研究数据的严谨性

数据的收集，管理与分享

“数据 ”的定义

- 纪录下来的信息；包括文字、影片、录音、复制的图片、图形、图样、或是其它以图形表达的东西，操作手册、表格、图标、工作流程图、设备仪器的描述、数据档案、数据处理、或计算机程序(软件)、统计纪录、与其它研究数据 。
  - [摘自美国国家卫生研究院（NIH）补助金政策说明，“ 对数据的权利”]

数据严谨性的一些问题

- 获取（acquisition）与纪录（record keeping）
- 处理（processing）
- 所有权（ownership）及控制（control）
- 保留（retention）及储存（storage）
- 使用（access）及分享（sharing）

数据严谨性相关的政策及指导方针

- 联邦政府（例如：公共卫生服务处PHS，食物及药物管理局FDA）
- 机关团体（例如：麻省理工学院）
- 研究室或是研究团体

  - （译注：不同的机构可能有不同的政策及指导方针）

数据的获取与纪录

纪录

- 食物及药物管理局的标准（生物制品、药物、仪器等最后核准的标准）
  - 在基础科学方面：＂良好的实验室惯例＂(GLP)
  - 在临床实验方面：＂良好的临床实验惯例＂(GCP)
  - 设计、进行、纪录及报告研究结果的标准—被核准使用的数据必须有＂最高度的品质与严谨性＂［食物及药物管理局FDA］

研究中数据的纪录

- 食物及药物管理局的＂良好的实验室惯例＂GLP：
  - 应该即刻纪录所有数据
  - 每笔数据必须有纪录日期与签名
  - 任何更改不应该使得原来的数据含混不清
  - 更改的原因一定要有清楚的记载、日期、签名

实验室中的笔记簿

- 纪录什么呢？
  - 方法
  - 结果 / 观察
  - 想法
  - 相关计算机与文件档案的所在地
  - 规程的修改
  - 实验室中的会议记录
  - 每个阶段的参与者

实验室中的笔记簿

如何纪录呢？

- 一有结果马上纪录
- 按照时间先后顺序纪录
- 每一笔纪录都清楚，易读

实验室中的笔记簿
—该做的和不该做的

该做的：

- 要用装订成册的笔记簿，并且每一页都有按顺序编的页码
- 用黑色的原子笔写（抗潮、抗溶解液）
- 每一笔数据都有日期与签名
- 以画一杠来更正错误，写上名字缩写与更正理由

实验室中的笔记簿
—该做的和不该做的

- 该做的：
  - 用胶带把附加文件黏贴在笔记簿上，并在胶带上写上名字缩写与日期
  - 保留笔记簿

实验室中的笔记簿
—该做的和不该做的

- 不该做的：
  - 擦掉资料
  - 空白的地方，没有画一杠、日期、与签名
  - 撕页

临床实验的数据的纪录

- 食物及药物管理局＂良好的实验室惯例＂(GLP)
- 维护个案的历史纪录
  - 个案报告的表格
  - 相关文件
- 确保数据的准确性、完整性、易读性、及时效性
- 更改个案的历史纪录时，不可以使得原来的数据含混不清（包括手写的、和计算机文件）

联邦公报〈Federal Register〉，第九册，第159号

临床实验的数据的纪录

- 个案的历史纪录应包括：
  - 实验对象的基本数据
  - 实验对象符合参与实验的条件的资料
  - 治疗的资料
  - 检查和测验的资料
    - 测验结果
    - X光
    - 体检等
- 影本、光盘、或计算机存盘均可以是纪录的形式

用计算机来纪录

计算机化的数据收集系统必须：

- 只准授权人员输入数据
- 有控制删除或修改数据的能力
- 提供稽核踪迹
  - 谁做修改
  - 什么时候
  - 为什么
- 防止擅改
- 确保数据保留（备分及还原）
- 有打印报表的系统

数据处理（Data Processing）

数据处理

- 尽量做下列两点
  - 完整性
  - 准确性
- 确认：
  - 数据输入正确
  - 每一个变量值都在合理范围内
  - 没有遗漏的数据（或是找出并处理有遗漏数据的个案）

数据的输入（entry）

- 使用双登录系统（double entry）
  - 先将观察的数据输入
  - 再重新输入以验证
- 直接从实验设备下载
- 输入后再做检查

超出范围的数值（Out-of-Range Values）

- 离群值（outlier），是指和数据组的其它数据不一致的观察数据
- 找出单一变量的离群值
  - 每一个变量的最大及最小值是不是如你所预期？
  - 数据的分布（distribution），有没有在分布线尾巴（tails）的外面？
- 如果可能，找出离群值的原因
- 在做统计分析之前要先决定如何处理离群值

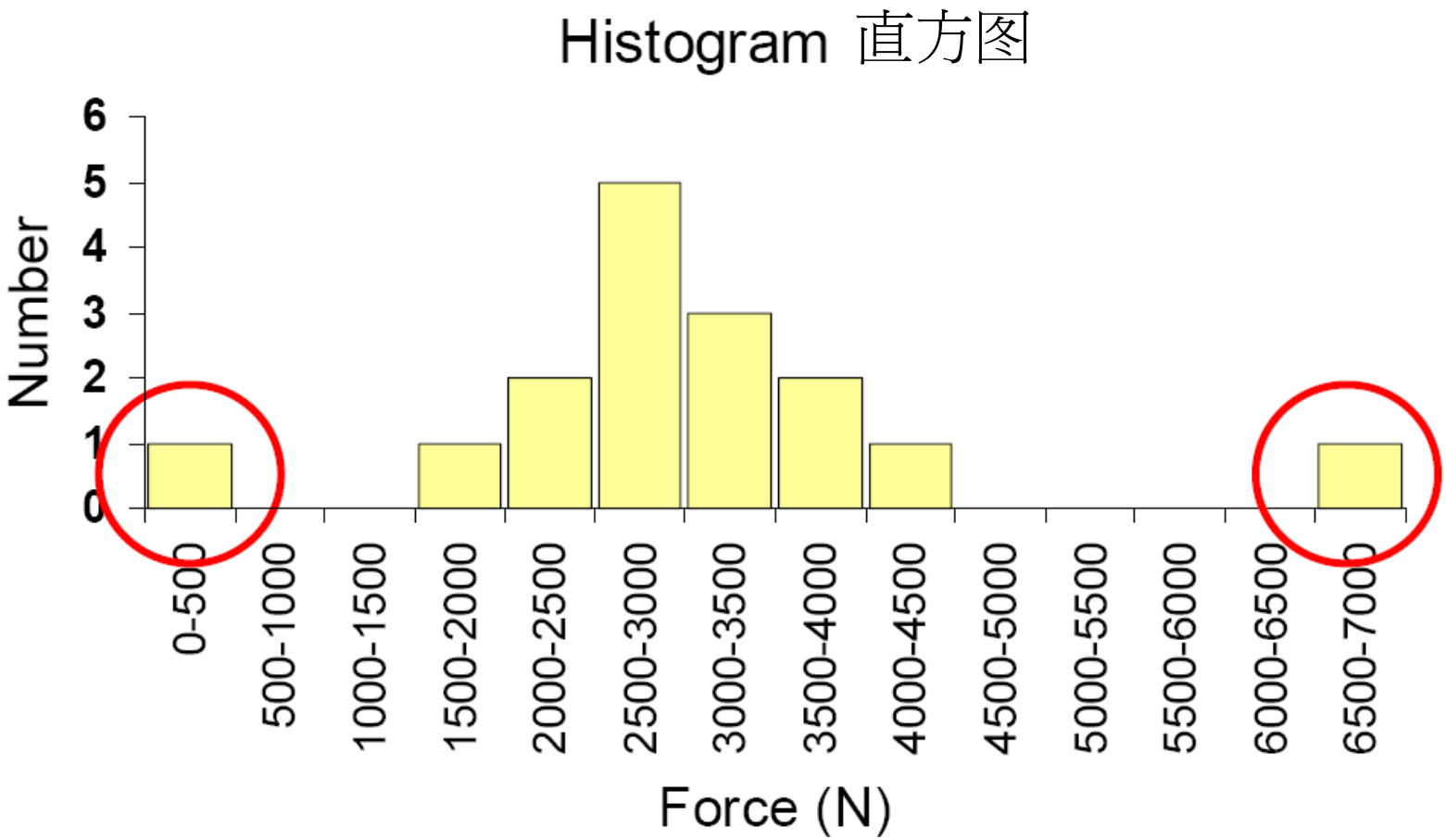
可能产生离群值的原因

- 在数据的取得、纪录、或输入时发生错误
- 有极端值的个案并不是从计划的母体中取样
- 极端的(真的)生物上或是环境上的变化所导致的极端值

例如

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

例如


Histogram 直方图
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)

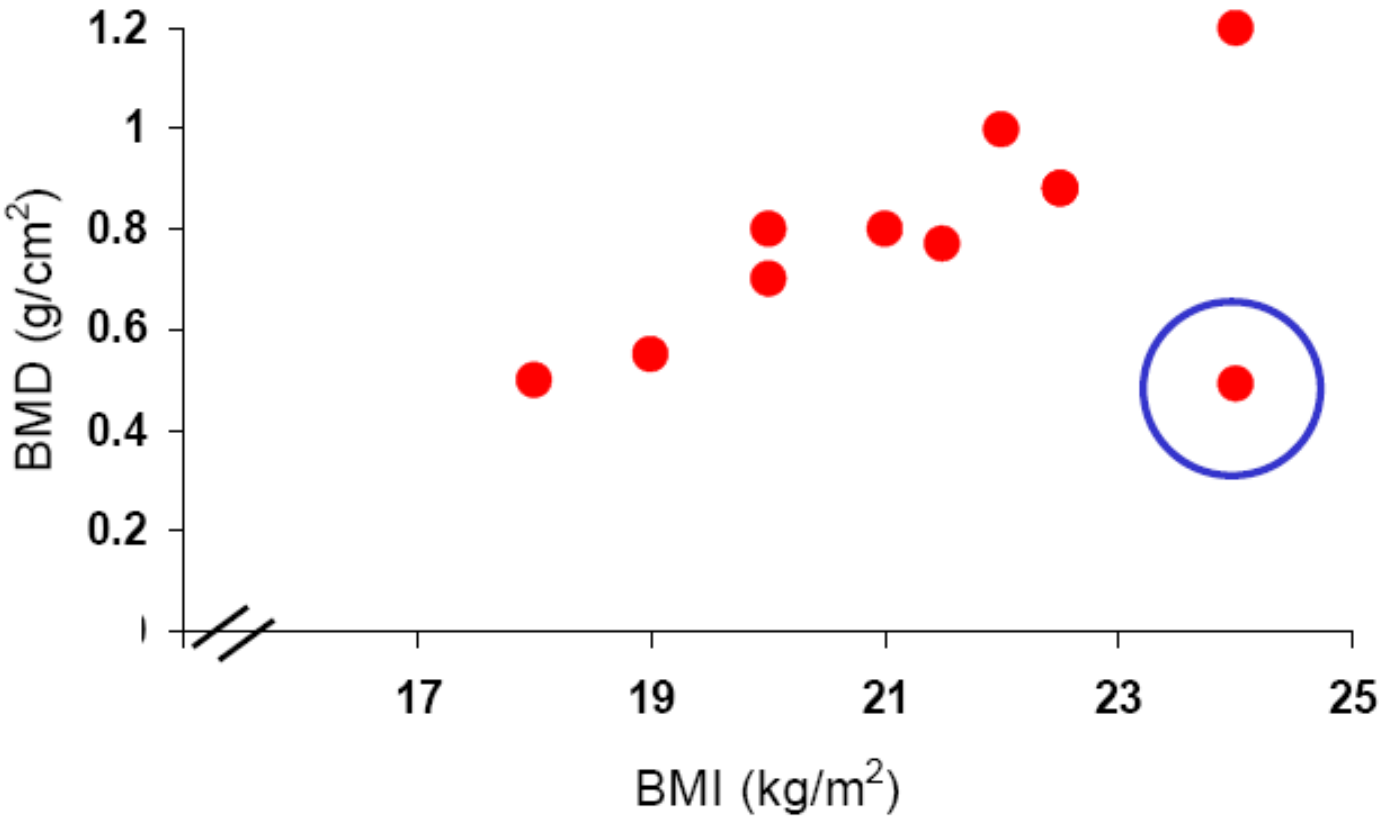
例如

- 当与原始数据对照时，发现案例1的值应该是2000N ->数据输入有错
- 和原始个案的描述比对之后，发现案例16号是男性。计划的样本群是女性Æ案例不是从计划的样本群中取得

几个变数在一起的离群值

- 不寻常的组合两个以上的测量
- 可以用散布图（scatter plots），来找寻连续变量中不寻常的走势（pattern）
- 统计的多变量分析（multivariate methods）也可找出离群值

多变量分析找出离群值的例子



若有离群值怎么办？

- 来源决定策略：
- 不正确数据的输入
  - 以正确的数据取代之
- 个案不是从计划的样本群中取得
  - 如果你知道是你取样不小心的错误，删除个案
  - 如果你只是怀疑，做两次统计分析，一次包括有问题的个案，一次不包括，然后比较两次的结果

若有离群值怎么办？

- 极端的差异
  - 做两次统计分析，一次包括有问题的个案，一次不包括，然后比较两次的结果
  - 在分析的过程时，用技巧来调整有偏差（skewed）的数据
  - 接受极端值及因其造成的任何偏差

遗漏的数据？

- 找出来并做上记号
- 如果可能，找出原因
  - 例如：病人退出，机器有问题
- 有多少遗漏的数据
- 寻找重复模式—应该是随机的
- 在做分析之前，就先决定如何处理遗漏的数据

遗漏的数据 - 怎么办?

- 有遗漏数据的变量不被列入计算
- 有遗漏数据的个案不被列入计算
- 接受数目不齐全的观察数据
- 对重复试验（repeated-measures）的设计要特别小心

骨质密度 （g/cm2）

个案编号	基线 （baseline）	第一年	第二年
1	112	112	119
2	097		101
3	086	088	099

遗漏的数据 - 怎么办?

- 计算遗漏的数据（要小心）
  - 用其它数据的平均值
    - 可能的问题：低估新数据组的变异性（variability）
  - 依某种顺序排列数据，然后采用遗漏数据之前的数值
  - 用遗漏的数据与其它变量的关系，来估计遗漏数据的值（不要用假设检定中的变量）
  - （译注：为什么呢？例如，你的实验假设，血压和骨质密度是有关联的。血压在这里是你的假设检定（hypothesis testing）变量。如果骨质密度有遗漏的数据，而你用血压来估计骨质密度的值，你就是承认血压和骨质密度是有关联的，那么就违反了实验的目的，因为你的目的是要证明两者的关系）

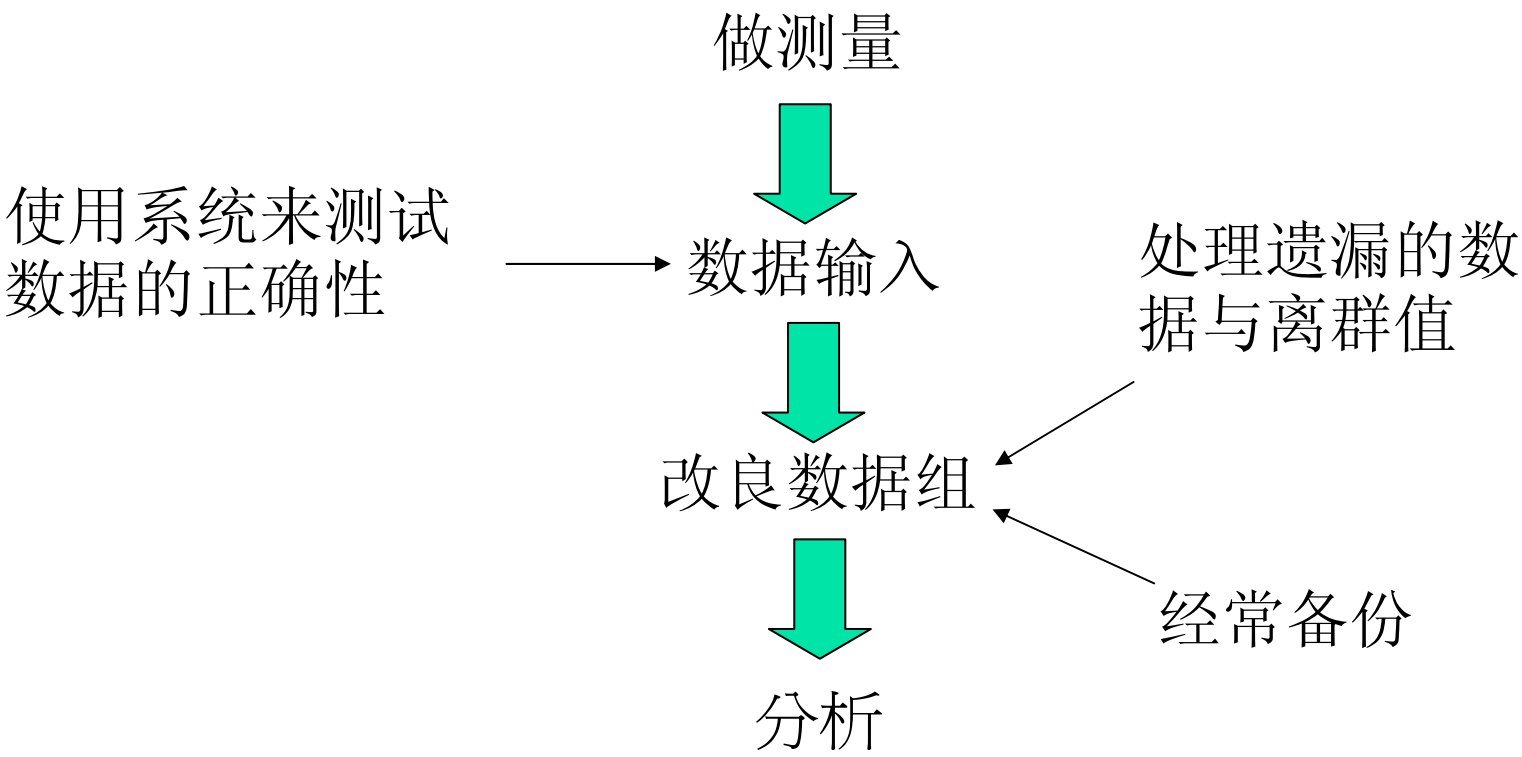
（下续）

遗漏的数据 - 怎么办?

- 为有遗漏的数据的变量建立新的虚拟变量，在分析时将其视同变量来处理
  - 完全＝ 0　遗漏＝ 1
- （译注：持续上面的例子，将同样的概念，用下列表格说明。所以在分析的时候，新的虚拟变量，就可以当一个变量来处理了。）

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

数据处理—总结



数据处理—负责任的行为

- 改良数据组可能需要删除或是修改
- 交代删除或是修改数据的理由

数据的所有权（ownership）与监护权（custody）

数据的所有权

- 数据的所有权的标准与政策，决定于计划的资助来源
  - 美国卫生署PHS：接受美国卫生署资助的机构拥有数据所有权
  - 接受资助的研究：研究合约决定数据的所有权及支配权
  - 没有接受资助的研究：一般而言，研究单位持有所有权

数据的所有权
—美国卫生署（PHS）的政策

- “一般而言，接受资助的机构拥有在资助款维持的计划内产生的研究数据”

〔摘自美国国家卫生研究院资助款政策说明，"对数据的权利"〕

数据的支配—受资助的研究

- 数据的支配通常在受资助的研究合约上＂出版物＂的部分
- 例子：
  - 受助机构有发表研究结果的自由，但要事先给资助单位足够的时间审查：
    - 资料是否牵涉到专利（patent）
    - 不小心泄漏了私权资产（proprietary information）

数据的监护权

- 监护权一般而言是在研究主持人（PI），或是实验室（代表所属机构）的手中

研究人员离职

- 离职的规定因研究机构不同而异
- 例如：
  - 如果计划主持人离职
    - 计划主持人可以移走原始数据，将来也允许继续取用
    - 计划主持人可以带拷贝
  - 如果是其它研究人员离职
    - 原始数据由计划的主持人保留，提供拷贝
    - 原始数据由计划的主持人保留，准许取用

数据的的保留及储存

- 原始数据纪录必须保留多久？
- 美国卫生署（PHS）：计划结束后为期三年［45 CFR Part 45］＊
- 美国食物及药物管理局（FDA）：上市后为期两年［21 CFR Part 48］
- 研究机构：研究资料发表之后为期不一，从两年至无尽期

＊译注：美国政府法规及公报之数据库

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

数据的的保留及储存

- 原始数据纪录应该储存在什么地方？
- 一般定没有明文规定
- 例如：
  - 在实验室里
  - 在研究机构的办公室里

数据的的保留及储存

- 原始数据纪录应该储存在什么样的环境下？
- 理想的环境
  - 温度及湿度都在控制下
  - 防范自然灾害及偷窃
  - 有限制、却便利的取用
- 现实情况
  - 档案柜

文件的储存

- 一般机构没有明文规定
- 建议：保持可追踪的目录来记载：
  - 有识别号码的数据种类（笔记本，录音带，X光片等）
  - 计划名称
  - 研究人员
  - 数据纪录或笔记本完成的日期
  - 储存的地方

数据的分享（Data Sharing）

数据的取用及分享

- 数据交换的目标之潜在的冲突
  - 研究员有权利因数据而获得荣誉（经由分析、在学术会议上发表、出版）

    相对于
  - 研究员有义务将数据与同侪分享

数据的取用及分享

- 我们为什么需要取用？
  - 信息的公开与交换是科学的基本宗旨
  - 在调查不当行为的指控时，取用可能有必要
  - 可能根据信息公开法，而请求取用权（译注：美国司法部　信息公开法 http://wwwusdojgov/04foia/）
  - 美国卫生署资助的研究计划均要求取用权

ＭＩＴ对数据分享的政策说明

- “及时与公开的传播ＭＩＴ的研究结果以及在学者之间自由的做信息交换，是ＭＩＴ在履行对卓越的教育及研究的承诺上很重要的一环”

ＭＩＴ科技的所有权、分布、及商业发展的指南

美国国家卫生研究院对数据分享的说明

- 美国国家卫生研究院期望及时发放与分享最后的研究数据给其它研究者的使用
- 美国国家卫生研究院要求申请者包括分享数据的计划，或是说明为什么不能分享数据（适用于超过某金额的计划）

美国国家卫生研究院的说明：为什么分享？

- 延伸美国国家卫生研究院在分享研究资源上的政策
- 加强公开的科学质询
- 鼓励不同的分析及意见
- 促进新的研究
- 支持新的或是不同假设的试验，与分析方法

为什么分享？

- 帮助教育新的研究者
- 促使最初的研究者并未想到的题目的探讨
- 容许从数据合并所产生的新数据组

什么样的信息应该被分享？

- 验证研究结果所必需的最后研究数据
- 不包括：
  - 实验室的笔记簿
  - 不完整的数据组
  - 初步的分析
  - 科学报告的草稿
  - 未来的研究计划
  - 同侪间的沟通
  - 实体的东西，例如凝胶或实验室标本等

适用于什么样的研究？

- 在美国国家卫生研究院资助下所产生的数据：
  - 基础研究
  - 临床实验
  - 问卷
  - 其它种类的研究
- 特别重要的分享：
  - 独特又不易复制的数据组
  - 庞大、昂贵的数据组

使用以人为对象的实验的警告

- 实验者必须小心：
  - 样本数目很小的研究
  - 收集敏感的数据的研究
- 然而，只要有安全措施以确保机密性、及防备透露参与者身分，即使这些数据也可以分享的

什么叫做及时？

- 在最终数据组的主要发现被接受发表之前或同时

结论（Summary）

结论

- 负责任的数据管理的政策不停的在改变
- 一些目前的看法
  - 确保精确与可信的数据组是研究者的责任，包括：
    - 主要来源（例如：笔记本）
    - 改良过的数据组

结论

- 研究机构常拥有数据
- 计划主持人或是实验室常拥有数据监护权
- 原始数据应该保存三年以上
- 美国卫生署要求受其资助的研究的数据的取用权(在结果发表后)，或许这也是其它种类的支持的要求
- 公开传播研究结果是研究界的根本宗旨

参考数据

- 美国国家统计委员会，美国国家研究院 《分享研究成果 》美国国家研究院出版
  http://wwwnapedu/books/030903499X/html/indexhtml
- MIT TLO（Technology Licensing Office）的政策
  http://webmitedu/tlo/www/guide2html
- 美国太空总署（NASA）确保信息品质的指导方针
  ftp://ftphqnasagov/pub/pao/reports/2002/NASA_data_quality_guidelinespdf
- 美国国家卫生研究院数据分享的网络主页
  http://grantsnihgov/grants/policy/data_sharing/