研究數據的嚴謹性

數據的收集，管理與分享

“數據” 的定義

- 紀錄下來的資訊；包括文字、影片、錄音、複製的圖片、圖形、圖樣、或是其他以圖形表達的東西，操作手冊、表格、圖示、工作流程圖、設備儀器的描述、數據檔案、資料處理、或電腦程式(軟體)、統計紀錄、與其他研究數據。
  - 〔摘自美國國家衛生研究院（NIH）補助金政策說明，“對數據的權利”〕

數據嚴謹性的一些問題

- 穫取（acquisition）與紀錄（record keeping）
- 處理（processing）
- 所有權（ownership）及控制（control）
- 保留（retention）及儲存（storage）
- 使用（access）及分享（sharing）

數據嚴謹性相關的政策及指導方針

- 聯邦政府（例如：公共衛生服務處PHS，食物及藥物管理局FDA）
- 機關團體（例如：麻省理工學院）
- 研究室或是研究團體
  - （譯註：不同的機構可能有不同的政策及指導方針）

數據的穫取與紀錄

紀錄

- 食物及藥物管理局的標準（生物製品、藥物、儀器等最後核准的標準）
  - 在基礎科學方面：”良好的實驗室慣例”(GLP)
  - 在臨床實驗方面：”良好的臨床實驗慣例”(GCP)
  - 設計、進行、紀錄及報告研究結果的標準－被核准使用的數據必須有”最高度的品質與嚴謹性”〔食物及藥物管理局FDA〕

研究中數據的紀錄

- 食物及藥物管理局的“良好的實驗室慣例”GLP：
  - 應該即刻紀錄所有數據
  - 每筆數據必須有紀錄日期與簽名
  - 任何更改不應該使得原來的數據含混不清
  - 更改的原因一定要有清楚的記載、日期、簽名

實驗室中的筆記簿

- 紀錄什麼呢？
  - 方法
  - 結果／觀察
  - 想法
  - 相關電腦與文件檔案的所在地
  - 規程的修改
  - 實驗室中的會議記錄
  - 每個階段的參與者

實驗室中的筆記簿

- 如何紀錄呢？
  - 一有結果馬上紀錄
  - 按照時間先後順序紀錄
  - 每一筆紀錄都清楚，易讀

實驗室中的筆記簿
－該做的和不該做的

- 該做的：
  - 要用裝訂成冊的筆記簿，並且每一頁都有按順序編的頁碼
  - 用黑色的原子筆寫（抗潮、抗溶解液）
  - 每一筆數據都有日期與簽名
  - 以畫一槓來更正錯誤，寫上名字縮寫與更正理由

實驗室中的筆記簿
—該做的和不該做的

- 該做的：
  - 用膠帶把附加文件黏貼在筆記簿上，並在膠帶上寫上名字縮寫與日期
  - 保留筆記簿

實驗室中的筆記簿
—該做的和不該做的

- 不該做的：
  - 擦掉資料
  - 空白的地方，沒有畫一槓、日期、與簽名
  - 撕頁

臨床實驗的數據的紀錄

- 食物及藥物管理局＂良好的實驗室慣例＂(GLP)
- 維護個案的歷史紀錄
  - 個案報告的表格
  - 相關文件
- 確保數據的準確性、完整性、易讀性、及時效性
- 更改個案的歷史紀錄時，不可以使得原來的數據含混不清（包括手寫的、和電腦文件）

聯邦公報〈Federal Register〉，第九冊，第159號

臨床實驗的數據的紀錄

- 個案的歷史紀錄應包括：
  - 實驗對象的基本資料
  - 實驗對象符合參與實驗的條件的資料
  - 治療的資料
  - 檢查和測驗的資料
    - 測驗結果
    - X光
    - 體檢等
- 影本、光碟、或電腦存檔均可以是紀錄的形式

用電腦來紀錄

電腦化的數據收集系統必須：

- 只准授權人員輸入數據
- 有控制刪除或修改數據的能力
- 提供稽核蹤跡
  - 誰做修改
  - 什麼時候
  - 爲什麼
- 防止擅改
- 確保數據保留（備分及還原）
- 有列印報表的系統

數據處理（Data Processing）

數據處理

- 儘量做下列兩點
  - 完整性
  - 準確性
- 確認：
  - 數據輸入正確
  - 每一個變數值都在合理範圍內
  - 沒有遺漏的數據（或是找出並處理有遺漏數據的個案）

數據的輸入（entry）

- 使用雙登錄系統（double entry）
  - 先將觀察的數據輸入
  - 再重新輸入以驗證
- 直接從實驗設備下載
- 輸入後再做檢查

超出範圍的數值（Out-of-Range Values）

- 離群值（outlier），是指和數據組的其他數據不一致的觀察數據
- 找出單一變數的離群值
  - 每一個變數的最大及最小值是不是如你所預期？
  - 數據的分佈（distribution），有沒有在分佈線尾巴（tails）的外面？
- 如果可能，找出離群值的原因
- 在做統計分析之前要先決定如何處理離群值

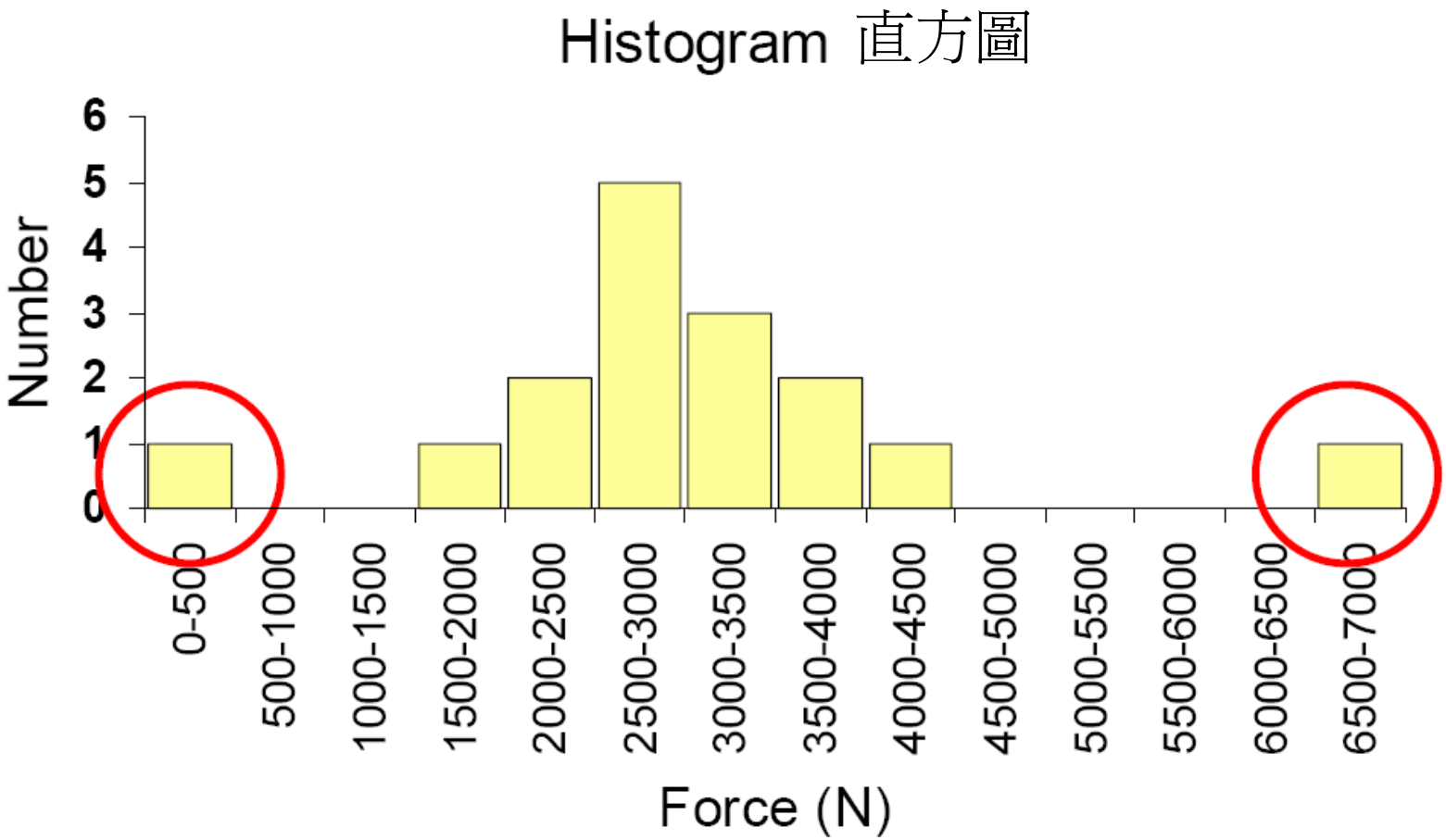
可能產生離群值的原因

- 在數據的取得、紀錄、或輸入時發生錯誤
- 有極端值的個案並不是從計劃的母體中取樣
- 極端的(真的)生物上或是環境上的變化所導致的極端值

例如

案例	F(N)
1	20
2	3400
3	4500
4	2010
…	…
16	7000

例如


Histogram 直方圖
Number
0
1
2
3
4
5
6
0-500
500-1000
1000-1500
1500-2000
2000-2500
2500-3000
3000-3500
3500-4000
4000-4500
4500-5000
5000-5500
5500-6000
6000-6500
6500-7000
Force (N)

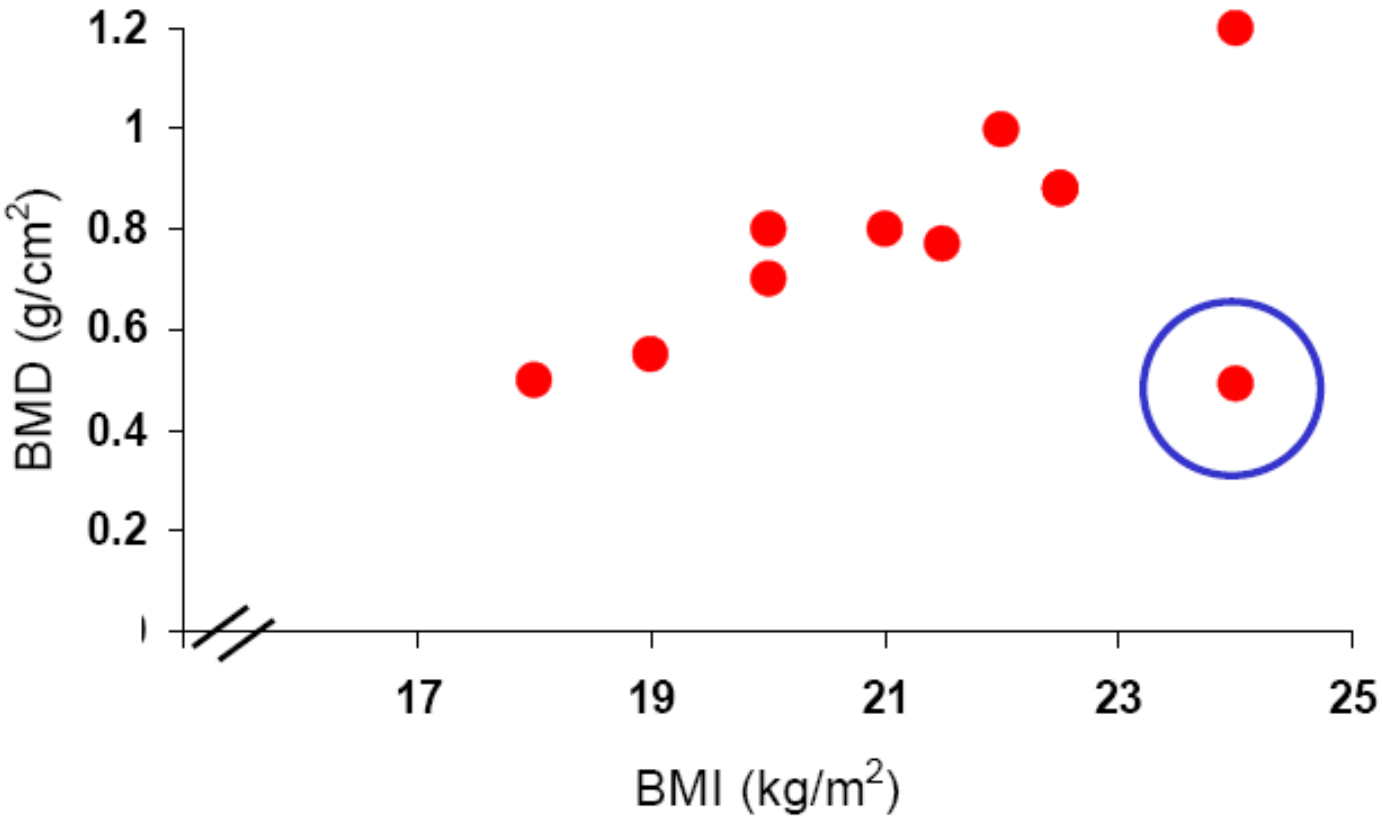
例如

- 當與原始數據對照時，發現案例1的值應該是2000 N→資料輸入有錯
- 和原始個案的描述比對之後，發現案例16號是男性。計劃的樣本群是女性Æ 案例不是從計劃的樣本群中取得

幾個變數在一起的離群值

- 不尋常的組合兩個以上的測量
- 可以用散佈圖（scatter plots），來找尋連續變數中不尋常的走勢（pattern）
- 統計的多變量分析（multivariate methods）也可找出離群值

多變量分析找出離群值的例子



若有離群值怎麼辦？

- 來源決定策略：
- 不正確數據的輸入
  - 以正確的數據取代之
- 個案不是從計劃的樣本群中取得
  - 如果你知道是你取樣不小心的錯誤，刪除個案
  - 如果你只是懷疑，做兩次統計分析，一次包括有問題的個案，一次不包括，然後比較兩次的結果

若有離群值怎麼辦？

- 極端的差異
  - 做兩次統計分析，一次包括有問題的個案，一次不包括，然後比較兩次的結果
  - 在分析的過程時，用技巧來調整有偏差（skewed）的數據
  - 接受極端值及因其造成的任何偏差

遺漏的數據？

- 找出來並做上記號
- 如果可能，找出原因
  - 例如：病人退出，機器有問題
- 有多少遺漏的數據
- 尋找重複模式－應該是隨機的
- 在做分析之前，就先決定如何處理遺漏的數據

遺漏的數據 - 怎麼辦？

- 有遺漏數據的變數不被列入計算
- 有遺漏數據的個案不被列入計算
- 接受數目不齊全的觀察數據
- 對重復試驗（repeated-measures）的設計要特別小心

骨質密度（g/cm2）

個案編號	基線（baseline）	第一年	第二年
1	112	112	119
2	097		101
3	086	088	099

遺漏的數據 - 怎麼辦？

- 將遺漏的數據當做數據
  - 缺乏數據的本身可能成爲你研究結果的預測變數（譯註：例如在一項癌症藥物的研究，觀察病人在服用藥物後的生存率(survival analysis）。研究期限是五年，五年當中，病人在服用藥物後的存活年數是所要收集的數據。五年到期，或有病人仍然活著，於是這個案例就有存活年數這個遺漏的數據－。這個遺漏的數據可以當作一種資訊：病人在服用藥物後的存活年數可能大於五年。)

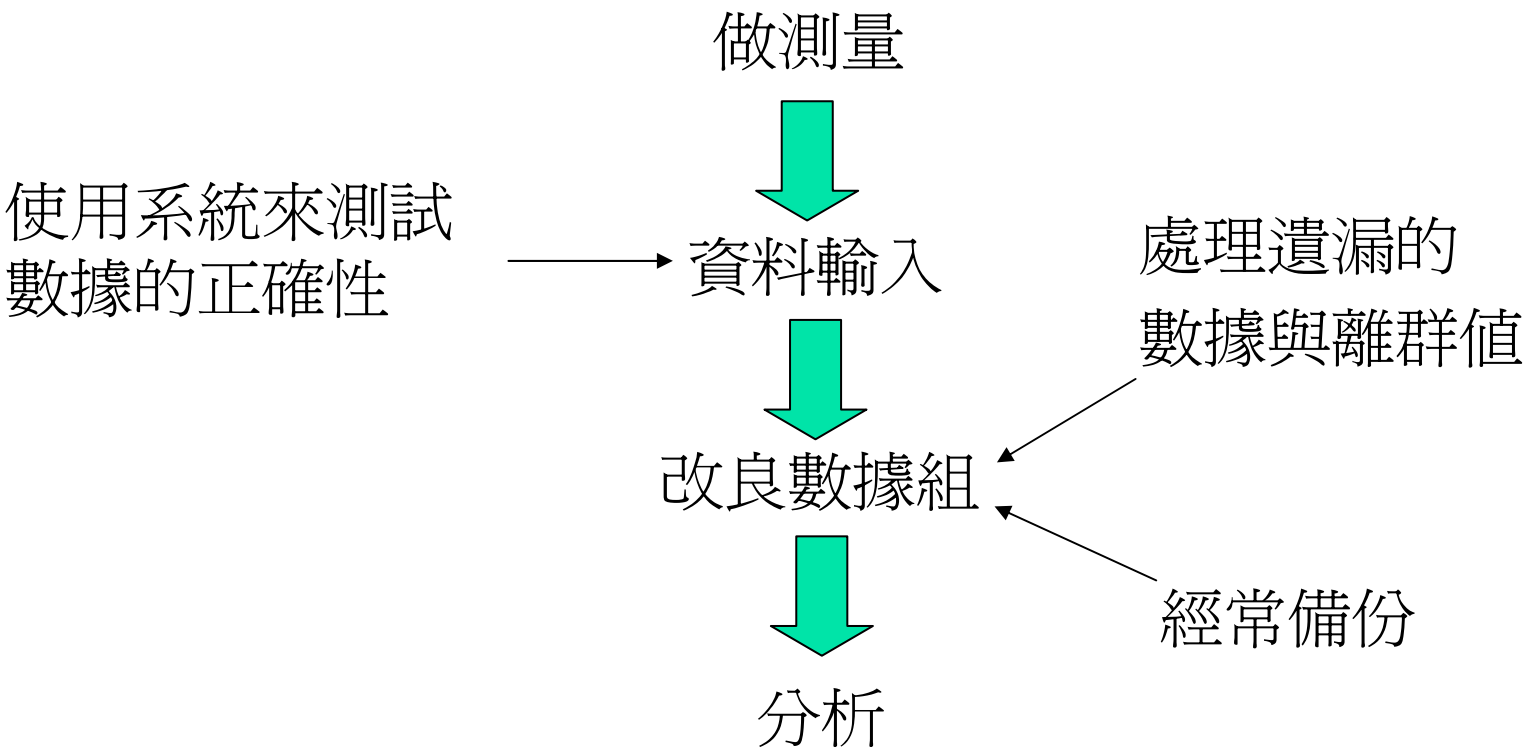
（下續）

遺漏的數據 - 怎麼辦？

- 爲有遺漏的數據的變數建立新的虛擬變數，在分析時將其視同變數來處理
  - 完全＝0　遺漏＝1
- （譯註：持續上面的例子，將同樣的概念，用下列表格說明。所以在分析的時候，新的虛擬變數，就可以當一個變數來處理了。）

個案	存活年數	虛擬變數
1	2	0
2	3	0
3		1
4	2	0
5	4	0

資料處理—總結



資料處理—負責任的行爲

- 改良數據組可能需要刪除或是修改
- 交代刪除或是修改數據的理由

數據的所有權（ownership）與監護權（custody）

數據的所有權

- 數據的所有權的標準與政策，決定於計畫的資助來源
  - 美國衛生署PHS：接受美國衛生署資助的機構擁有數據所有權
  - 接受資助的研究：研究合約決定數據的所有權及支配權
  - 沒有接受資助的研究：一般而言，研究單位持有所有權

數據的所有權—
美國衛生署（PHS）的政策

- “一般而言，接受資助的機構擁有在資助款維持的計劃內產生的研究數據”

〔摘自美國國家衛生研究院資助款政策說明，“對數據的權利”〕

數據的支配—受資助的研究

- 數據的支配通常在受資助的研究合約上”出版物”的部分
- 例子：
  - 受助機構有發表研究結果的自由，但要事先給資助單位足夠的時間審查：
    - 資料是否牽涉到專利（patent）
    - 不小心洩漏了私權資產（proprietary information）

數據的監護權

- 監護權一般而言是在研究主持人（PI），或是實驗室（代表所屬機構）的手中

研究人員離職

- 離職的規定因研究機構不同而異
- 例如：
  - 如果計劃主持人離職
    - 計劃主持人可以移走原始數據，將來也允許繼續取用
    - 計劃主持人可以帶拷貝
  - 如果是其他研究人員離職
    - 原始數據由計劃的主持人保留，提供拷貝
    - 原始數據由計劃的主持人保留，准許取用

數據的的保留及儲存

- 原始數據紀錄必須保留多久？
- 美國衛生署（PHS）：計劃結束後爲期三年〔45 CFR Part 45〕＊
- 美國食物及藥物管理局（FDA）：上市後爲期兩年〔21 CFR Part 48〕
- 研究機構：研究資料發表之後爲期不一，從兩年至無盡期

＊譯註：美國政府法規及公報之資料庫

Code Of Federal Regulations http://www.legalbooksdepot.com/

數據的的保留及儲存

- 原始數據紀錄應該儲存在什麼地方？
- 一般定沒有明文規定
- 例如：
  - 在實驗室裡
  - 在研究機構的辦公室裡

數據的的保留及儲存

- 原始數據紀錄應該儲存在什麼樣的環境下？
- 理想的環境
  - 溫度及濕度都在控制下
  - 防範自然災害及偷竊
  - 有限制、卻便利的取用
- 現實情況
  - 檔案櫃

文件的儲存

- 一般機構沒有明文規定
- 建議：保持可追蹤的目錄來記載：
  - 有識別號碼的數據種類（筆記本，錄音帶，Ｘ光片等）
  - 計劃名稱
  - 研究人員
  - 數據紀錄或筆記本完成的日期
  - 儲存的地方

數據的分享（Data Sharing）

數據的取用及分享

- 數據交換的目標之潛在的衝突
  - 研究員有權利因數據而穫得榮譽（經由分析、在學術會議上發表、出版）

    相對於
  - 研究員有義務將數據與同儕分享

數據的取用及分享

- 我們爲什麼需要取用？
  - 資訊的公開與交換是科學的基本宗旨
  - 在調查不當行爲的指控時，取用可能有必要
  - 可能根據資訊公開法，而請求取用權（譯註：美國司法部　資訊公開法 http://wwwusdojgov/04foia/ ）
  - 美國衛生署資助的研究計劃均要求取用權

ＭＩＴ對數據分享的政策說明

- “及時與公開的傳播ＭＩＴ的研究結果以及在學者之間自由的做資訊交換，是ＭＩＴ在履行對卓越的教育及研究的承諾上很重要的一環”

ＭＩＴ科技的所有權、分佈、及商業發展的指南

美國國家衛生研究院對數據分享的說明

- 美國國家衛生研究院期望及時發放與分享最後的研究數據給其他研究者的使用
- 美國國家衛生研究院要求申請者包括分享數據的計劃，或是說明爲什麼不能分享數據（適用於超過某金額的計劃）

美國國家衛生研究院的說明：為什麼分享？

- 延伸美國國家衛生研究院在分享研究資源上的政策
- 加強公開的科學質詢
- 鼓勵不同的分析及意見
- 促進新的研究
- 支持新的或是不同假設的試驗，與分析方法

爲什麼分享？

- 幫助教育新的研究者
- 促使最初的研究者並未想到的題目的探討
- 容許從數據合併所產生的新數據組

什麼樣的資訊應該被分享？

- 驗証研究結果所必需的最後研究數據
- 不包括：
  - 實驗室的筆記簿
  - 不完整的數據組
  - 初步的分析
  - 科學報告的草稿
  - 未來的研究計劃
  - 同儕間的溝通
  - 實體的東西，例如凝膠或實驗室標本等

適用於什麼樣的研究？

- 在美國國家衛生研究院資助下所產生的數據：
  - 基礎研究
  - 臨床實驗
  - 問卷
  - 其他種類的研究
- 特別重要的分享：
  - 獨特又不易複製的數據組
  - 龐大、昂貴的數據組

使用以人爲對象的實驗的警告

- 實驗者必須小心：
  - 樣本數目很小的研究
  - 收集敏感的數據的研究
- 然而，只要有安全措施以確保機密性、及防備透露參與者身分，即使這些資料也可以分享的

什麼叫做及時？

- 在最終數據組的主要發現被接受發表之前或同時

結論（Summary）

結論

- 負責任的數據管理的政策不停的在改變
- 一些目前的看法
  - 確保精確與可信的數據組是研究者的責任，包括：
    - 主要來源（例如：筆記本）
    - 改良過的數據組

結論

- 研究機構常擁有數據
- 計劃主持人或是實驗室常擁有數據監護權
- 原始資料應該保存三年以上
- 美國衛生署要求受其資助的研究的數據的取用權(在結果發表後)，或許這也是其他種類的支持的要求
- 公開傳播研究結果是研究界的根本宗旨

參考資料

- 美國國家統計委員會，美國國家研究院《分享研究成果》美國國家研究院出版 http://wwwnapedu/books/030903499X/html/indexhtml
- MIT TLO（Technology Licensing Office）的政策 http://webmitedu/tlo/www/guide2html
- 美國太空總署（NASA）確保資訊品質的指導方針 ftp://ftphqnasagov/pub/pao/reports/2002/NASA_data_quality_guidelinespdf
- 美國國家衛生研究院數據分享的網路主頁 http://grantsnihgov/grants/policy/data_sharing/