驗的進行能及早發現問題，並使問題的發生或再發生減至最小。

由廠商發起的確認性臨床試驗，通常有兩種不同的監測類型。一種類型是監測整個試驗的品質，另一種是以解盲方式進行治療組間的比較(如期中分析)。這兩種類型的監測方法，除人員職責不同外，所用的數據類型和資訊也不同，須以不同原則來控制統計上與操作上的可能偏差。

為了監督試驗的品質，試驗監測所涉及的檢查可能包括：是否按計畫書進行，所收集到的

數據是否可接受，是否達到預期收案的目標，設計的假設是否適當，以及是否成功地讓病人保留在試驗中等(見第 4.2 至 4.4 章節)。這種類型的監測既不需要獲取治療組間比較的資訊，也不需要將資料解盲，因此對型一誤差沒有影響。出於此一目的而對試驗進行監測是廠商的職責(見 ICH E6)，它可以由廠商或廠商指定的獨立小組來進行。這種類型的監測一般是從選定試驗地點開始，到收集和清理最後一位受試者的數據結束。

另一類的試驗監測(期中分析)涉及到收集治療組間比較的結果。期中分析需要解盲以獲取治療組別的分派訊息以及治療組別間比較的總結訊息。因此，計畫書(或首次分析前適當的計畫書變更)中須包含期中分析的統計計畫，以防止某些類型的偏差。這將在第 4.5 和 4.6 章節中討論。

4.2. 變更納入與排除條件

在整個受試者招募期間，納入與排除標準應按照計畫書的規定保持不變。偶爾對其做些變更也是合適的。例如，在長期的臨床試驗中，從外部或從本試驗期中分析中不斷獲得的醫學知識，就有可能建議對納入條件進行變更。此外，納入排除標準的改變，可能來自於監測人員的發現，如他們在監測中發現，試驗常常不能按納入的條件選擇受試者，或由於太嚴格的納入條件導致收納速度非常慢。納入排除條件的變更，應該在不解盲的情況下進行，並應於計畫書變更中描述其改變，內容應包括後續對統計上的影響，如不同的事件發生率所致的樣本數調整；或者對預定統計分析的修改，如根據變更的納入排除條件進行分層分析。

4.3. 受試者收納速度

在受試者收納時間較長的試驗中，應監視收納速度，如果收納速度明顯低於預期，應查明理由並採取補救措施，以維持試驗的檢定力並減少對選擇性納入(selective entry)和其他試驗品質方面的疑慮。在多中心試驗中，這些考量適用於每一個中心。

4.4. 樣本數調整

在長期的臨床試驗中，通常會對原設計和樣本數計算所依據的假設進行審視。如果試驗的規劃是根據初步的或不確定的資訊制定的，這種調整尤為重要。在不解盲的情況下，對數據進行期中檢查，可能發現所得總反應變異數(overall response variances)、事件發生率或存活狀況不如預期。則應適當地修訂假設條件，重新計算樣本數，並且應在計畫書變更和臨床試驗報告中清楚說明它的合理性並做成記錄。應闡述為保持盲性而採取的步驟，以及

對型一誤差和信賴區間後續的影響。試驗有重新估算樣本數的潛在需要時，宜盡可能於計畫書中預先設定(見 3.5 節)。

4.5.期中分析與試驗提早終止

期中分析是指在臨床試驗正式結束之前的任何時間，為了比較治療組間的有效性或安全性而進行的任何分析。因為這些治療組間比較的次數、方法及結果將對試驗結論的闡釋產生影響，因此所有的期中分析都必須預先仔細地計畫，並在試驗計畫書中闡明。一些特殊情況下，可能需要執行未事先計畫的期中分析。在這些情況下，應於解盲分析前變更試驗計畫書增加期中分析的章節。當期中分析是為了決定是否終止試驗而設計時，這通常採用以統計監督計畫為指導原則的群組序貫設計(見第 3.4 章節)。此類試驗期中分析的目標是當試驗組的較優性已被清楚地建立，或有意義的療效差異已經不可能被建立，或出現了無法耐受的不良反應時能及早停止試驗。一般而言，與監測安全性而設定的界限相比，監測療效而設定的界限需要更多的證據才能讓試驗提早終止(即療效監測更保守)。當試驗計畫或監測目標包含了多個評估指標時，則須有多重檢定的考量。

試驗計畫書中應描述期中分析的時間表，或執行期中分析的相關考量，例如採用彈性的 α 消耗函數(alpha spending function)，須於計畫書或第一次期中分析前之計畫書變更中寫明。試驗終止的規則及其特性須在試驗計畫書或變更中詳述。試驗提前終止對其他重要指標的分析的潛在影響也需要考量。當試驗有數據監測委員會時，這些內容應該由數據監測委員會撰寫或核准(見第 4.6 章節)。偏離預先計畫的程序常有使試驗結果變成無效的可能。如果試驗計畫書需要變更，則任何相應的統計方法須儘早在計畫書變更中寫明，應特別就此變更可能對任何分析或推論所產生的影響進行討論。所選定的方法必須確保整體型一誤差機率得到控制。

期中分析的執行應該是一個完全保密的過程，因為它可能涉及解盲的數據及結果。參與試驗的所有人員必須對這些分析的結果保持盲性，否則他們對試驗的態度可能改變，進而影響新納入受試者特徵的改變或者在治療組別之間的比較產生偏差。這一原則可應用於除了直接參與執行期間分析的人員之外的所有研究人員和廠商聘雇人員。研究者應僅被告知是繼續試驗、或停止試驗，或是對試驗程序進行調整。

大部分用以支持研究用藥的療效和安全性的臨床試驗，須完成預計收納的樣本數。只有在遇到倫理方面的原因，或者檢定力大小不再是可以接受時，應當提早終止試驗。儘管如此，藥物研發計畫由於各種理由，廠商有查閱治療組間比較性數據的需要，比如要設計其他的

試驗。大家也公認，有一部分的試驗會採用嚴重威脅生命的指標或死亡率，基於倫理考量，這些都需要對不斷積累的比較性療效結果進行持續的監測。在以上這些情況中，統計相關的期中分析內容應在解盲前於試驗計畫書或計畫書變更中載明，以避免可能產生的統計上和操作上的偏差。

對於許多新藥的臨床試驗，特別是與公眾健康關係重大的新藥，應該另外指定一個外部獨立小組，負責監督有效性和安全性結果的比較。這個小組通常被稱為獨立數據監測委員會 (Independent Data Monitoring Committee, IDMC)，或數據與安全性監測委員會 (Data and Safety Monitoring Board)，或數據監測委員會 (Data Monitoring Committee)，其職責應明確地說明。

當廠商負起監測有效性或安全性比較的職責，並因此從中獲得解盲的比較性結果時，須特別注意保持試驗的完整性，並適當地管理和限制資訊分享。因此，廠商必須保證並記錄內部監測委員遵守了書面的標準操作程序，以及保存含有期中分析結果記錄在內的決策會議記錄。

任何設計不良的期中分析(不管是否提早終止試驗)都可能導致偏差的產生，所得的結論也缺乏可信度。因此，應該避免這樣的分析。如果執行了計畫外(unplanned)的期中分析，必須在試驗報告中解釋其必要性與解盲的程度，並評估可能導致偏差的嚴重程度以及對試驗結果解讀的影響。

4.6.獨立數據監測委員會(IDMC)的角色(見 ICH E6 第 1.25 與 5.52 章節)

獨立數據監測委員會(IDMC)可以由廠商成立，它主要的任務是每隔一段時間會對臨床試驗的進展、安全性資料和關鍵療效指標進行評估，並向廠商建議這個試驗是可以繼續進行、須調整設計或是須終止。該委員會(IDMC)應當有書面的操作程序書，並保存所有的會議記錄，包括期中分析結果；當試驗結束時，這些紀錄應可供查閱。該委員會(IDMC)的獨立性旨在它能控制重要的比較性試驗數據的分享，又能防止因可接觸試驗資料而對試驗的完整性造成不利的影響。該委員會(IDMC)是與人體試驗委員會(Institutional Review Board, IRB)或獨立倫理委員會(Independent Ethics Committee, IEC)相獨立的機構，它的成員應包括通曉統計學等相關學科專業知識的臨床試驗科學家。

當獨立數據監測委員會(IDMC)中有廠商代表參與時，這些代表的權責應在委員會的操作程序書(operating procedures)中加以明確規定(例如，在關鍵議題上是否具有投票權)。由於這些

廠商代表能接觸到解盲的資訊，因此這些操作程序還應探討如何控制期中分析結果在廠商內部傳播。

5. 資料分析的考量

5.1. 預先設定分析方法

在設計臨床試驗時，最終的統計分析之主要特徵(principal features)須在計畫書的統計章節中進行描述。這個章節應包括主要指標的確認性分析之所有主要特徵，以及對預期分析問題的解決方法。至於探索性試驗，這一章節可描述更一般的原則和方向。

統計分析計畫(見詞彙表)可以在試驗計畫書完成後單獨成文撰寫。在該文件中，可以更技術性地和更詳細地闡述在計畫書中所描述的主要特徵(見第 7.1 章節)。該計畫可包括對主要指標、次要指標及其他資料進行統計分析的詳細步驟。在對試驗的資料進行盲性評估時(見 7.1 的定義)，應對統計分析計畫再檢視並進行可能的變更，但須在解盲之前完成。統計分析計畫完成的時間點與各解盲的時間點，都應保留正式的記錄。

如果盲性評估的結果是建議更改計畫書的主要特徵，則須在試驗計畫書變更中記錄這些更改。否則可根據盲性評估結果的建議，對統計分析計畫做出修改。唯有按照試驗計畫書(包括所有計畫書變更)中所設定的統計分析來執行，其結果才能被視為是確認性的。

在臨床試驗報告的統計章節中，應清楚地描述統計方法，包括在臨床試驗過程中何時做出方法學的決策(見 ICH E3)。

5.2. 分析群體

用於主要分析之受試者資料須在計畫書中的統計章節明確定義。另外，紀錄所有受試者開始進入試驗時(例如，納入期間; run-in period)的資料是有用的。該記錄的內容須取決於試驗各自的特性，但是，應盡可能收集人口統計學以及關於疾病狀態的基礎值資料。

如果所有接受隨機分派進入臨床試驗的受試者均符合所有納入條件、完全遵循試驗程序且無失訪，並提供完整的數據記錄，很明顯地，這樣的受試者都會包括在分析中。試驗設計與執行的目標應該是盡可能地達到這一理想狀態，但實際上，它是否能完全達到是值得懷疑的。因此，計畫書中的統計章節應前瞻性地處理預期會發生的問題，並說明這些問題如何影響受試者和需要分析的數據。計畫書中還須訂出一些做法，以減少試驗執行中任何預

期的違規行為的發生進而損害分析的完善。這些行為包括各種類型的違反計畫書、退出及缺失值。計畫書應考慮降低這類問題發生頻率的方法以及解決分析中發生這類問題的處理方式。在盲性評估時間，應該確定統計分析對處理違反計畫書事件方式的可能變更。最好是確定任何重大違反計畫書事件的發生時間、原因及對試驗結果的影響。違反計畫書事件、缺失值以及其他問題的頻率和類型均應記錄於試驗報告中，並且應當描述它們對試驗結果的潛在影響(見 ICH E3)。

關於分析群體的決定應該遵循以下兩個原則：(1)使偏差減到最小；(2)避免型一誤差的擴增。

5.2.1. 全分析群體

意圖治療(intention-to-treat，見詞彙表)的基本原則是：主要分析應包括所有接受隨機分派的受試者。遵循此一原則需要對所有接受隨機分派受試者進行完整的追蹤(follow-up)以獲得試驗結果。但實際上這一理想難以達到，理由後述。因此，在此指引中，「全分析群體」這個術語被用來描述盡可能完整的分析群體，以及盡可能接近包括所有接受隨機分派受試者意圖治療之理想狀態的分析群體。在分析中保持初始的隨機分派，對於防止偏差以及為統計檢定提供安全的基礎是很重要的。在許多臨床試驗中，使用「全分析群體」是保守的策略。在許多情況下，它提供的治療效果估計更能反映爾後實際上的情況。

在一些有限的情況下，可能導致從「全分析群體」中排除已接受隨機分派的受試者，包括不符合主要納入條件(違反納入資格)、未接受過至少一次試驗用藥以及缺乏隨機分派後的任何資料。這些被排除的受試者都需要說明其合理性。只有在以下情況下，不符合納入條件的受試者可以從分析中排除而不會造成偏差：

- (1) 在隨機分派之前已判定了納入條件。
- (2) 可以完全客觀地檢查相關的資格不符。
- (3) 所有受試者接受相同的資格違反審查(這在開放性試驗中很難保證，即使在雙盲試驗中，若資料在審查之前被解盲，同等審查就很難保證，所以要強調盲性評估的重要性)。
- (4) 違反特定納入條件的所有受試者已被排除。

在某些情況下，從所有接受隨機分派受試者中，排除任何未服用過試驗藥物的受試者可能是合理的。例如，是否開始給予治療的決定，不受知曉受試者所接受的治療組別的影響時，雖然排除這些受試者，但意圖治療原則仍將被保持。在其他情況下，可能需要從所有接受隨機分派受試者中排除任何隨機分派後沒有資料的受試者。除非來自這些特定排除的或任何其他原因所引起的可能偏差都已得到解決，否則沒有一個分析是完整的。

當使用全分析群體時，在隨機分派後的違反計畫書可能會對資料和結論造成影響，尤其是當它們的發生與治療組別有關時。在很多時候，將這些病人的資料納入到分析中是合適的，這與意圖治療原則相一致。然而，特殊疑慮在於：接受一次或多次劑量後退出治療且在此後不提供任何資料的受試者，以及因其他原因而失訪的受試者，因為不把這些受試者納入到全分析群體中，可能會嚴重地破壞這個方法的基礎。因此，無論基於任何理由在病人失訪下測量的主要指標，或是隨後根據計畫書預計日程所收集到的主要指標，都具有分析價值；在主要指標是死亡率或嚴重發病率的研究中，後續的資料收集特別重要。若意圖以這種方法收集資料應在計畫書中載明。插補的技巧，從將最後觀察值遞延插補(last observation carrying forward)，到使用複雜的數學模型的方法，都可以用來插補缺失值。其他用於全分析群體中主要指標的插補方法，可能會要求對受試者的結果做一些假設，或者對結果做較簡單的選擇(例如，成功或失敗)。使用這些策略應在計畫書的統計章節描述並說明其選取之合理性，所採用的數學模型其假設也應解釋清楚。特別是當所述的策略本身可能導致療效的估計有偏差時，證實相關的分析結果的穩健性也很重要。

由於一些問題的不可預測性，有時針對試驗期間所出現的無規律性問題，推遲到試驗結束對數據做盲性評估後，再做出詳細考量會更好。若打算這樣做，應在試驗計畫書中清楚說明。

5.2.2. 依計畫書群體

依計畫書群體的受試者有時又稱為有效病例(valid cases)、療效樣本(efficacy sample)或可評估受試者(evaluable subjects)樣本，它是全分析群體的一個子集，這些受試者對計畫書更具順從性，並符合以下特徵：

- (i) 完成了對治療療程的某個預定的最小暴露量；
- (ii) 具有主要指標的量測值。
- (iii) 無任何重大的違反計畫書事件，包括違反納入條件。

在按照適合於特定試驗情況的方式解盲前，應該充分地定義並紀錄將受試者排除在依計畫書群體之外的確切理由。

使用依計畫書群體可以使新的治療在分析中展現出額外有效性的機會最大化，並且更密切地反映了計畫書基礎下的科學模型。然而，依據試驗的不同，相應的假說檢定和治療效果的估計可能是保守或是不保守；當對計畫書的順從性(adherence)與接受的治療和結果有關

時，由此依計畫書群體分析所產生的偏差可能是嚴重的。

將受試者排除於依計畫書群體之外的理由，以及其他的違反計畫書事件，應當清楚記載並加以總結。相關的違反計畫書事件可能包括治療組別分派的錯誤、使用禁用的藥物、順從性不佳、失訪和缺失資料。評估各治療組間關於這些事件發生頻率與發生時間的模式，是一種良好的做法。

5.2.3. 不同分析群體的角色

一般說來，顯示出不同分析群體對主要試驗結果缺乏敏感性是有利的。在確認性試驗中，同時對全分析群體及依計畫書群體進行分析，一般來說是恰當的，由此可以明確地討論和解釋他們之間的任何差異。在某些情況下，可能期望進一步探討不同分析群體的選擇對結論的敏感性。當全分析群體和依計畫書群體得出實質上相同的結論時，試驗結果的可信度會增加。然須銘記在心的是，若需要將相當大比例的受試者排除於依計畫書群體之外，則會對試驗的整體可信性(overall validity)產生疑慮。

在較優性試驗(試圖顯示研究產品的效果是更好的)和相等性或不劣性試驗(試圖顯示研究產品具有可比性，見第 3.3.2 章節)中，全分析群體和依計畫書群體發揮不同的角色。在較優性試驗中，全分析群體用於主要分析(除了例外情況)，因為它傾向於避免依計畫書群體所導致的對療效過度樂觀的估計，這是因為全分析群體包含了順從性不佳的受試者，這類受試者一般會降低療效的估計。然而，在相等性或不劣性試驗中，採用全分析群體一般而言並不保守，應當非常仔細地考慮它的角色。

5.3. 缺失值與離群值(outliers)

缺失值是臨床試驗中一個潛在偏差的來源。因此，應盡一切努力符合試驗計畫書中關於資料收集和管理的所有要求。然而，實際上，試驗幾乎不可避免地會有缺失值。雖然如此，只要處理缺失值的方法是合理的，並且在計畫書中預先定義，則試驗可以被認為是可信的。在盲性評估期間，於統計分析計畫中更新此方面的內容，可以使統計方法的定義更完善。遺憾的是，尚無一個普遍適用的缺失值處理方法可供推薦。須探討分析結果對不同缺失值處理方法的敏感性，特別是當缺失值的數量很大時。

應當採用類似的方法探討離群值的影響。離群值在統計學上的定義，從某種程度上來講是主觀認定的。當從醫學上和統計上共同判斷某一特定值明確地確定是離群值，才最具有說服力，並且醫學上往往會定義適當的後續行動。在計畫書或統計分析計畫中訂定的任何離

群值的處理方式不應當對任何一個治療組別有偏袒。同樣地，在盲性評估期間可以有效地更新這方面的分析。如果在試驗計畫書中未事先訂定處理離群值的方法，則應執行一個使用實際資料分析所得的結果，以及至少執行一個排除或減少離群值影響的其他分析，並對結果之間的差異加以討論。

5.4.數據轉換

對重要的指標是否要進行轉換，最好根據早期臨床試驗的類似資料，並在試驗設計階段就決定好。擬採用的轉換(如平方根、對數)及其合理性須在試驗計畫書中說明，特別是對於主要指標。在標準教科書上可以找到變數轉換的一般原則，以確保資料能滿足統計方法所依據的假設；一些特定指標常用的轉換方法已在某些特定的臨床領域中使用。一個指標是否轉換，以及如何轉換的決定應取決於是否有助於臨床解讀。

類似的考量也適用於其他衍生指標(derived variables)，例如，從基礎值的改變量(change from baseline)、從基礎值的改變百分比(percentage change from baseline)、重複測量的「曲線下的面積」，或者是兩個不同指標的比值。惟須仔細考量後續在臨床上的解讀，並在計畫書中說明衍生的合理性。與此密切相關的重點已在 2.2.2 章節中討論過。

5.5.參數估計、信賴區間及假說檢定

為了滿足試驗的主要目標，應該在計畫書的統計章節中詳細說明所要檢定的試驗假說和所要估計的治療效果。為了完成這些任務，針對主要指標(最好亦針對次要指標)的統計方法，以及所依據的統計模型應清楚闡述。只要有可能，治療效果的估計應盡可能附上信賴區間，並應確認它們的計算方法。應說明是否打算使用基礎值資料來提高估計精準度或調整對潛在基礎值差異的估計，例如，採用共變數分析。

明確說明所採用的假說檢定是單尾還是雙尾是重要的，尤其是要採用單尾檢定時，須事先說明其合理性。如果認為假說檢定不合適，則須提出可以得到統計結論的其他方法。關於統計推論採用單尾還是雙尾的問題是有爭議的，在統計文獻中可以找到不同的觀點。從法規的觀點來看，將單尾檢定的型一誤差設定為雙尾檢定型一誤差的一半是較合適的方式。這與通常用於估計兩種治療組間差異大小的雙尾信賴區間相一致。

所選用的統計模型應當要反映目前醫學和統計學關於所分析的指標以及試驗設計的知識現況。所有在分析中納入的影響因子應完整地說明(例如在變異數分析模型)，並應解釋根據初步結果修改這組因子的方法(如果有)。同樣的考量也適用於在共變數分析中所納入的共變

數集合(見第 5.7 章節)。在選擇統計方法時，應注意到主要和次要指標的統計分布。在選擇時(例如，選擇母數還是無母數方法)，重要的是要記得提供治療效果大小統計估計值及其信賴區間(除了提供統計顯著性檢定之外)。

應當清楚地區分主要指標的主要分析與主要或次要指標的支持性分析。在計畫書的統計章節或統計分析計畫中，也應概述主要和次要指標之外的資料總結和報告的方法。為了在一系列試驗中達到分析一致性的目的，例如對安全性資料，應當包括所採用方法的介紹。

關於已知的藥理學參數、每個受試者對計畫書的順從性或其他基於生物學上的數據，整合這些訊息以建立模型的方法，可以對實際上或潛在的療效提供有價值的見解，特別是對於療效的估計。這類模型所依據的假設應清楚地說明，並仔細描述任何結論的侷限性。

5.6.顯著水準與信賴水準的調整

當存在多重性(multiplicity)時，頻率學派常用的方法會是調整型一誤差。多重性可能來自於，例如，多個主要指標(見第 2.2.2 章節)、治療組間的多重比較、隨時間的重複量測或期中分析(見第 4.5 章節)等。在可行的情況下，最好採用避免或減少多重性的方法，例如，指定一個關鍵的主要指標(當具有多個指標)、選擇關鍵的治療組比較(當存在多重比較)、運用綜合性測量如「曲線下面積」(當具有重複量測)。於確認性分析中，採取此類步驟之後，若仍有多重檢定的問題，也須在試驗計畫書中載明；應該始終考慮調整，並應在分析計畫中寫明任何調整型一誤差的詳細步驟，或是解釋為什麼沒有必要調整。

5.7.次群體、交互作用與共變數

主要指標通常系統性地受到治療以外的其他因素影響，例如，主要指標可能與年齡或性別等共變數有關，或者在多中心試驗中，不同中心接受治療的受試者之特定次群體間可能存在差異。在有些情況下，對共變數影響的調整或者對次群體效應的調整是分析計畫中不可或缺的部分，應在試驗計畫書中陳述。試驗執行前須確定那些預計會對主要指標產生重要影響的共變數和因子，並應思考如何在分析中將這些因素納入考量，以提高估計的精準度以及改善治療組間的任何不平衡。如果使用一個或多個因子對試驗設計進行分層，則在分析中考慮這些因子是適當的。當對調整的潛在價值有疑問時，通常建議將未經調整的分析作為主要關注的分析，把調整後的分析作為支持性的證據。應特別注意中心效果(center effects)和主要指標基礎值的作用。不建議在主要分析中對隨機分派後測量的共變數進行調整，因為它們可能受治療的影響。

治療效果的大小可能會因次群體或共變數的不同而不同，例如，治療效果可能隨年齡的增

加而減少或者可能在特定診斷類別(particular diagnostic category)的受試者中更大。在某些情況下，這些交互作用可被預期或者特別感興趣(如老年醫學)。因此，次群體分析或包含交互作用項的統計模型，是計畫中確認性分析的一部分。然而，在大多數情況下，次群體分析和交互作用分析都是探索性的，並且應當明確地認定為是探索性的；它們應當探索整體發現的任何治療效果的一致性。一般來說，這些分析應該首先在所討論的統計模型中添加交互作用項來進行，再以在相關受試者的次群體內或者由共變數定義的層內進行額外的探索性分析。當進行探索性分析時，應謹慎解讀這些分析；僅僅基於探索性次群體分析的治療有效性(或缺乏有效性) 或安全性的任何結論，都不太可能被接受。

5.8.數據的完整性與電腦軟體的效度

資料分析結果的可靠程度取決於用於數據管理(包括數據輸入、儲存、查核、更正和檢索)以及在統計上資料處理的方法和軟體(包括內部和外部編寫)的品質(quality)和有效性(validity)。因此，數據管理應當基於全面和有效的標準作業程序。用於資料管理和統計分析的電腦軟體應當是可靠的，並應提供適當的軟體測試程序(software testing procedures)的文件。

6. 安全性與耐受性評估

6.1.評估的範圍

在所有的臨床試驗中，安全性和耐受性(見詞彙表)的評估是重要的。在試驗的早期階段，這種評估主要是探索性質的，並且只對藥物直接造成的毒性具敏感性(only sensitive to frank expressions of toxicity)，而在晚期階段，可由較多的受試者經驗更加全面地描述所建立的藥物安全性和耐受性概況。晚期階段的對照組試驗代表了以無偏(unbiased)的方式探索任何新的潛在藥物不良反應的重要方法，即使這些試驗在這方面通常都缺乏檢定力。

出於對安全性與耐受性的較優性或相等性做出特定主張的目的，可以設計某些試驗，與其他藥物或與研究藥物的其他劑量相比。應當通過來自確認性試驗的相關證據來支持這些特定主張，這跟宣稱療效要求所需要的證據是一樣的。

6.2.指標的選擇與數據的收集

在任何臨床試驗中，用於評估藥物的安全性和耐受性的方法及量測準則將取決於許多因素，包括對與之相近藥物的不良反應的了解、來自非臨床試驗或早期臨床試驗的資訊、以及特定藥物藥效學及藥動學的可能結果、給藥方式、待研究的受試者類型以及試驗持續的時間。

有關臨床化學和血液學、生命徵象(vital signs)和臨床不良事件(疾病、徵象和症狀)的實驗室檢驗，通常構成了安全性與耐受性資料的主體。發生嚴重不良事件以及因不良事件導致停止治療在法規審查上是特別重要的(見 ICH E2A 及 ICH E3)。

此外，建議在整個臨床試驗規劃中採用一致的方法來收集和評估資料，以便日後可以合併不同試驗的資料。使用通用的不良事件分類詞典(common adverse event dictionary)尤為重要。該詞典的編排應將不良事件資料整理為三種不同層級的分類，即系統-器官分類(system-organ class)、選用詞(preferred term)或納入詞(included term)(見詞彙表)。選用詞是處在整合不良事件的層級上，在資料的描述性呈現中，將屬於同一系統-器官級的選用詞進行合併(見 ICH M1)。

6.3.分析群體與數據的呈現

評估整體安全性與耐受性時，用於總結的分析群體常被定義為至少接受了一劑研究藥物的受試者。應盡可能全面地從這些受試者中收集安全性及耐受性指標，包括不良事件類型、嚴重程度、發病時間和持續時間(見 ICH E2B)。可能也需要在特定的次族群，如女性、老年人(見 ICH E7)、嚴重疾病或接受常見的併用藥物之受試者，進行額外的安全性及耐受性評估。這些評估可能需要探討更具體的議題(見 ICH E3)。

在評估過程中需要重視所有安全性及耐受性指標，所用的分析方法須在計畫書中載明。所有的不良事件都應報告，無論它們是否被認為與治療有關。在評估中應當考慮試驗群體中所有可用的資料。應當謹慎地定義測量值的單位和實驗室指標的參考值範圍，如果在同一試驗出現不同單位及不同的參考值範圍(例如，如果涉及一個以上的實驗室)，則測量值須進行適當的標準化，以便進行統一評估。毒性分級尺度也必須事先訂定，並說明其合理性。

某種不良事件的發生率通常是以出現不良事件的病人數與暴露病人數之比來表示。然而，如何評估發生率並不總是顯而易見的。例如，根據不同情況，可考慮把暴露的受試者人數或暴露程度(用人年表示，person-years)作為分母。無論目的是用來估計風險還是進行治療組間的比較，在試驗計畫書中定義計算的方式是很重要的。尤其是當計畫進行長期治療，並預期會有較大的退出治療或死亡的比例時，這一點非常重要。對於這些情況，應考慮採用存活分析方法，並計算累積不良事件率，以避免低估的風險。

在一些情況下，當存在大量徵象(signs)和症狀(symptoms)的背景雜訊干擾時(例如，在精神疾病的臨床試驗)，我們在估計不同的不良事件風險時，應該考慮對此進行解釋的方法。其中一種方法是利用「治療後發生(treatment emergent)」(見詞彙表)的概念，只有當該不良事

件發生，或相對於治療前之基礎值發生惡化時，才記錄它們。

減少背景雜訊影響的其他方法，如忽略輕度不良事件，或要求在重複訪視時觀察到的不良事件方可列入分子計算，也是合適的。這些方法應在計畫書中解釋並說明其合理性。

6.4.統計評估

安全性與耐受性的研究是一個多方面(multidimensional)的問題。儘管某些藥物的不良反應通常可被預期而進行監測，但藥物可能的不良反應範圍很廣，而新的、未預期的不良反應總是有可能發生。此外，在違反試驗計畫書後(例如，使用禁用藥物)才發生的不良事件，也可能導致偏差。此種情況造成了統計在分析與評估藥物安全性及耐受性上的困難，這意味著由確認性臨床試驗得到有關安全性的結論只是一種例外而不是通例。

在大多數試驗中，最好採用描述性統計來評估藥物安全性與耐受性，並輔以信賴區間以利解讀。利用圖形來呈現治療組內及受試者間的不良事件型態(patterns)亦有價值。

計算 p 值有時也有幫助，既有利於評估感興趣的差異，對於大量安全性與耐受性指標又可作為一種標記工具以突顯出值得進一步注意的差異。這對於實驗室數據特別有用，因為很難有其他適當的彙整方法。建議對實驗室數據要進行定量(quantitative)分析，例如估計治療效果的平均數，也要進行定性(qualitative)分析，計算高於或低於某一閾值的受試者個數。

如使用假說檢定，則在進行多重比較時，須進行統計上的調整以控制型一誤差，但通常也須留意型二誤差的大小。如未對多重檢定問題進行調整時，在解讀是否具統計上有意義的結果時須特別謹慎。

在大多數臨床試驗中，研究者希望確定，與活性對照藥及安慰劑相比時，在安全性及耐受性方面未出現臨床上不可接受的差異。這與不劣性或相等性方式評估療效的情況相同，在這種情況下，使用信賴區間比假說檢定更為適當。如此，因發生頻率較低而造成較大的估計不精確性可被清楚地呈現。

6.5.綜合性總結

在藥物的研發過程中，特別是進行上市申請時，通常會在不同試驗間持續地總結藥物安全性與耐受性的特性。然而，這個總結的可用性須基於適當、良好控制且具有高品質數據的個別試驗。

藥物的整體可用性(overall usefulness)始終是風險與利益之間的平衡問題；即使風險利益的評估通常是在整個臨床試驗規劃的總結中進行，但在單一試驗中也可考慮這個觀點(見 7.2.2 章節)。

有關安全性與耐受性報告的更多細節，請見 ICH E3 第 12 章。

7. 報告撰寫

7.1. 評估與報告

如緒論所述，臨床試驗報告的格式與內容是 ICH E3 的主題，該指引涵蓋了適當整合臨床及其他資料的統計工作報告。因此，本章節只作簡單的討論。

如第 5 章節所述，在試驗設計階段，統計分析的主要特點應在計畫書中載明。當試驗結束，收集資料並進行初步檢視時，對在第 5 章節中所描述的計畫好的分析進行盲性評估是有價值的。這種對治療組別保持盲性的預先分析，應當包括以下一些問題的決定，例如，從分析群體中排除受試者或資料；檢視指標的可能轉換方式；以及定義離群值；將在其他最新研究中確定的重要共變數添加到統計模型中；可以重新考慮使用母數或無母數方法。此時所做的決定應當在試驗報告中加以描述，並且應當與統計人員在獲得隨機編碼之後做出的決定加以區別，因為在盲性下所做的決定通常會減少偏差的可能性。參與解盲性期中分析的統計人員或其他人員，不應參與盲性評估或修改統計分析計畫。數據中可能會顯示出治療所誘發的效應，當這種可能性會破壞盲性設計時，盲性評估時將需要特別謹慎。

許多更詳細的報告和列表方式應當在盲性評估時完成最終版，以便在實際分析資料時，整個計畫的各方面都已確立，包括受試者的選擇、數據的選擇與修改、數據的匯總與列表、參數估計及假說檢定。一旦完成資料的驗證(data validation)，應按照預定的計畫進行，愈能依循這些分析計畫執行，所得結論的可信度就愈高。當實際分析有別於試驗計畫書、計畫書變更以及基於數據進行盲性評估時所更新的統計分析計畫中所描述的計畫中(planned)的分析時，須特別注意。應對偏離計畫的統計分析提供詳盡的解釋。

所有進入臨床試驗的受試者，無論是否納入統計分析，都應在研究報告中呈現。排除在分析之外的所有原因都應被記錄，對於納入全分析群體但未納入依計畫書群體的受試者，或是排除在依計畫書群體的原因也應被記錄。同樣地，對於被納入分析集的所有受試者，在所有相關的時間點上的所有重要指標測量值都應紀錄。

應仔細考慮所有受試者的或數據的所有缺失、退出治療組別和重要違反計畫書等情況對主要指標分析的影響。受試者的失訪、退出治療或嚴重違反計畫書必須一一指出，並對其進行描述性分析，包括缺失的原因及其與治療和結果的關連。

描述性統計是報告中不可或缺的部分。應用適當的表格或圖形清晰地呈現主要和次要指標以及關鍵預後因子和人口統計學指標的重要特徵。應當在報告中特別仔細地描述與試驗目標有關的主要分析結果。當撰寫統計顯著性檢定結果時，應當寫出精確的 p 值(如 $p=0.034$)，而不是只依臨界值列出顯著或不顯著的結論。

儘管臨床試驗分析的主要目的應當是回答主要目標所提出的問題，但在解盲性分析過程中，基於觀察到的資料很可能會出現一些新問題。後果是，可能導致需要用額外且相對複雜的統計分析方法來處理。在報告中，應該嚴格區分這種額外的分析內容與計畫書中預定分析的內容。

對於在原計畫的分析中，未被預先定義為共變數，但仍然具有某些預後重要性的基礎值，由於隨機的作用，可能會導致它們在治療組之間出現無法預期的不平衡。處理這種不平衡的最佳方法是，用於解讀這些不平衡的補充分析得出了與原計畫的分析一致的結論來證明。如果不能得出一致的結論，則應討論這種不平衡對結果的影響。

一般而言，計畫外的分析應儘量少用。當認為治療效果可能隨某個或某些其他因素的改變而有不同時，常會進行這類計畫外的分析。這時可能是企圖找出治療效果特別好的次族群，然而過度解讀計畫外的次族群分析結果的潛在危險是眾所皆知的(見 5.7 節)，應謹慎避免。儘管類似的解讀問題也出現在治療無效或次族群有不良反應時，但我們應對其可能性做出適當的評估並加以報告。

最後，應根據臨床試驗結果的分析、解讀及呈現做出統計判斷。為此，試驗統計人員應該是負責臨床試驗報告的小組成員之一，並核准臨床試驗報告。

7.2. 臨床試驗資料的總結

藥品申請上市時，需要對所提供的臨床試驗之安全性和療效證據進行全面總結和綜合結論(如歐盟的專家報告、美國的整合報告和日本的概要)。某些情況，也會附上合併的統計分析結果。

在總結中，有一些特定的統計關注的領域：如描述在臨床試驗規劃期間接受試驗藥物之受試者群體的人口統計學和臨床特徵；根據相關(通常是有對照組的)試驗的結果，並強調它們互相一致或不一致的程度，以探討療效方面的關鍵問題；以及總結所有為上市申請的臨床試驗之安全性資訊，並找出潛在的安全問題。在臨床試驗設計規劃時，應注意量測值的統一定義和收集，這將有助於後續一系列試驗結果的解讀，特別是進行跨試驗合併分析時。應該選用可記錄用藥細節、病史及不良事件的通用詞典。對主要和次要指標採用共同的定義往往是有利的，並且它對統合分析是極其重要的。針對關鍵療效指標的測量方法、相對於隨機分派/進入試驗的評估時間、違反或偏離試驗計畫書的處理方法以及可能的預後因子的定義，都應該保持一致，除非有合理的理由不這麼做。

應詳細描述用於跨試驗間資料合併的任何統計程序。應注意因試驗選擇而導致偏差的可能性、試驗結果的同質性、以及各種變異來源下的適當模型。亦應探討分析結果對所做的假設和選擇的敏感性。

7.2.1. 療效資料

每一個臨床試驗的樣本數應該足夠多，以確保達到它們預期的目標。透過整合一系列本質上探討相同的重要療效問題的臨床試驗，也可以獲得額外有價值的資訊。這一系列試驗的主要結果，應當以統一的格式呈現以便於比較，通常是用表格或圖形的方式呈現估計值和信賴區間。使用統合分析的技術來合併這些估計值常常是一個有用的方法，因為它允許對整體治療效果提供一個更精準的估計，並提供一個完整而簡潔的試驗結果總結。在特殊情況下，統合分析也可能是通過整體假說檢定提供充分的整體療效證據之最適當或唯一的方式。當用於此目的時，統合分析應該有其預先規劃的專屬計畫書。

7.2.2. 安全性資料

在總結安全性資料時，應徹底檢查安全性資料庫以尋找潛在毒性的任何跡象，並通過尋找相關的支持性或觀察性分析來追蹤這些跡象。將所有暴露於新藥的受試者的安全性資料合併起來，可提供重要的資訊來源，因為其較大的樣本數提供了最佳機會以檢測出較罕見的不利事件，也許還可估計它們的約略發生率。然而，由於缺乏對照組，難以評估來自該資料庫的發生率數據；而來自有對照組試驗的資料在克服這種困難方面就顯得特別有價值。應合併具有相同對照組(安慰劑或特定的活性對照組)的試驗結果，並單獨呈現每個提供足夠數據的對照組的結果。

所有透過資料探索所發現的潛在毒性的跡象都應報告。在評估這些潛在不良反應時，應考慮到由於多次比較所產生的多重檢定問題。還應適當地使用存活分析方法，以探討不良事件的發生率與暴露時間或追蹤時間(follow-up)的潛在關係。並應適當地量化已辨識出的不良反應相關的風險，以便正確地評估風險與利益關係。

詞彙表

貝氏方法(Bayesian Approaches)

一種資料分析的方法，針對某些參數(如治療效果)，由觀察資料以及先驗機率分布所推得出的後驗機率分布作為統計推論之基礎。

偏差(統計上和操作上的)(Bias [Statistical and Operational])

與臨床試驗設計、執行、分析和結果解讀相關的任何因素，使得治療效果估計值偏離真實值的系統性傾向。由於執行上的偏離所造成的偏差稱為操作上的偏差，其他原因造成的偏差則稱為統計上的偏差。

盲性評估(Blind Review)

在試驗完成(最後一例病人的最後一次觀察)與解盲之間對資料進行查核和評估，目的是為了最終確定計畫中的分析。

內容效度(Content Validity)

一個指標(如評分尺度指標)測量它應該測量的內容的程度。

雙虛擬(Double-Dummy)

在臨床試驗中當兩組治療不能做到一樣時，使提供的藥物仍能維持盲性的一種技術。所準備的藥物必須為活性藥物 A 及其相同無法分辨之安慰劑、活性藥物 B 及其相同無法分辨之安慰劑，受試者將接受 A、B 兩種藥物：活性藥物 A 及安慰劑 B，或是安慰劑 A 和活性藥物 B。

退出(Dropout)

臨床試驗中的受試者，由於任何原因不能繼續進行試驗到依計畫所規定的最後一次回診。

相等性試驗(Equivalence Trial)

一個試驗的主要目的是要顯示兩種或多種治療的差異並無臨床上重要的意義。這通常以真

實的治療組間差異落在臨床上可接受的相等性臨界值上下限之間來推論相等性。

頻率學派方法(Frequentist Methods)

在假設重複實現相同試驗情境時，用某些結果發生的頻率來解釋的統計方法，如顯著性檢定和信賴區間。

全分析群體(Full Analysis Set)

盡可能接近按意圖治療原則的理想病例集。由所有接受隨機分派的受試者，以最少且合理的方法排除病例得出。

外推性(Generalisability, Generalisation)

一個臨床試驗的結果可以被可信地由參加試驗的受試者族群，外推到更廣大的病人族群和更廣泛的臨床環境的程度。

整體評估指標(Global Assessment Variable)

單一指標，通常是把客觀的指標和研究者對受試者的狀況或者狀態的改變情形整合起來的順序類別評分量表。

獨立數據監測委員會(資料和安全性監測小組、監測委員會、數據監測委員會)(Independent Data Monitoring Committee [IDMC], Data and Safety Monitoring Board, Monitoring Committee, Data Monitoring Committee)

一個獨立的數據監測委員會可以是由廠商成立，定期評估試驗進度、安全性數據與重要的療效指標，並可以向廠商提出建議是否繼續、修正或停止試驗。

意圖治療原則(Intention-To-Treat Principle)

一種依安排好的治療方案而不是依據實際接受的治療為基礎之評估方式，這是可以對治療效果作出最好評估的原則。其結果是分到任一組別的受試者即應作為該組的成員接受訪視、評估與分析，而不論他們是否遵從計畫中的治療。

交互作用(定性和定量)(Interaction-Qualitative and Quantitative)

一種治療差異(如研究產品與對照組的差異)會由另一個因素(如試驗中心)而決定的情況。定量交互作用指的是治療差異的大小在因素的不同水平時(如不同試驗中心)會有不同，而定性交互作用指的是治療差異的方向性至少在因素的某一水平上(如某一試驗中心)不同。

評估者間信度(Inter-Rater Reliability)

不同評估者在不同情況下產生相同結果的性質。

評估者內信度(Intra-Rater Reliability)

同一評估者在不同情況下產生相同結果的性質。

期中分析(Interim Analysis)

在試驗正式結束前，於任何時間點為了比較治療組間的療效或安全性結果，所執行的任何分析。

統合分析(Meta-Analysis)

對同一個問題的兩個或更多試驗的定量證據進行正式評估。這常常是從各試驗的摘要統計量進行統計上的合併分析，但此名詞有時也用於原始數據的合併分析。

多中心試驗(Multicenter trial)

按照同一個計畫書在多個試驗中心所進行的臨床試驗。因而，由多個試驗主持人執行。

不劣性試驗(Non-inferiority trial)

主要目的為顯示研究產品的反應在臨床上不劣於對照組(活性或安慰劑對照)的試驗。

選用詞與納入詞(Preferred and Included Terms)

在一個多層次的醫學名詞分類法中，例如 MedDRA，納入詞(included terms)是最低階的醫學名詞，直接以研究者對不良事件的描述詞進行編碼。選用詞(preferred terms)是對納入詞進行分類與統整，用於通報事件發生的頻率。例如，研究者寫的是「左臂疼痛」，納入詞編碼為「關節疼痛(Joint pain)」，在選用詞可通報為「關節痛(Arthralgia)」。

依計畫書群體(有效病例，有效樣本，可評估受試者樣本)(Per Protocol Set [Valid Cases, Efficacy Sample, Evaluable Subjects Sample])

由能充分遵從計畫書以確保這些數據會按所基於的科學模型而表現出治療效果之病例子集所產生的數據集。順從性包括以下的考量，如使用過治療藥物、有可用的測量值，以及沒有重大違反計畫書事件。

安全性和耐受性(Safety and Tolerability)

藥物的安全性關係到病人醫療風險時，試驗中通常以實驗室檢查(包括臨床化學與血液學)、

生命徵象、不良事件(疾病、徵象和症狀)，以及其他專門的安全性檢查(例如心電圖、眼科檢查)等來評估。藥物的耐受性代表了病人所能忍受明顯不良反應的程度。

統計分析計畫(Statistical analysis plan)

統計分析計畫是一份文件，其中包含對計畫書中所描述統計分析的主要特徵進行更技術性的詳細闡述，並包含執行主要和次要指標及其他數據統計分析的詳細過程。

較優性試驗(Superiority Trial)

主要目的是顯示研究產品的反應優於對照組(活性或安慰劑對照)的試驗。

替代指標(Surrogate Variable)

當直接測量臨床效果是不可能或不實際時，所提供的間接測量的指標。

治療效果(Treatment Effect)

在臨床試驗中歸因於治療所產生的效果。在大多數臨床試驗中感興趣的治療效果是兩個或多個治療組間的比較。

治療後發生的事件(Treatment Emergent)

用藥前沒有，但在試驗用藥期間出現的事件；或是用藥前已有，但用藥期間更加惡化的事件，都稱為治療後發生的事件。

試驗統計人員(Trial statistician)

經過教育、培訓並有足夠經驗以貫徹本指導原則，並負責試驗統計方面的統計學家。