「健保資料資料庫管理系統解決方案」工作坊

SQL Server 與健保資料處理

陳育群

國立陽明大學醫學系/醫務管理所助理教授

國立陽明大學附設醫院教學研究部副主任

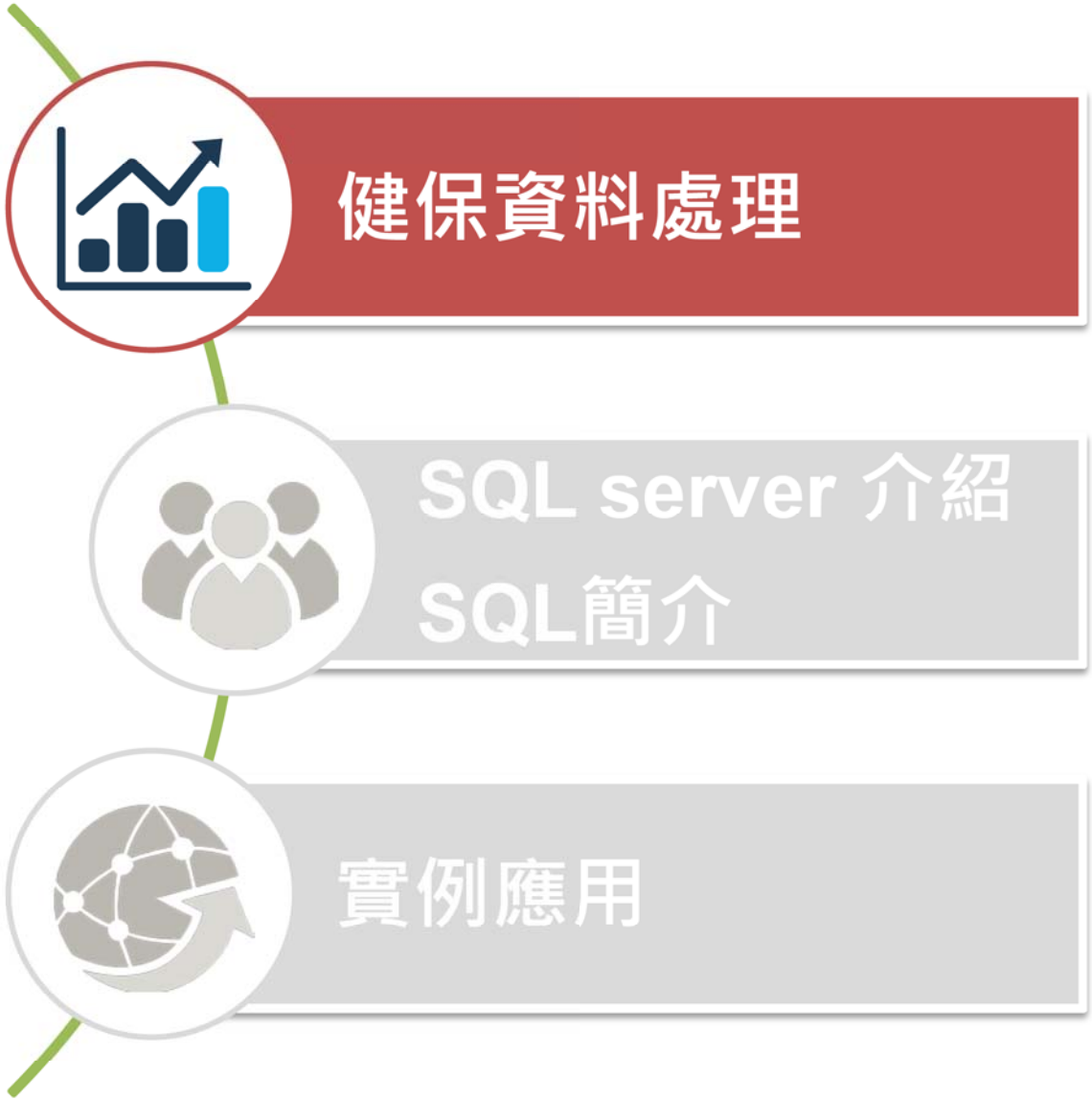
大綱

SQL Server
健保資料處理

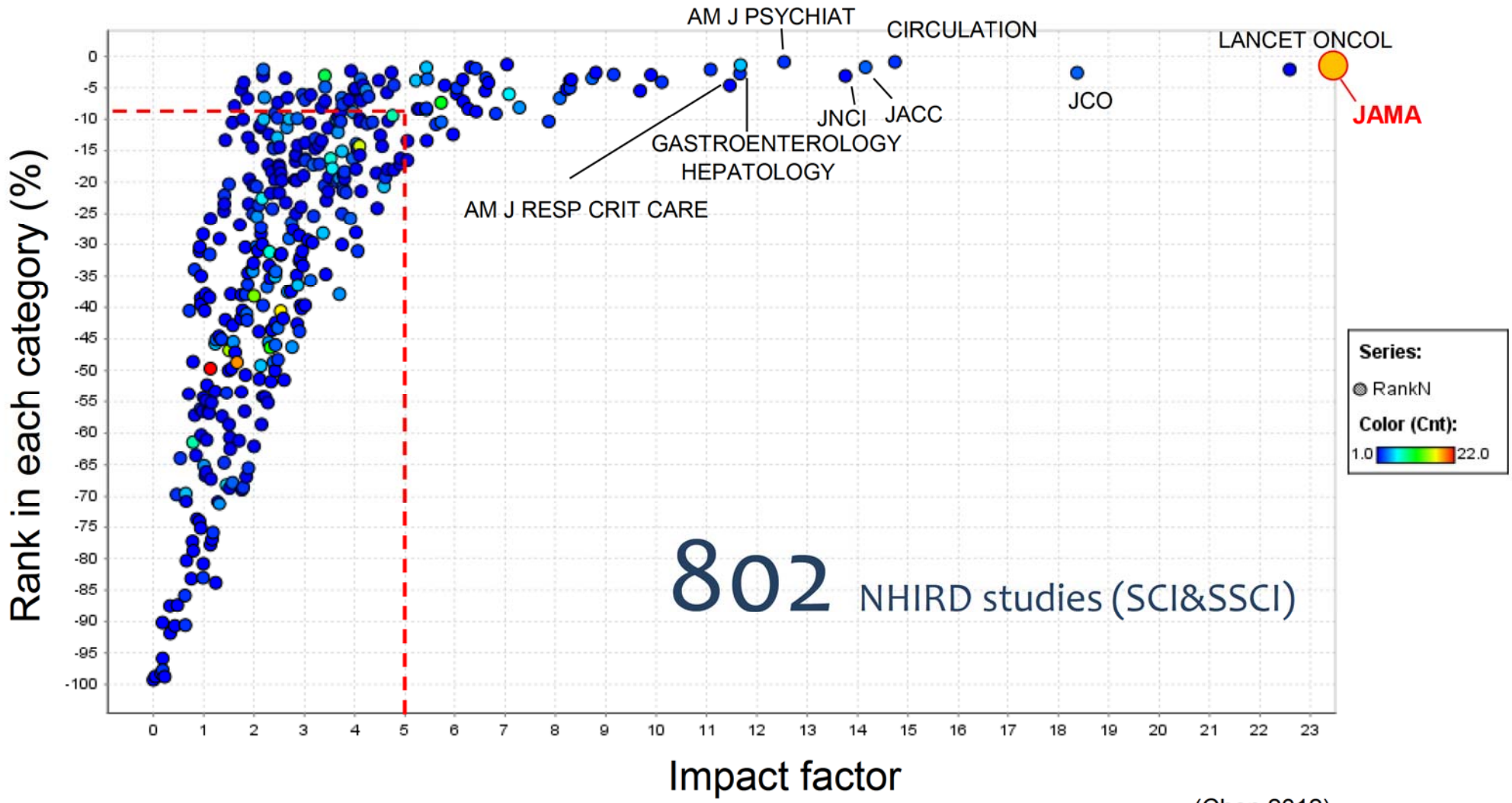
- 健保資料處理
- SQL server 介紹
  SQL簡介
- 實例應用

健保資料處理

SQL Server
健保資料處理



∵ 2001-2012年健保資料庫 研究分布



(Chen 2012)

JAMA: a new milestone

PRELIMINARY COMMUNICATION

JAMA. 2012/11

B型肝炎相關之肝癌
切除術後復發因子

ONLINE FIRST

Association Between Nucleoside Analogues and Risk of Hepatitis B Virus–Related Hepatocellular Carcinoma Recurrence Following Liver Resection

Chun-Ying Wu, MD, PhD, MPH
Yi-Ju Chen, MD, PhD
Hsiu J. Ho, PhD
Yao-Chun Hsu, MD, MS
Ken N. Kuo, MD
Ming-Shiang Wu, MD, PhD
Jaw-Town Lin, MD, PhD

Context Tumor recurrence is a major issue for patients with hepatocellular carcinoma (HCC) following curative liver resection.

Objective To investigate the association between nucleoside analogue use and risk of tumor recurrence in patients with hepatitis B virus (HBV)–related HCC after curative surgery.

Design, Setting, and Participants A nationwide cohort study between October 2003 and September 2010. Data from the Taiwan National Health Insurance Research Database. Among 100 938 newly diagnosed HCC patients, we identified 4569 HBV-related HCC patients who received curative liver resection for HCC between Oc-

Lancet oncology: another milestone

Articles

Risk of ovarian cancer in women with pelvic inflammatory disease: a population-based study PID 與 卵巢癌有關

Hui-Wen Lin, Ying-Yueh Tu, Shiyng Yu Lin, Wei-Ju Su, Wei Li Lin, Wei Zer Lin, Shen-Chi Wu, Yuen-Liang Lai

Lancet Oncol 2011; 12: 900–04

Published Online August 10, 2011 DOI:10.1016/S1470-2045(11)70165-6

See Comment page 833

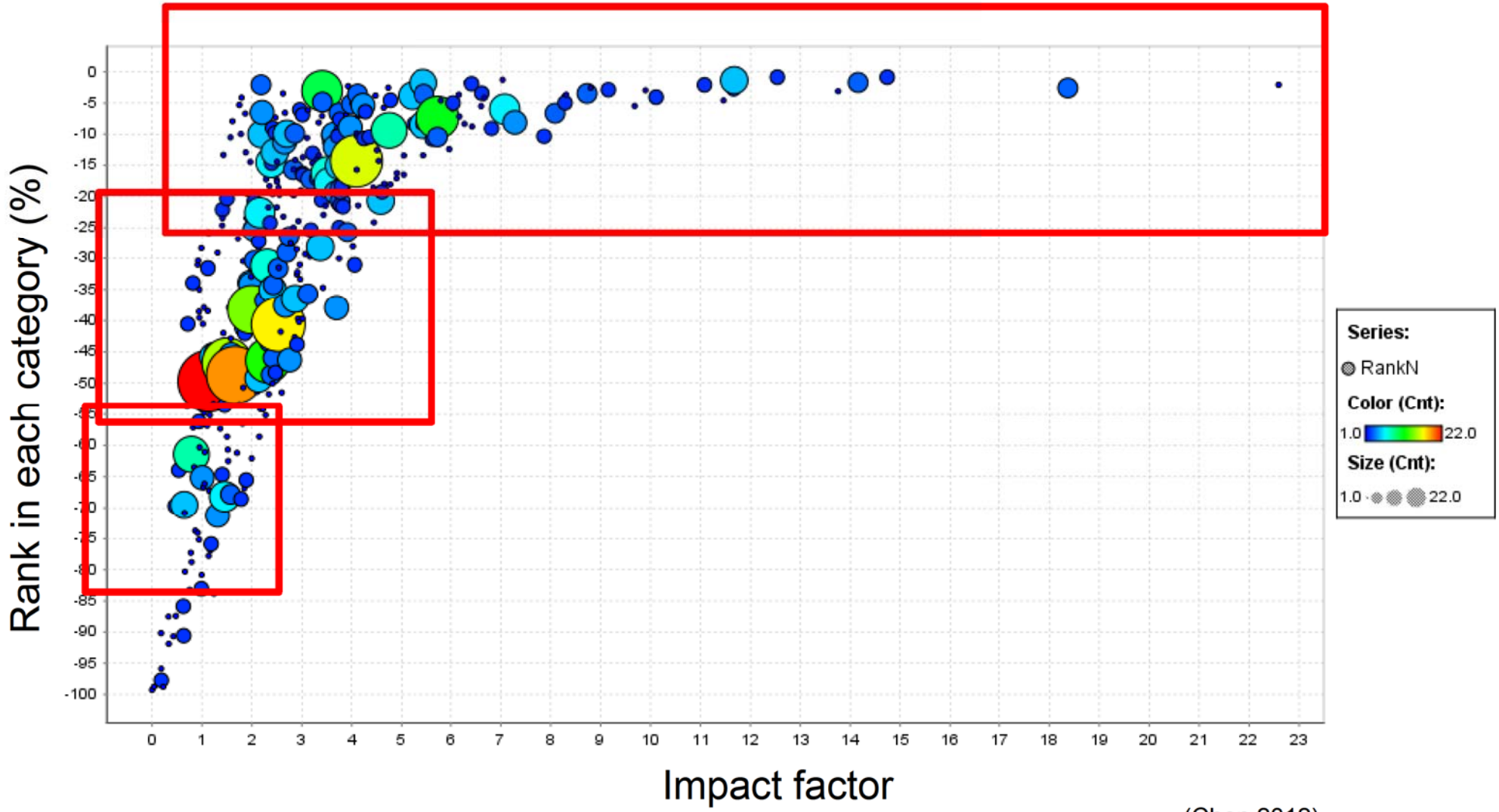
Department of Mathematics, Soochow University, Taipei, Taiwan (H-W Lin PhD); Biostatistics and Research Consultation Centre (H-W Lin), School of Medicine (S Y Lin, W L Lin BA, Y-L Lai MD), and

Summary

Background Ovarian cancer is commonly fatal and incidence has persistently risen in Taiwan over the past 20 years. Prevention strategies, however, are limited. Pelvic inflammatory disease (PID) has been suggested to increase the risk of developing ovarian cancer, but the results of studies have been inconsistent. Therefore, we investigated whether PID increases the risk of developing ovarian cancer in a large, nationwide cohort.

Methods From the Longitudinal Health Insurance Database 2005 (LHID2005) in Taiwan, we obtained data for women aged 13–65 years for whom a diagnosis of PID, confirmed by multiple episodes, had been recorded between Jan 1, 2004, and Dec 31, 2005. We also obtained data for two controls per patient, matched for age and the year of first entry into the LHID2005. All patients were followed up from the date of entry in the LHID2005 until they developed ovarian cancer or to the end of 2006, whichever was earlier. We used Cox's regression models to assess the risk of developing ovarian cancer, with adjustment for age, comorbid disorders, and socioeconomic characteristics.

Segmentation of SCI studies and SCI journals (2001-2012)

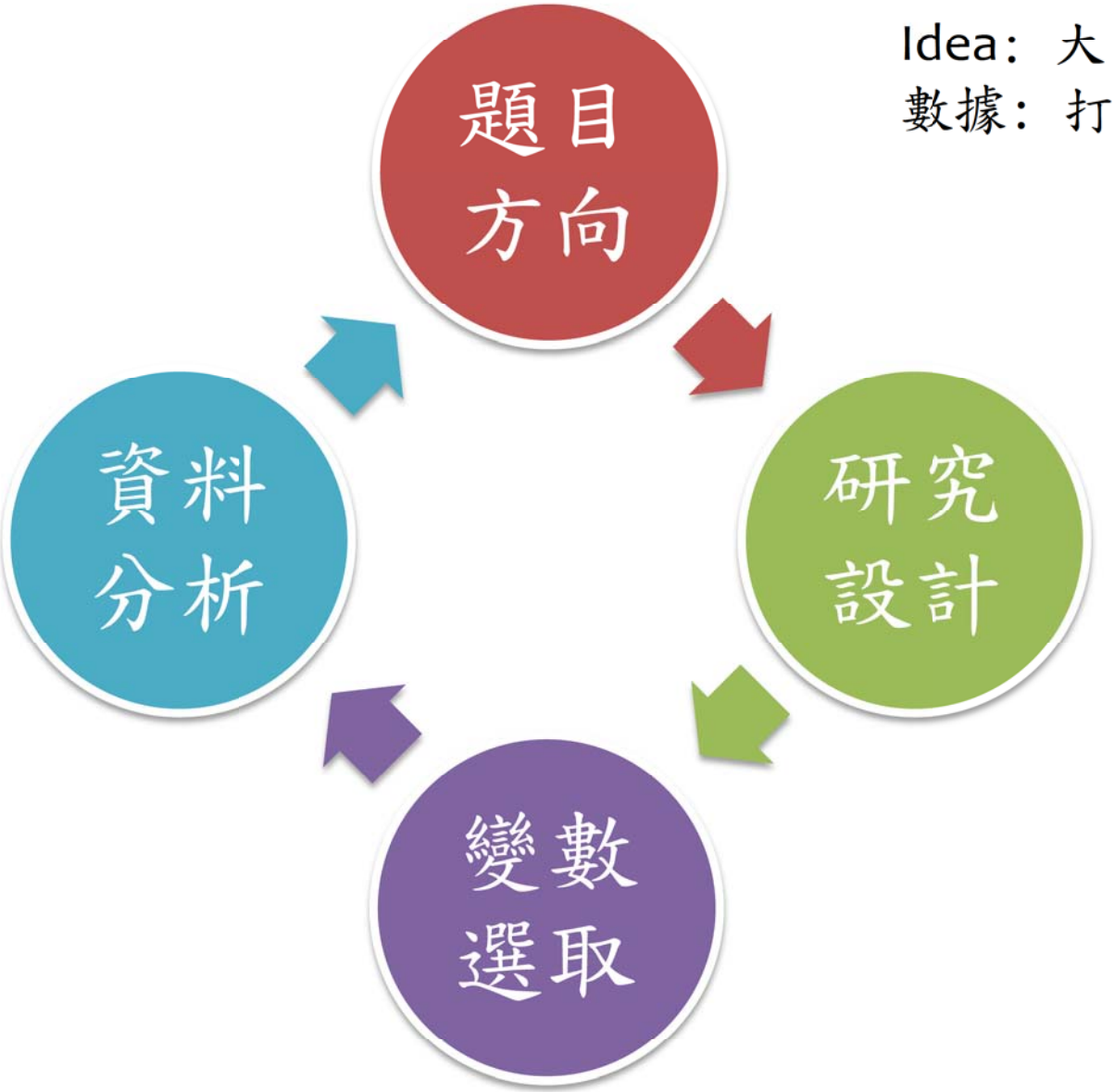


(Chen 2012)

Features of good studies

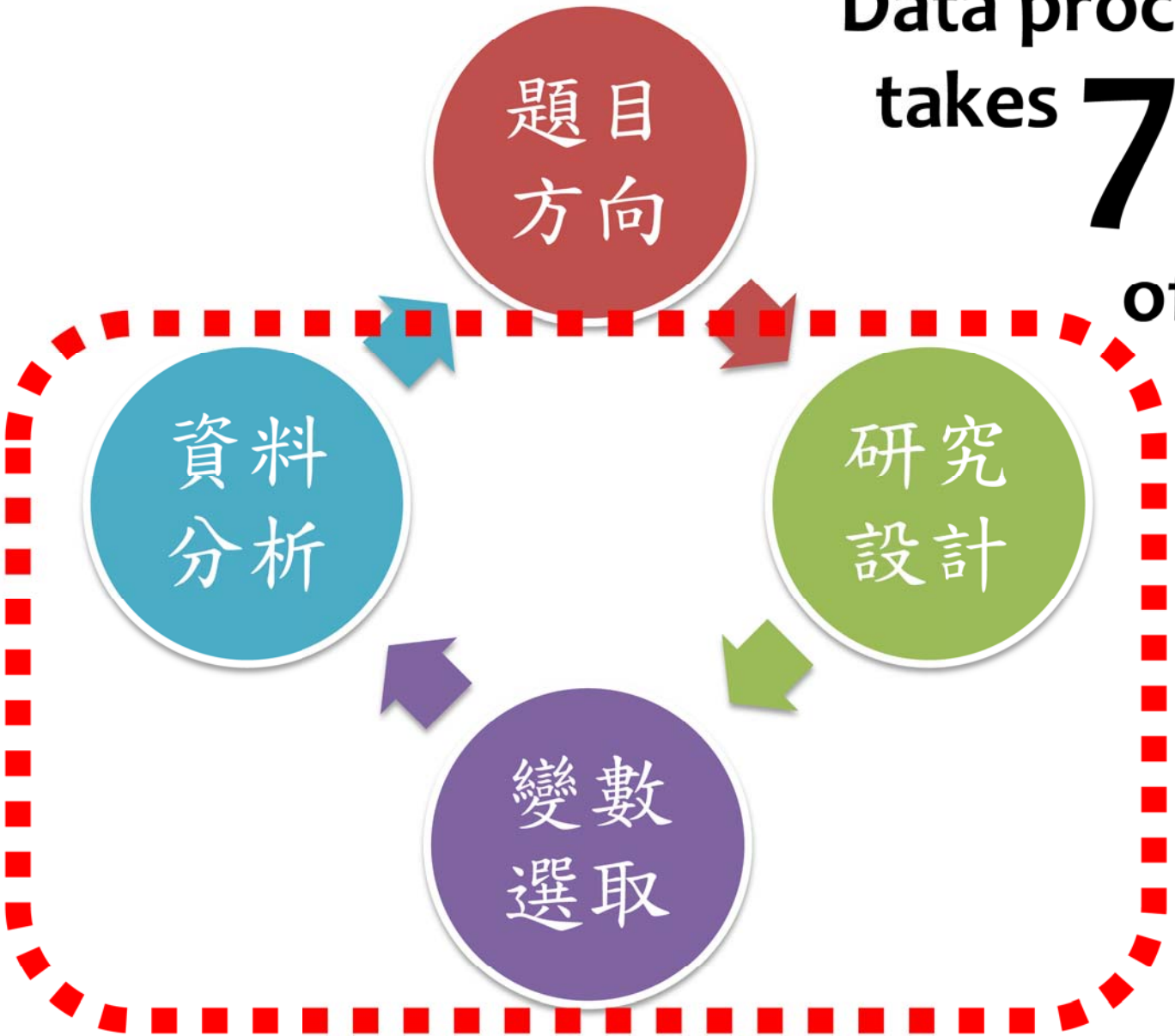
In terms of NHIRD…

Idea：大家都有
數據：打開電腦就可以跑



NHIRD studies

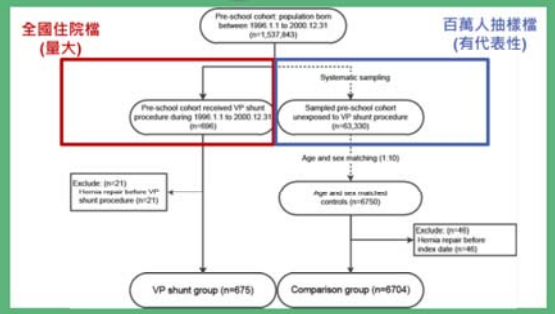
Data processing takes 75% of the task



A journey from idea to result

1. Idea

2. Study

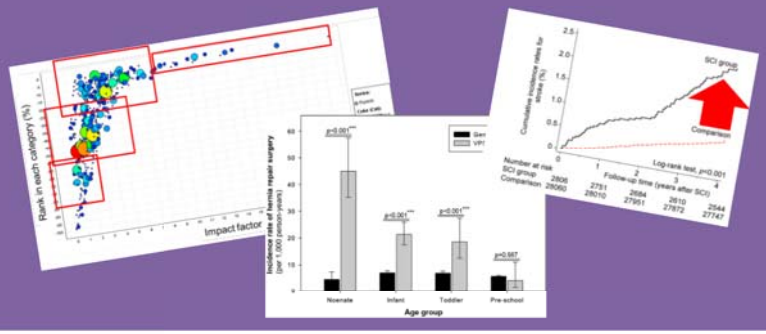


全民健康保險研究資料庫
National Health Insurance Research Database

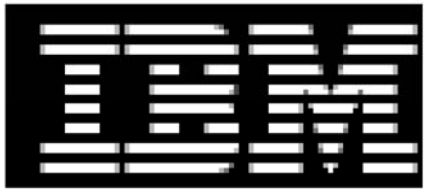
3. Data

- (1)
- (2)
- (3)

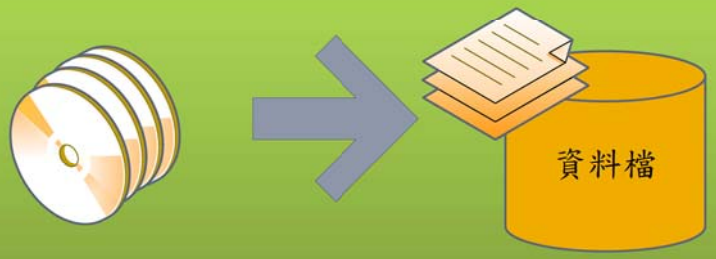
4. Result



常使用於健保資料庫研究分析工具



Three main components of data processing



1.資料轉換

READ

2.變數處理

SELECT, FILTER
APPEND, UNION
JOIN, SORT

3.結果分析

AGGREGATE
OUTPUT
STATISTICS

1. 資料轉換

- 由資料來源讀取出資料，將它們轉換成適合分析的型態，並且將它們匯入資料庫內。
- 通常還要搭配著資料清潔（Data Cleaning）將系統源頭許多未經整合的、不允許的、遺失的或者錯誤的資料，在匯入之前重新整頓(Garbage in Garbage out)。

2. 變數處理

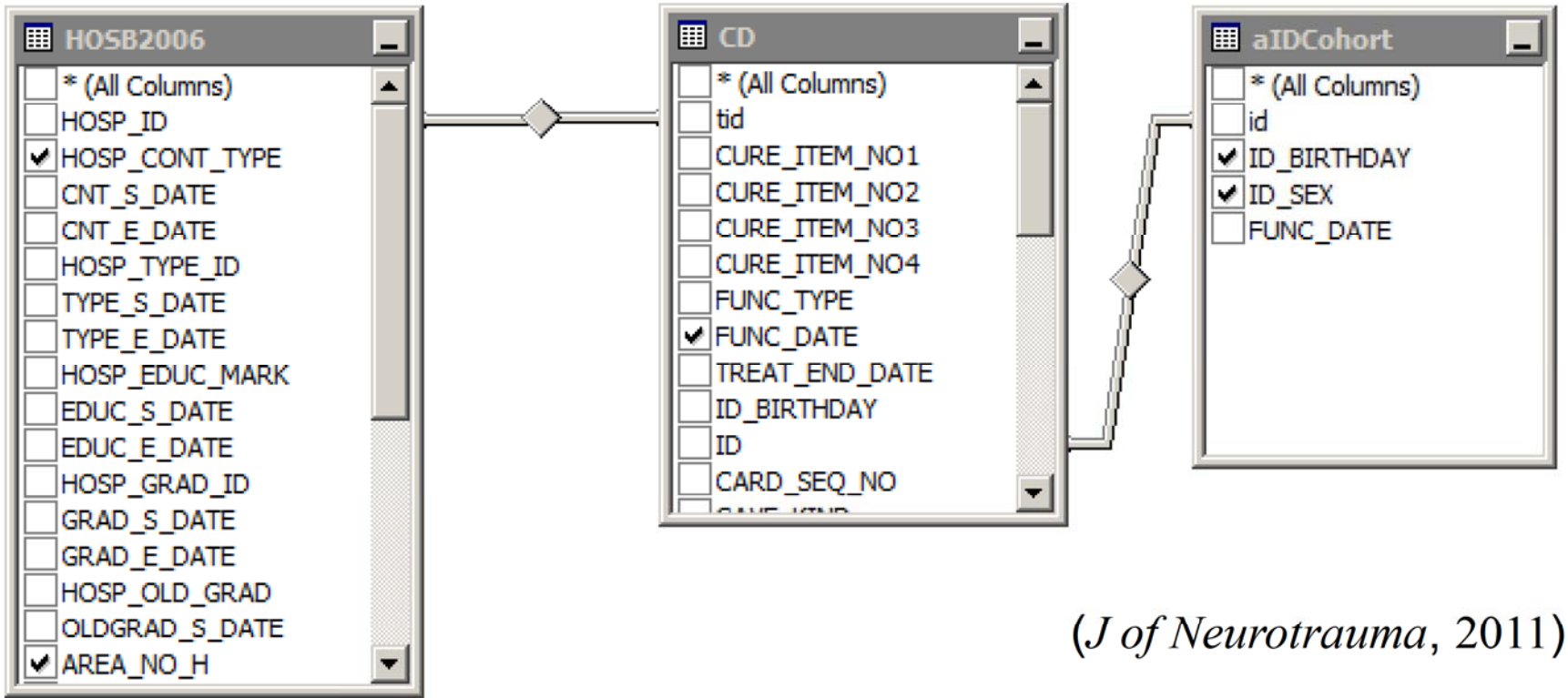
- 把分別儲存於不同表格的原始資訊，如醫院層級、藥物分類、疾病分類、重大傷病等等「串連」許多個資料表。
- 選取適當的變數(proxy variable)，進行資料加值。
- 敏感度測試。

3. 結果分析

- 把整理好的資料，以表格方式 (tabular) 輸入分析軟體 (資料探勘/統計/繪圖軟體)，呈現有意義的結果發現。

資料串連加值:
通常是最重要/耗時的步驟

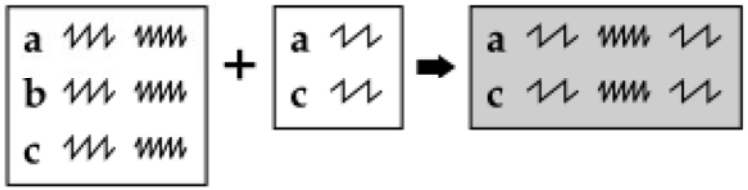
- 已知某個族群(ID_Cohort)，想了解這個族群在2000年到2007年的就診情形(就診科別、醫院層級、醫院所在地點、看診日期)



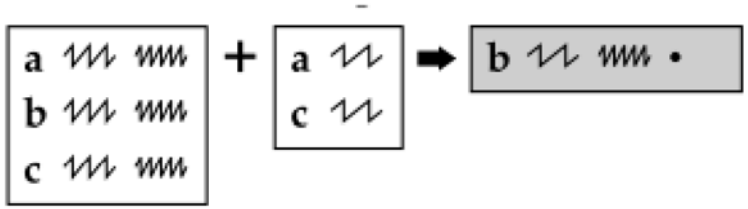
(*J of Neurotrauma*, 2011)

Basic idea for data join / merge

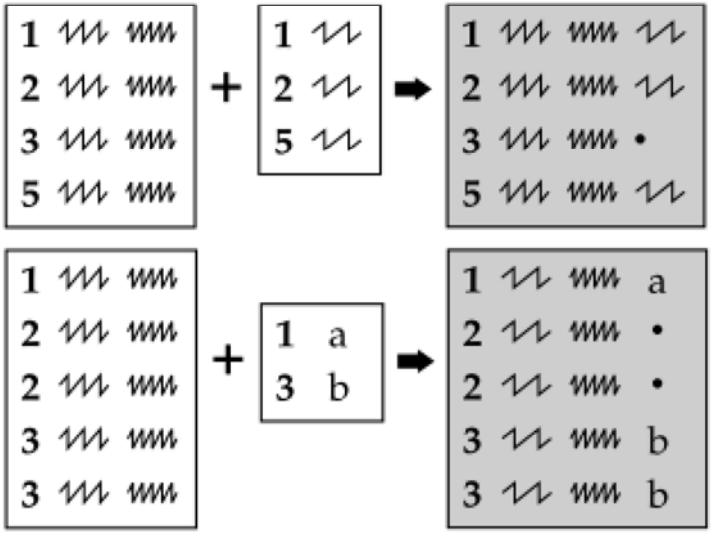
1. Matching obs.



2. Non-matching obs.



3. Combined (Matching + non-matching)

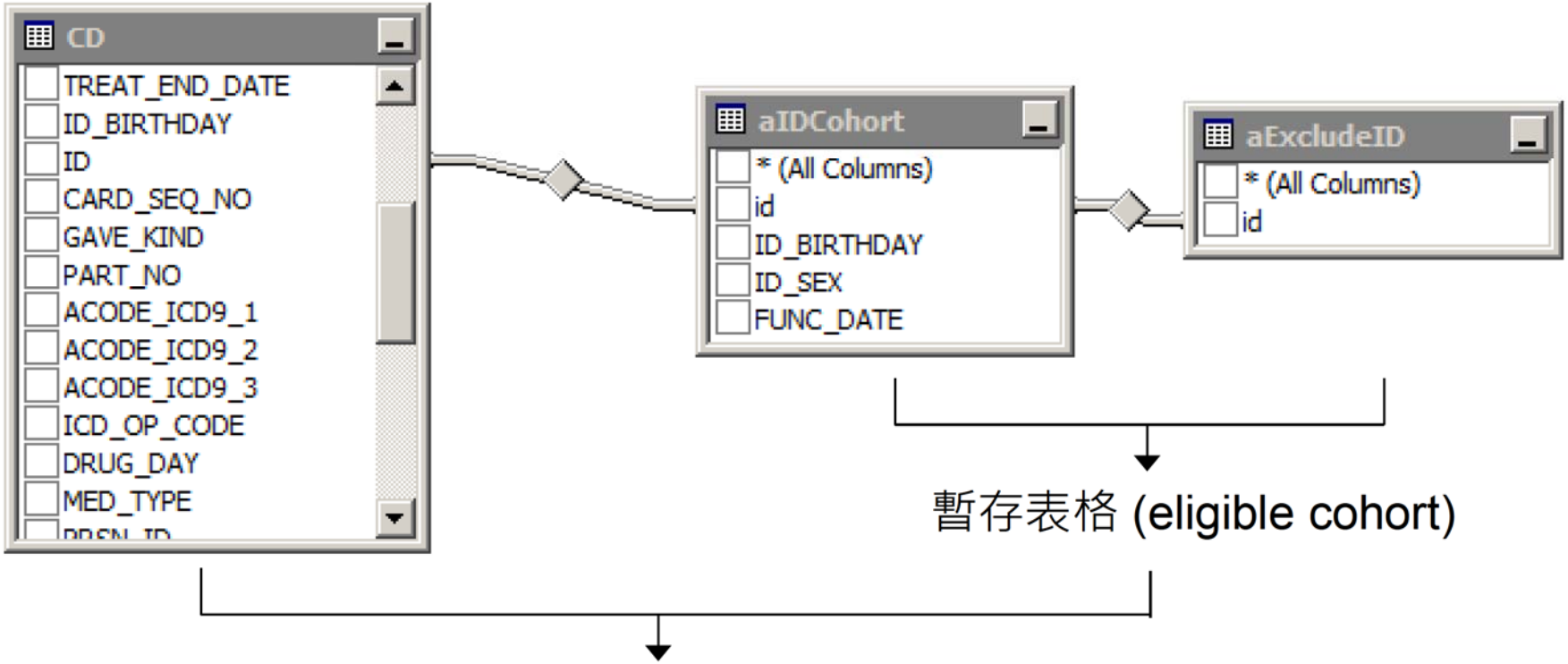



```
DATA new-data-set;
   MERGE data-set-1 data-set-2;
   BY variable-list;
```

```
DATA both;
   MERGE state (IN = InState) county (IN = InCounty);
   BY StateName;
```

資料串連加值:

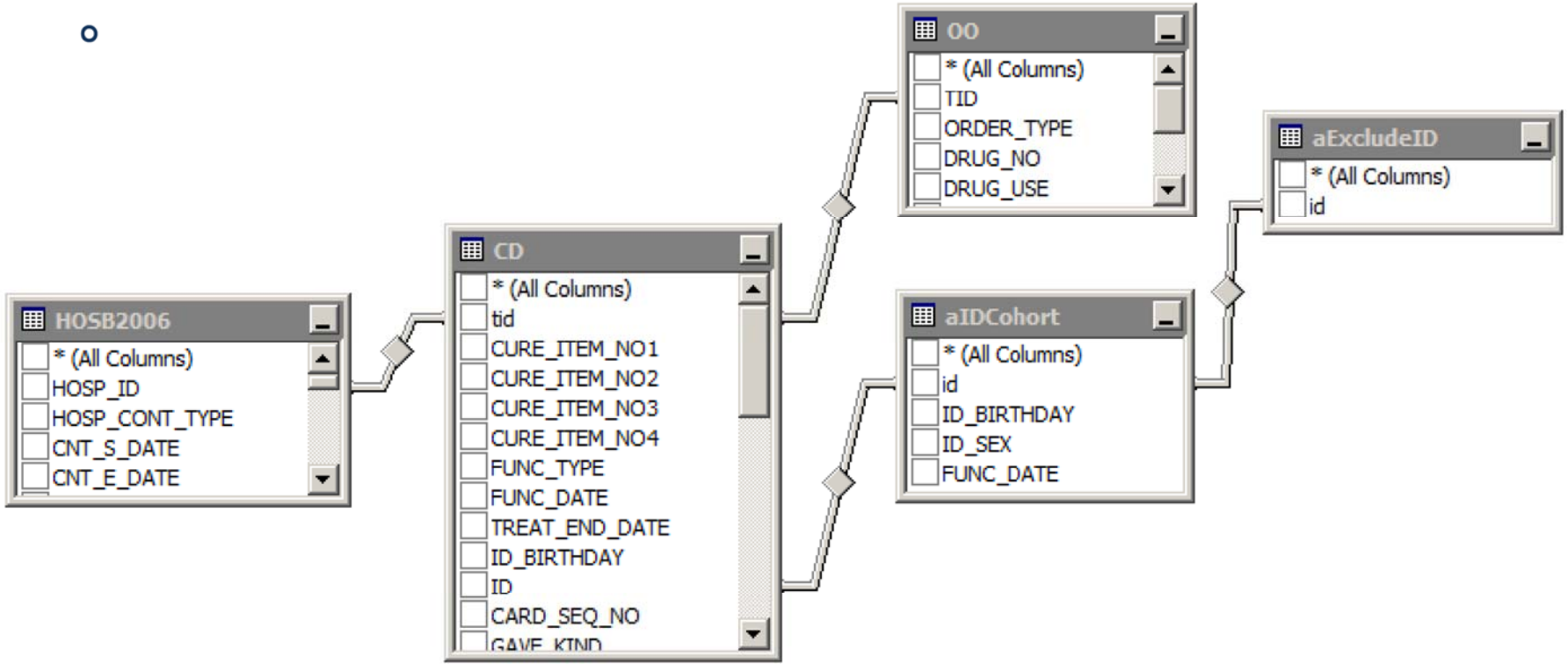
- 已知某個族群(ID_Cohort)，排除掉(exclusion criteria)，想抓取該族群在2000年到2007年內第一次就診科別



(*J of Neurotrauma*, 2011)

資料串連加值:

- 已知某個族群(ID_Cohort)，扣掉不符合的案例(exclusion criteria) ，想了解2000年到2007年內最後一次**就診日期**及**就醫地點科別**、**治療時期**及**使用藥物**。



基本串連功能不足以應付複雜處理資料需求

- 需要有一種工具能夠儲存大量分散且相關連的資料。
- 能夠有一種指令，像英文一樣直覺將大筆資料互相串連，並擷取所需的部份資料。
- 最好還能夠提供基本的統計加總功能。

→ Invention of RDBMS and SQL

關聯式資料庫管理系統 (RDBMS) 與結構式查詢語言 (SQL)

1970年代, 因應大量資料處理，IBM 推出：

RDBMS (Relational database management system)

- 關聯式資料庫管理系統被發展用來處理/儲存銀行公司等大量錯綜複雜資料。

SQL (Structured Query Language)

- 用於大型資料庫中資料的定義/操作/查詢/控制。

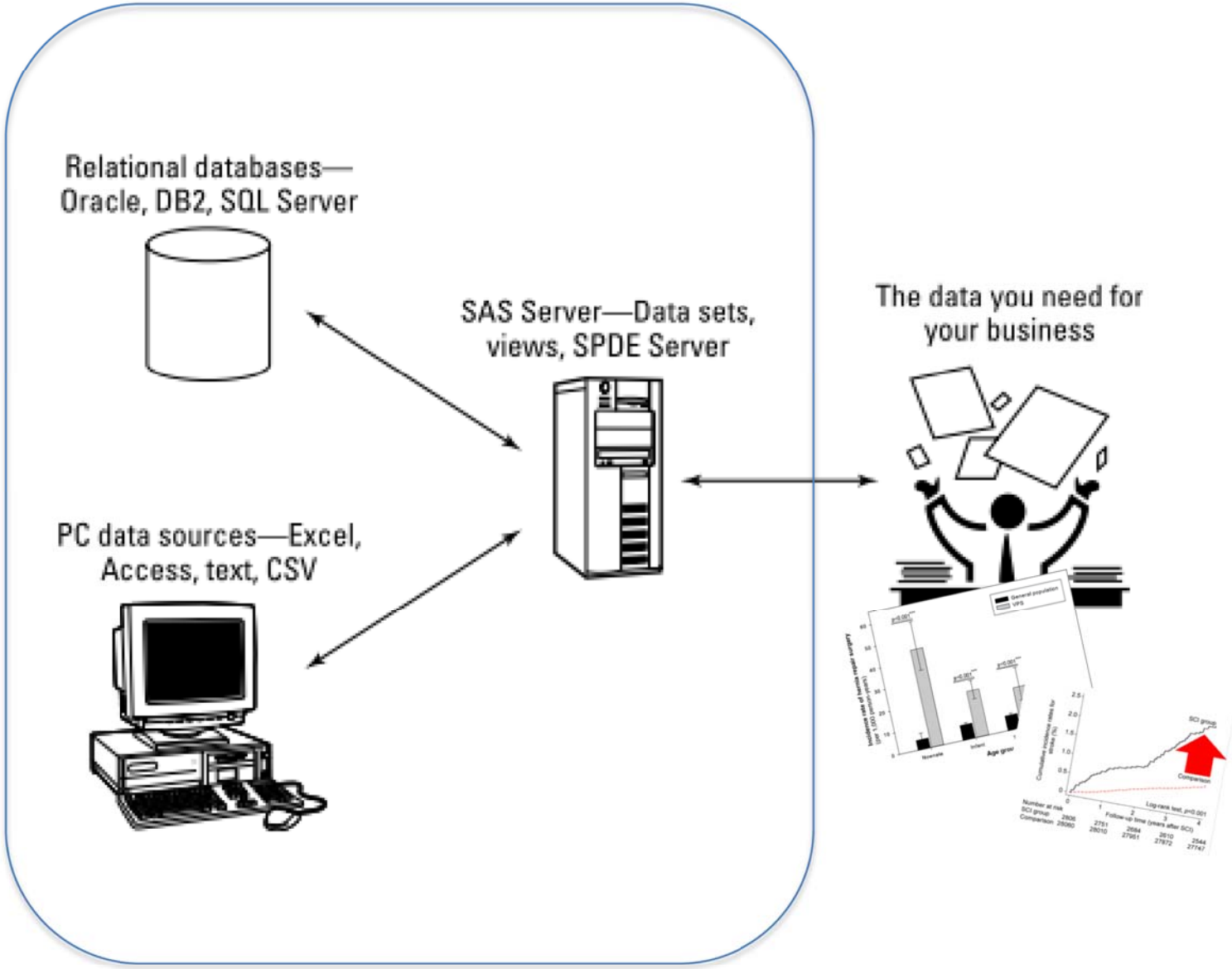
→ 成為現代資料庫系統標準。

Data Preparation

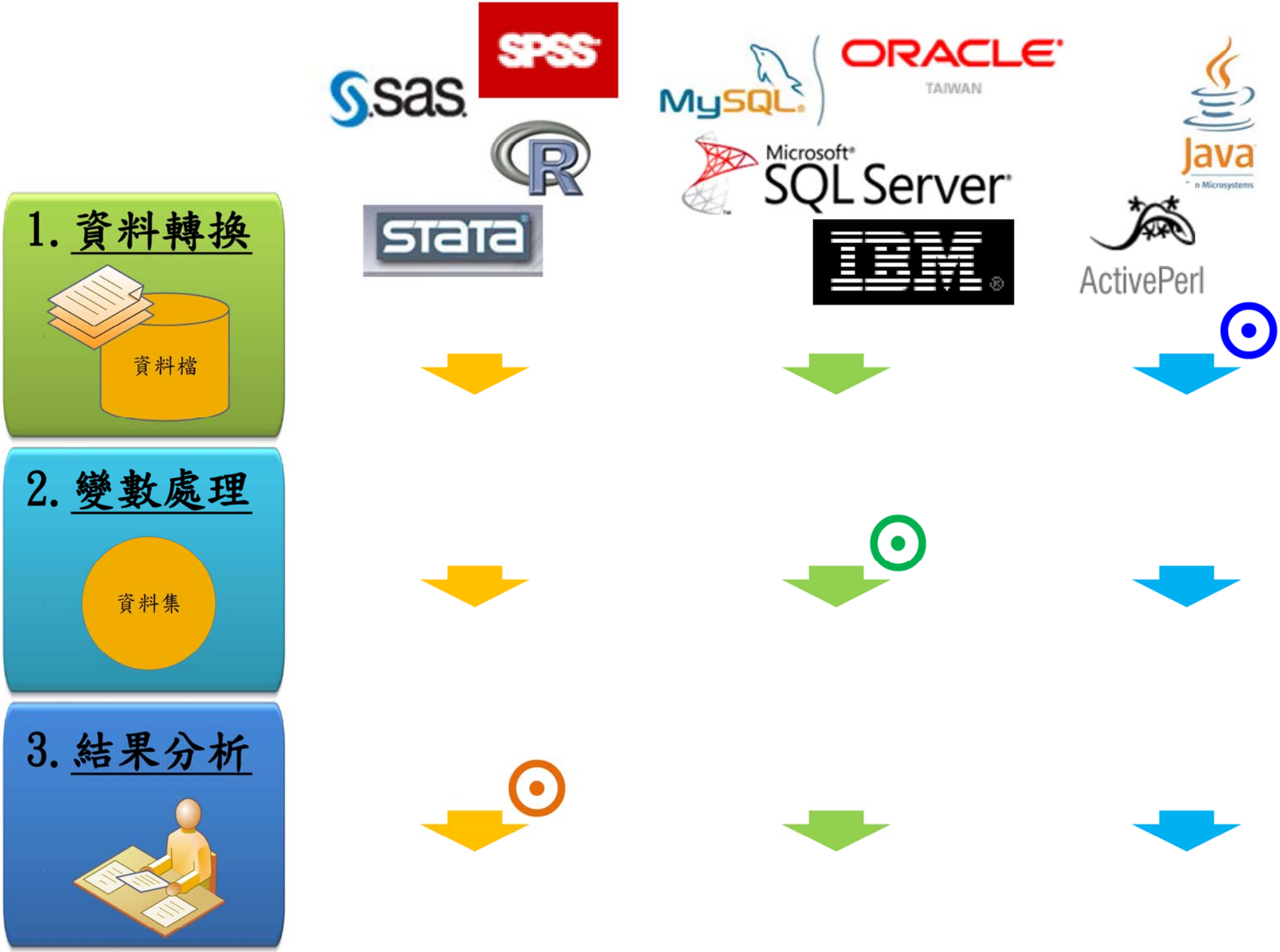
1. 資料轉換

2. 變數處理

3. 結果分析




SPSS
sas
STATA
R
ORACLE
TAIWAN
MySQL
IBM
Microsoft
SQL Server
Java
Sun Microsystems
ActivePerl
統計套裝軟體
資料管理系統
程式語言


1. 資料轉換
資料檔
2. 變數處理
資料集
3. 結果分析
sas
SPSS
R
STATA
MySQL
ORACLE
TAIWAN
Microsoft
SQL Server
IBM
Java
ActivePerl

Current trend: 統計軟體-結合後端資料庫處理系統

統計軟體 (友善)

資料處理軟體(強悍)

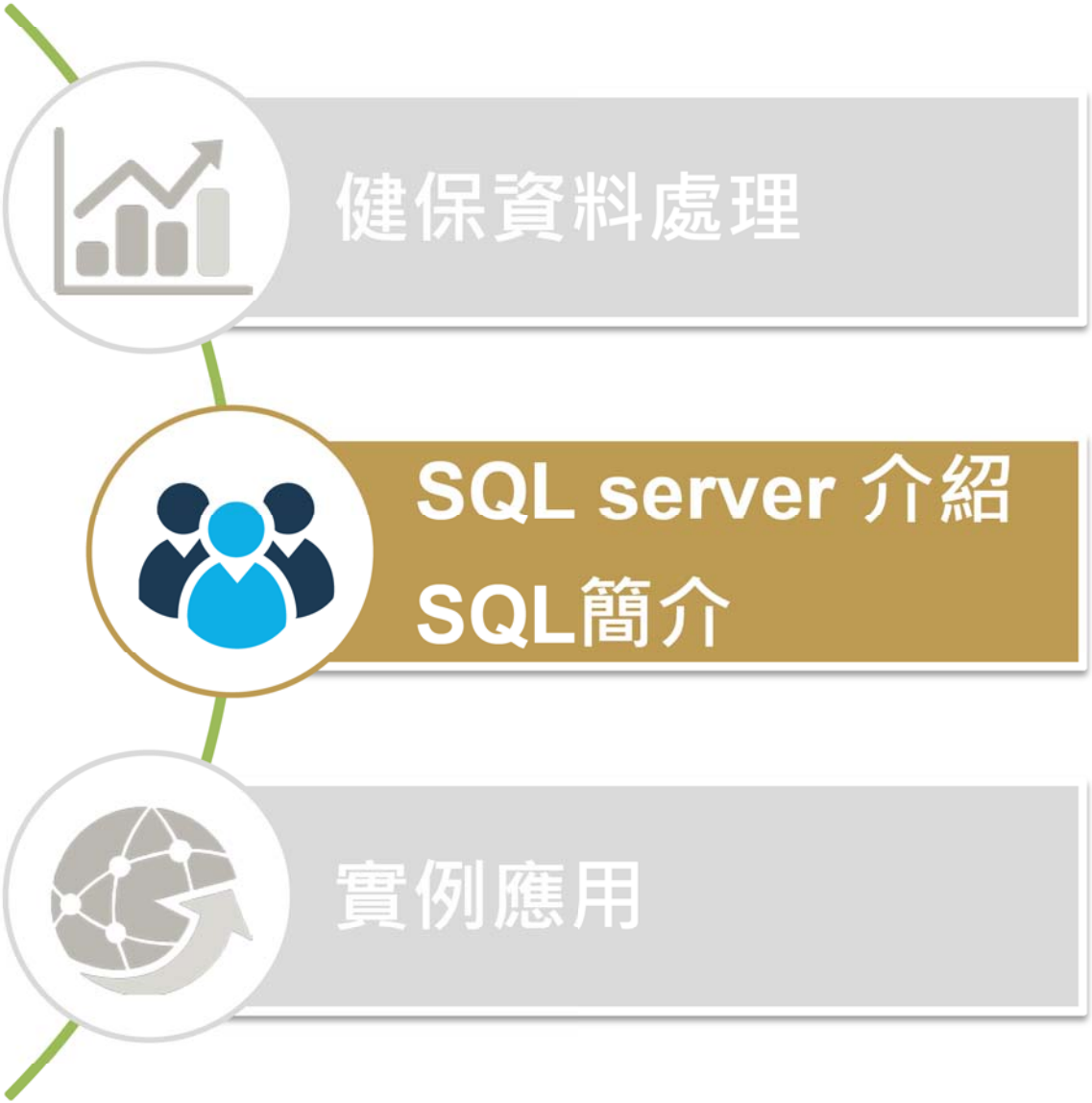
1.直接連接

2. 內建SQL語法 (Proc SQL)

- 作者合作促進健保資料庫研究學術產出
- 善用不同軟體之長處，可以減少學習摸索門檻

SQL server & SQL簡介

SQL Server
健保資料處理



Current available solution to data management

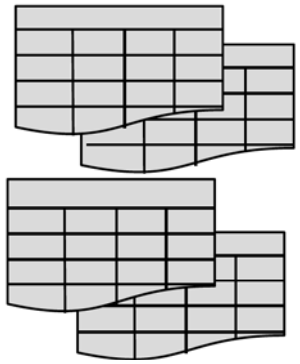
1.資料轉換 >

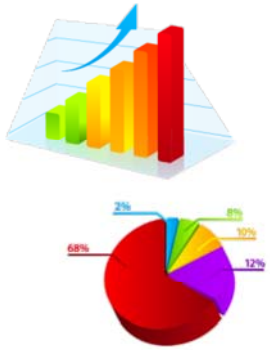
2.變數處理 >

3.統計分析 >

健保資料庫

資料庫





MySQL

IBM

Microsoft SQL Server

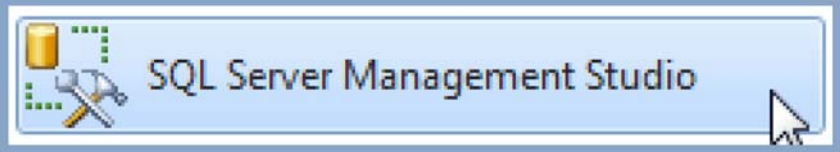
ORACLE

SPSS
STATA
SAS
EXCEL....

SQL 資料儲存概念圖

SQL	SAS	SPSS
❶ DATABASE	LIBRARY	x
❷ TABLE	TABLE	FILE
VIEW	x	X

SQL Server Management Studio



SQLQuery1.sql - 180.218.247.126.SrcCohort1M2k (sa (53))* - Microsoft SQL Server Management Studio (Administrator)

File Edit View Query Project Debug Tools Window Help

New Query | SrcCohort1M2k | Execute | Debug

Object Explorer

Connect

- 180.218.247.126 (SQL Server 11.0.3000 - sa)
  - Databases
    - System Databases
    - Database Snapshots
    - SrcATC (Read-Only)
    - SrcCohort1M2k (Read-Only)
      - Database Diagrams
      - Tables
        - System Tables
        - FileTables
        - dbo.CD1996
          - Columns
            - tid (char(57), null)
            - CURE_ITEM_NO1 (char(
            - CURE_ITEM_NO2 (char(
            - CURE_ITEM_NO3 (char(
            - CURE_ITEM_NO4 (char(
            - FUNC_TYPE (char(2), nu
            - FUNC_DATE (char(8), nu
            - TREAT_END_DATE (char
            - ID_BIRTHDAY (char(8), r
            - ID (char(32), null)
            - CARD_SEQ_NO (char(4),
            - GAVE_KIND (char(1), nu
            - PART_NO (char(3), null)
            - ACODE_ICD9_1 (char(5),
            - ACODE_ICD9_2 (char(5),
            - ACODE_ICD9_3 (char(5),
            - ICD_OP_CODE (char(4),
            - DRUG_DAY (char(2), nul
            - MED_TYPE (char(1), null
            - PRSN_ID (char(32), null)

1

SQLQuery1.sql - 18...ohort1M2k (sa (53))*

```
-- 逛醫院: 定義為一年內看超過兩家醫院的人
-- 選取  2005 逛醫院者 的就醫次數

SELECT id, count(distinct substring(tid,8,34) ) as [就診次數]
FROM cd2005 GROUP BY id
HAVING count(distinct substring(tid,8,34) ) > 2
```

100 %

2

Results

	id	就診次數
1	00002a71971c6e590278ebd780804dd1	4
2	0000cb3ca917553de5a152d1ed0c261a	4
3	0001ab4b6b8a4390179a78c67fade9b1	6
4	0001e0755255eec8d3ec439df3745922	5
5	00023816852df5976494ac8a32796435	5
6	00035b68449ac9c9c120009c2f8129f1	10
7	000366aa317980cc51cd6710648b7cb7	4
8	0003d93e378f2856ea6c23022f91e897	3
9	00041000b81f6642244e30a626e168d5	10
10	00046733593fff4463ce8ad188d31072	9
11	00048b666bda35af8d3b8265fc168c90	5
12	00059f45dc4e061fce05a4d4ed695a78	10
13	0006f2b1e07b50241fc3b5d6a8e4d03d	3
14	0007d85d90c02ab1fbb52053a442011e	4
15	00080035058781f0c867abd4c8ea635a	9
16	000975d9d5ad3eac6664d43067fe6403	7
17	0009e9c11c27c8f46058e61309bce1f4	6
18	000a12f59abb45acb0a96a3f5c04e6d9	8
19	000a596e06380f85fc9c97cc5bca8139	8

Query executed successfully. 180.218.247.126 (11.0 SP1) | sa (53) | SrcCohort1M2k | 00:05:49 | 597082 rows

Ready | Ln 1 | Col 1 | Ch 1 | INS

SQL:專為資料處理而設計的語言 → 簡單/易讀/能夠串連大量表格

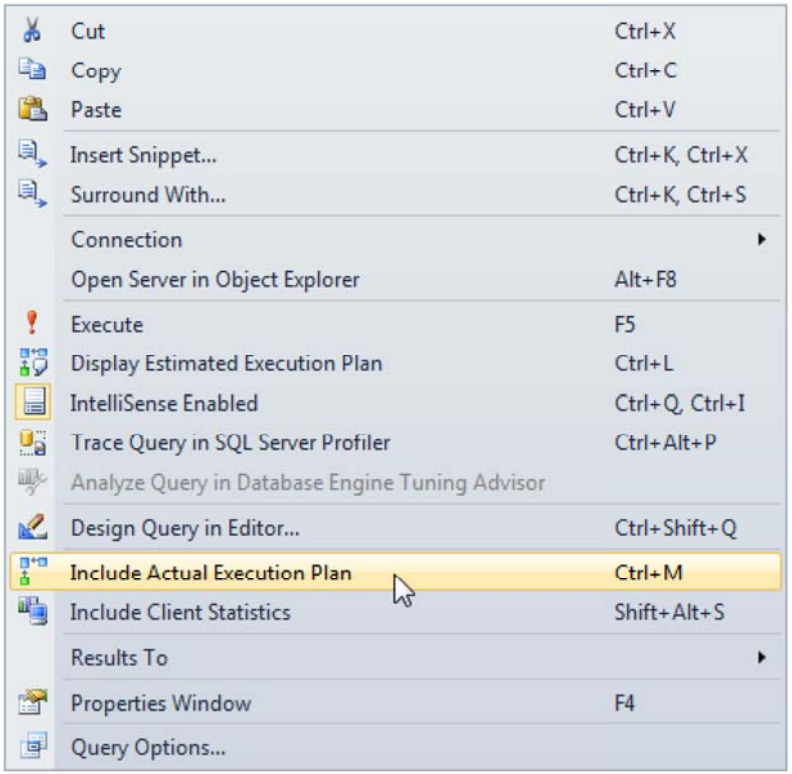
- Data Definition Language (DDL)
  - 建置資料 / 維護資料 / 定義資料
  - 總共只有三大命令
  - CREATE
  - ALTER, DROP [知道即可]
- Data Manipulation Language (DML)
  - 操作 / 查詢 / 控制 資料
  - 總共有四大命令
  - SELECT
  - INSERT, UPDATE, DELETE [知道即可]

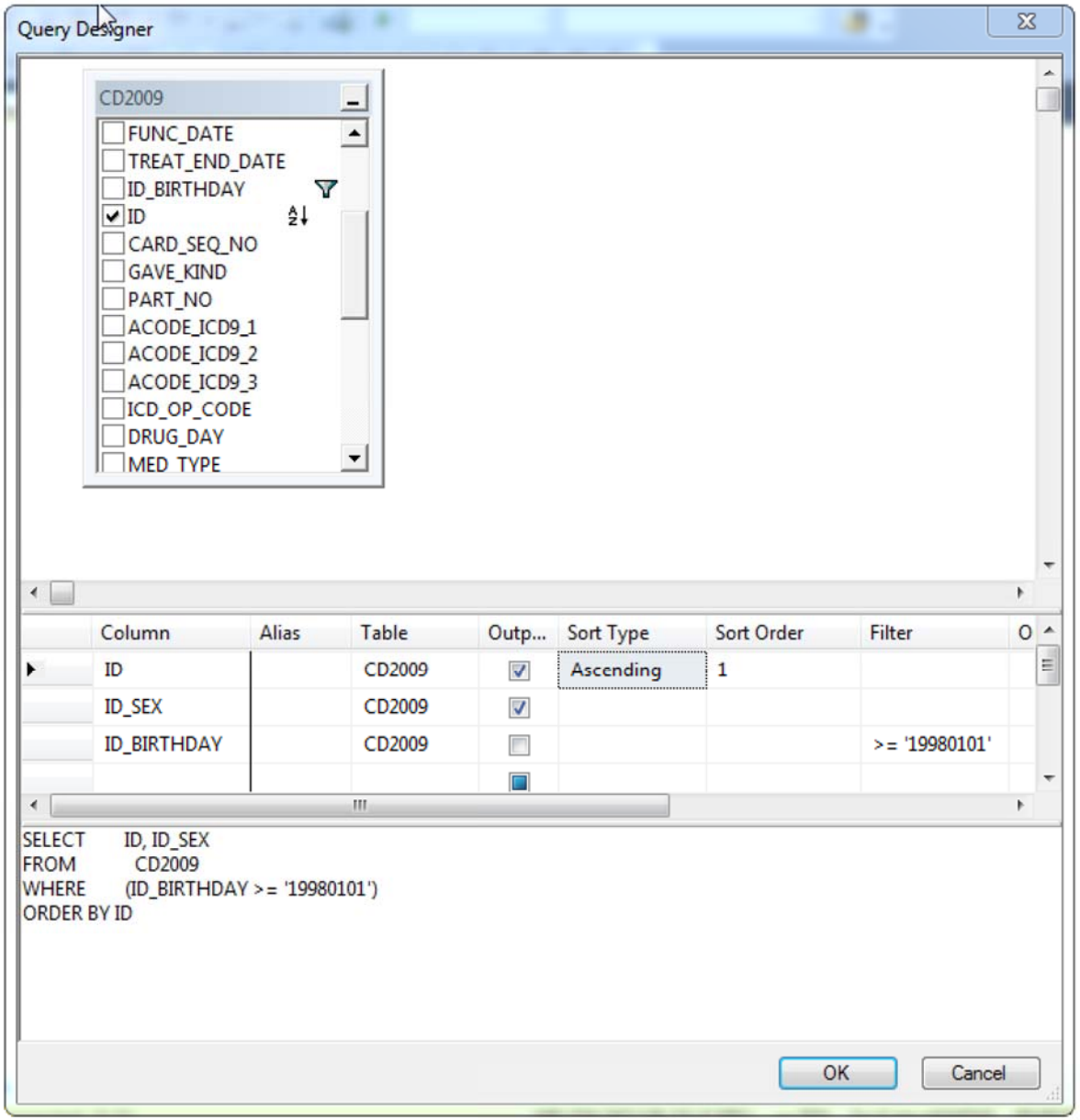
SQL Syntax

- SQL statement: 總是由動詞開始...
- Comment (註解),
- 句後加上分號 (;)

```
-- 從2009年門診檔中擷取1998年1月1日以後出生的病患紀錄並且依照 生日排序
SELECT      id,  id_sex
  FROM      cd2009
  WHERE     id_birthday >= '19980101'
  ORDER BY id;
```

查詢設計工具 (Query Designer)





選取變數 (SELECT)

選取… SELECT …

從… FROM …

依條件… WHERE …

排序… ORDER BY …

範例:
選取...從...依照(條件)...排序

SQLQuery1.sql - 1...N-NB\zCBRAIN (51))*

```
-- 2009 年門診檔中有幾筆紀錄 ?  (598574)
select id, id_birthday from cd2009 order by ID_BIRTHDAY;

-- 請計算 2009 年門診檔中, 主診斷為糖尿病 (ICD=250.x) 的紀錄有幾筆 ? (7799)
select id, id_birthday from cd2009 where ACODE_ICD9_1 like '250%';

-- 請計算 2009 年門診檔中, 主診斷為糖尿病 (ICD=250.x) 的紀錄有幾人 ? (1322)
select distinct id from cd2009 where ACODE_ICD9_1 like '250%';
```

SELECT …

FROM …

WHERE …

ORDER BY …

100 %

Results | Messages

	id	id_birthday
1	00000000000000000000000000008165	19050817
2	00000000000000000000000000008165	19050817
3	00000000000000000000000000008165	19050817
4	00000000000000000000000000008165	19050817
5	00000000000000000000000000008165	19050817
6	00000000000000000000000000008165	19050817
7	00000000000000000000000000008165	19050817
8	00000000000000000000000000008165	19050817
9	00000000000000000000000000008165	19050817
10	00000000000000000000000000008165	19050817
11	00000000000000000000000000026782	19060103
12	00000000000000000000000000026782	19060103

Query executed successfully. | 127.0.0.1 (11.0 SP1) | zCBRAIN-NB\zCBRAIN (51) | NHIRD | 00:00:05 | 598574 rows

SQL 的特色

- DISTINCT

  如： 診次 vs. 人次

- LIKE 字串比對

  糖尿病 ICD: 250x → ‘250%’

  464.0-2 → ‘464[0-2]%’

- RLIKE 正規表示式比對 (mysql特有)

SELECT 結合 DISTINCT 與 函數功能

SELECT …

FROM …

WHERE …

ORDER BY …

SQLQuery1.sql - 1...N-NB\zCBRAIN (51))

```
--選取 2009 年門診花費最高與最低值
select min(t_amt) as [年度最低花費], max(t_amt) as [年度最高花費] from cd2009;

--選取 2009 年 糖尿病門診花費最高與最低值
select min(t_amt) as [糖尿病-年度最低花費], max(t_amt) as [糖尿病-年度最高花費] from cd2009
       where ACODE_ICD9_1 like '250%';

--選取 2009 年門診人數與診次
select count(id) as [年度門診診次], count(distinct id) as [年度門診人數] from cd2009;

--選取 2009 年糖尿病門診人數與診次
select min(t_amt) as [糖尿病-門診診次], max(t_amt) as [糖尿病-門診人數] from cd2009
where ACODE_ICD9_1 like '250%';
```

100 %

Results | Messages

	糖尿病-年度最低花費	糖尿病-年度最高花費
1	00000000	00119339

SELECT 結合 分組功能

選取… SELECT …

從… FROM …

依條件… WHERE …

分組… GROUP BY …

分組條件… HAVING …

SELECT 分組結合 DISTINCT 與 函數功能

SELECT …
FROM …
WHERE …
GROUP BY …
HAVING …

SQLQuery1.sql - 1...N-NB\zCBRAIN (52))

```sql
-----------------------------------------------
-- 2009 年內 每個人看診幾次 ?
select id, count(hosp_id) as [看診次數] from cd2009
group by id order by 2 desc;

-- 2009 年內 每個人 每個科別 看診幾次 ?
select id, func_type, count(hosp_id) as [看診次數] from cd2009
group by id, func_type order by 3 desc, 2 asc, 1 asc;

-- 2009 年 每家醫院就診 門診次數多少 ? 就診人數多少 ?  申報費用多少 ?
select hosp_id, count(id) as [門診診數],
                count(distinct id) as [門診人數],
                sum(cast(t_amt as money)) as [總花費]
from CD2009 group by HOSP_ID order by 3 desc;
```

100 %

Results | Messages

	hosp_id	門診診數	門診人數	總花費
1	0000000000000000000000000000012079	38	14	14480.00
2	0000000000000000000000000000002574	18	10	9951.00
3	0000000000000000000000000000013652	26	6	12700.00
4	0000000000000000000000000000017614	51	13	12255.00
5	0000000000000000000000000000011772	7	4	6770.00
6	0000000000000000000000000000002873	6	3	11531.00
7	0000000000000000000000000000014606	8	7	3149.00

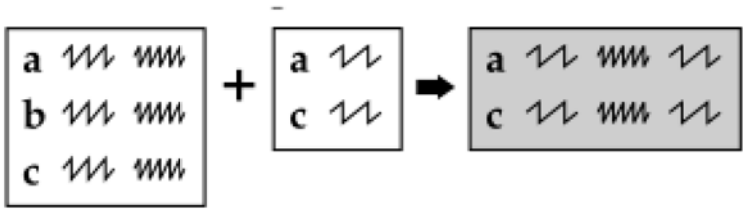
SELECT 分組結合 暫存表 與 分組條件

SQLQuery1.sql - 1...N-NB\zCBRAIN (52))

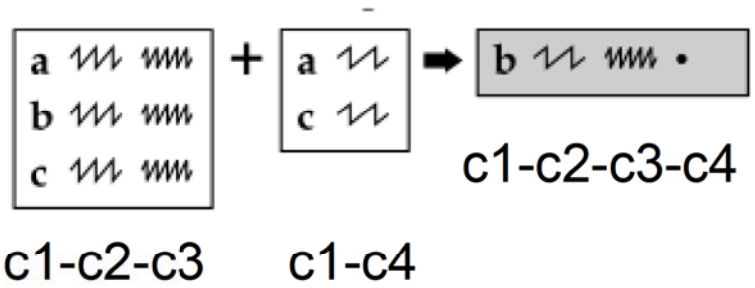
```sql
-----------------------------------------------
-- 病人逛醫院 doctor shopping 是台灣就醫特殊現象,
-- 如果將逛醫院定義為一年看超過 4 位醫師
-- 請問 2009 年有多少人是 doctor shopping ? 請問平均門診次數多少 ?請問平均花費多少 ?
-- 非 doctor shopping 者的平均門診次數多少 ?請問平均花費多少 ?
-- doctor shopper
with shopper as (
  select id, count(id) as [門診診次], sum(cast(t_amt as money)) as [門診費用]
    from CD2009 group by id having count(distinct PRSN_ID) > 4
), nonShopper as (
 select id, count(id) as [門診診次], sum(cast(t_amt as money)) as [門診費用]
    from CD2009
    where id not in (select id from shopper)
    group by id
)
 select 'Shopper', count(distinct id) as [人數],
        avg(cast([門診診次] as decimal)) as [平均門診診次],
        avg([門診費用]) as  [平均門診費用]
from shopper
union
select 'nonShopper', count(distinct id) as [人數],
        avg(cast([門診診次] as decimal)) as [平均門診診次],
        avg([門診費用]) as  [平均門診費用]
from nonShopper;
```

串連 (MERGE, JOIN)

1. Matching obs.

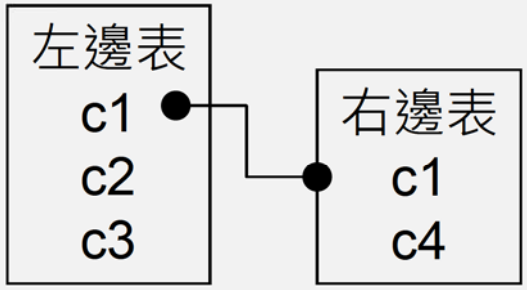


2. Non-matching obs.



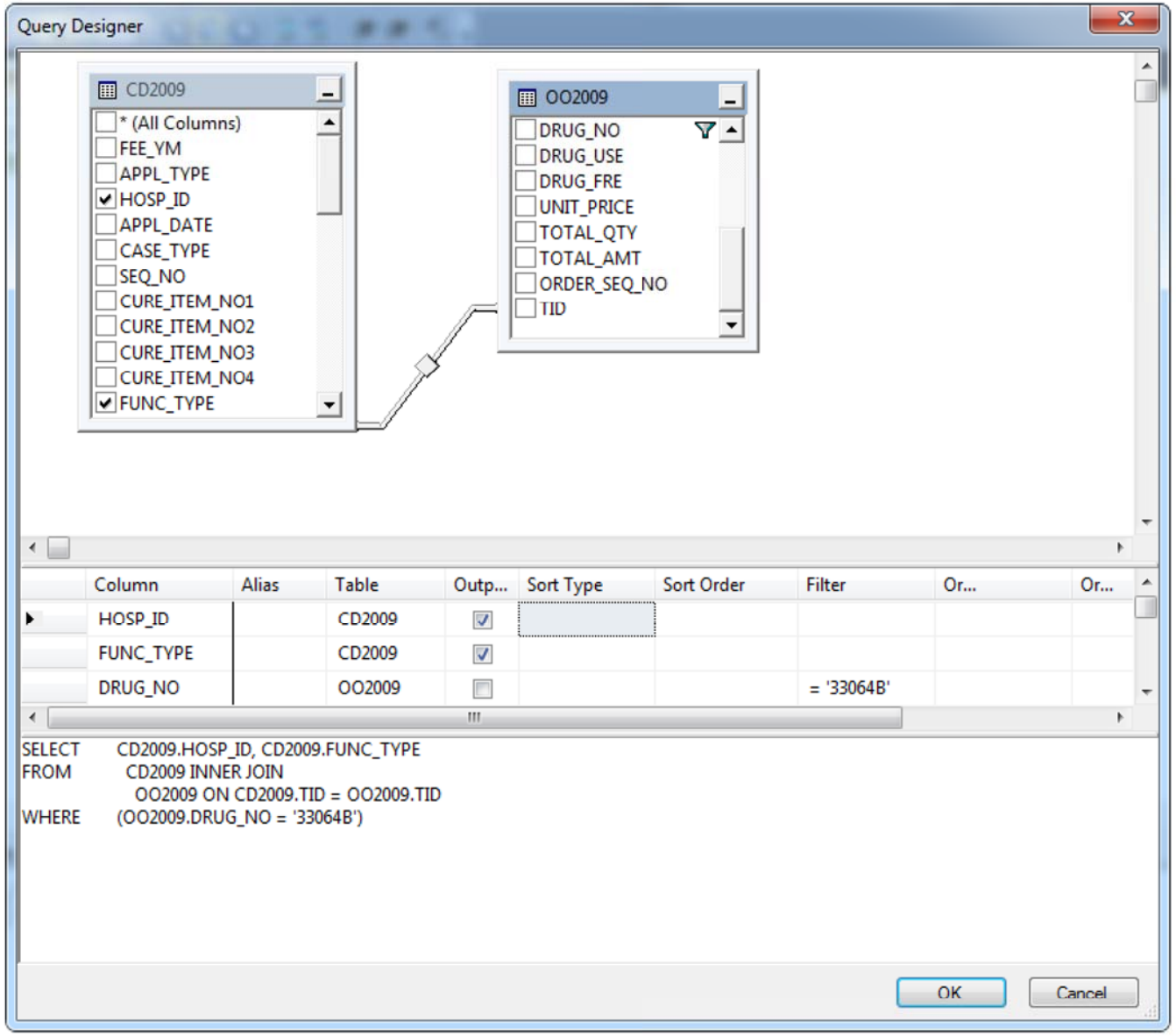

```
DATA new-data-set;
   MERGE data-set-1 data-set-2;
   BY variable-list;
```

```
SELECT c1,c2,c3,c4
FROM   左邊表 JOIN 右邊表
ON       左邊c1=右邊c1
```



串連 結合分組 與 統計函數

- 電腦斷層 (CT, 健保代碼 33064B) 屬於高貴檢查，請問2009年電腦斷層利用情形: (1) 幾次 (2) 那些科開立 (3) 那些醫院？



SQLQuery1.sql - 1...N-NB\zCBRAIN (52))

```sql
-----------------------------------------
-- 電腦斷層 (CT, 健保代碼 33064B) 屬於高貴檢查, 請問2009年
-- 電腦斷層利用情形: (1) 幾次 (2) 那些科開立 (3) 那些醫院 ?


SELECT      CD2009.HOSP_ID as [醫院代碼], CD2009.FUNC_TYPE as [科別], count(cd2009.TID) as [次數]
FROM            CD2009 INNER JOIN
                        OO2009 ON CD2009.TID = OO2009.TID
WHERE        (OO2009.DRUG_NO = '33064B')
group by    CD2009.HOSP_ID, CD2009.FUNC_TYPE;
--
```

100 %

Results | Messages

	醫院代碼	科別	次數
1	0000000000000000000000000000000381	04	1
2	0000000000000000000000000000004228	05	1
3	0000000000000000000000000000000381	06	1
4	0000000000000000000000000000002703	06	2
5	0000000000000000000000000000016735	06	1
6	0000000000000000000000000000002703	14	1
7	0000000000000000000000000000011273	AE	1

跨年度串連

SQLQuery1.sql - 1...N-NB\zCBRAIN (53))

```sql
--- 如果將逛醫院定義為一年看超過 4 位醫師
--- 針對 2009 年逛醫院者: 請問 2009 年看病次數 與 2010 年看病次數 各為何 ??
with shopper as (
  select id, count(tid) as [2009年門診次數] from CD2009 group by id having count(distinct PRSN_ID) > 4
), visit2010 as (
  select id, count(tid) as [2010年門診次數] from cd2010 group by id
)
  select shopper.ID, [2009年門診次數], [2010年門診次數]
  from shopper join visit2010 on shopper.ID=visit2010.ID;
```

100 %

Results | Messages

	ID	2009年門診次數	2010年門診次數
1	00000000000000000000000000028195	15	16
2	00000000000000000000000000019412	24	9
3	00000000000000000000000000018663	30	18
4	00000000000000000000000000024961	14	21
5	00000000000000000000000000025187	9	7
6	00000000000000000000000000034876	6	9
7	00000000000000000000000000027226	25	34
8	00000000000000000000000000007407	12	26
9	00000000000000000000000000033313	14	8
10	00000000000000000000000000039516	49	45

Power Pivot: 使用 EXCEL 進行資料處理

Microsoft

Sign in

Download Center

產品 類別 資訊安全 支援服務

Microsoft® SQL Server® 2012 PowerPivot for Microsoft® Excel® 2010

快速連結

- 總覽
- 系統要求

Microsoft PowerPivot for Microsoft Excel 2010 提供突破性的技術，例如，快速操作大型資料集、簡化資料的整合，而且能夠透過 Microsoft SharePoint 輕鬆共用您的分析結果。

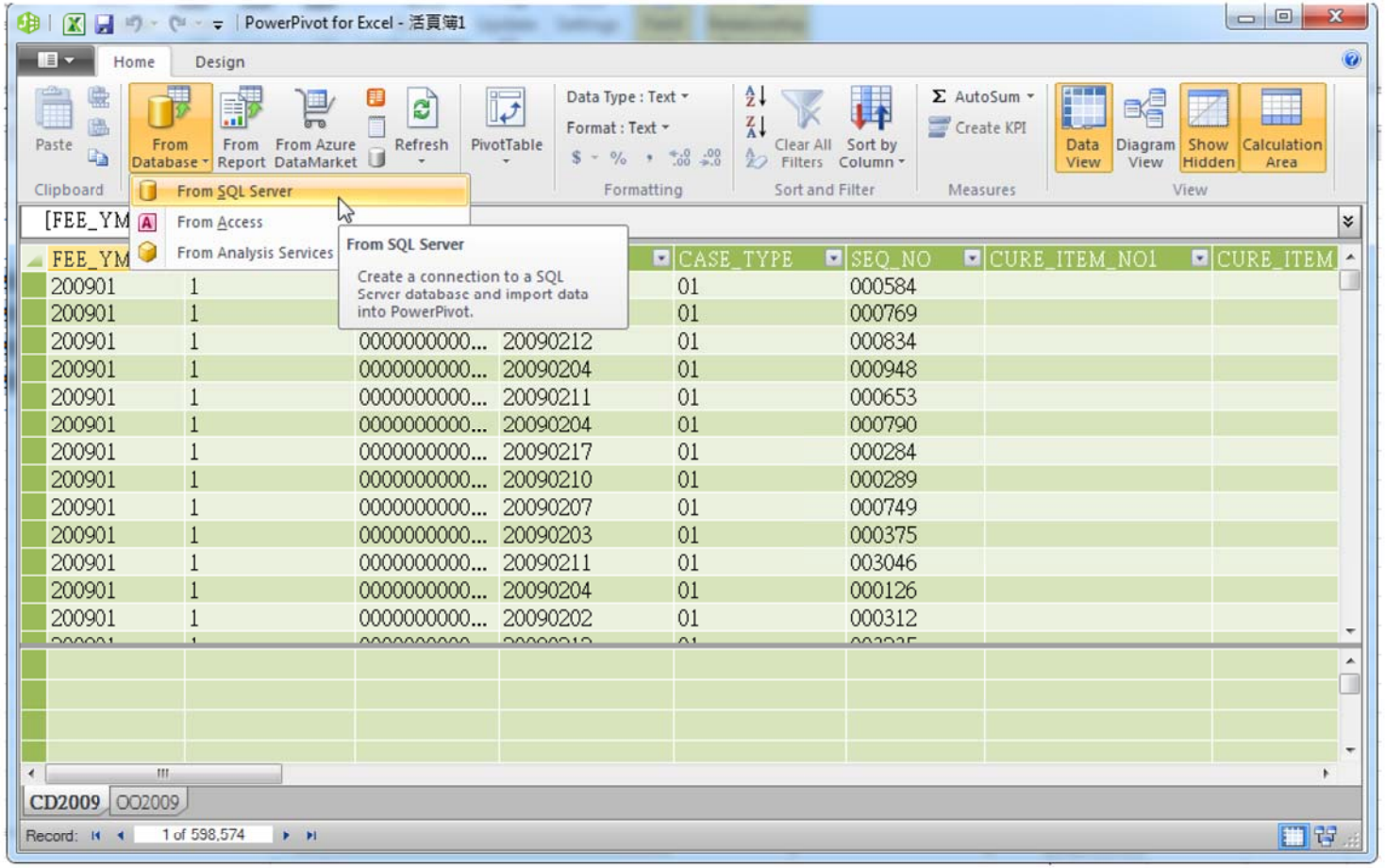
http://www.microsoft.com/zh-tw/download/details.aspx?id=29074

EXCEL視窗

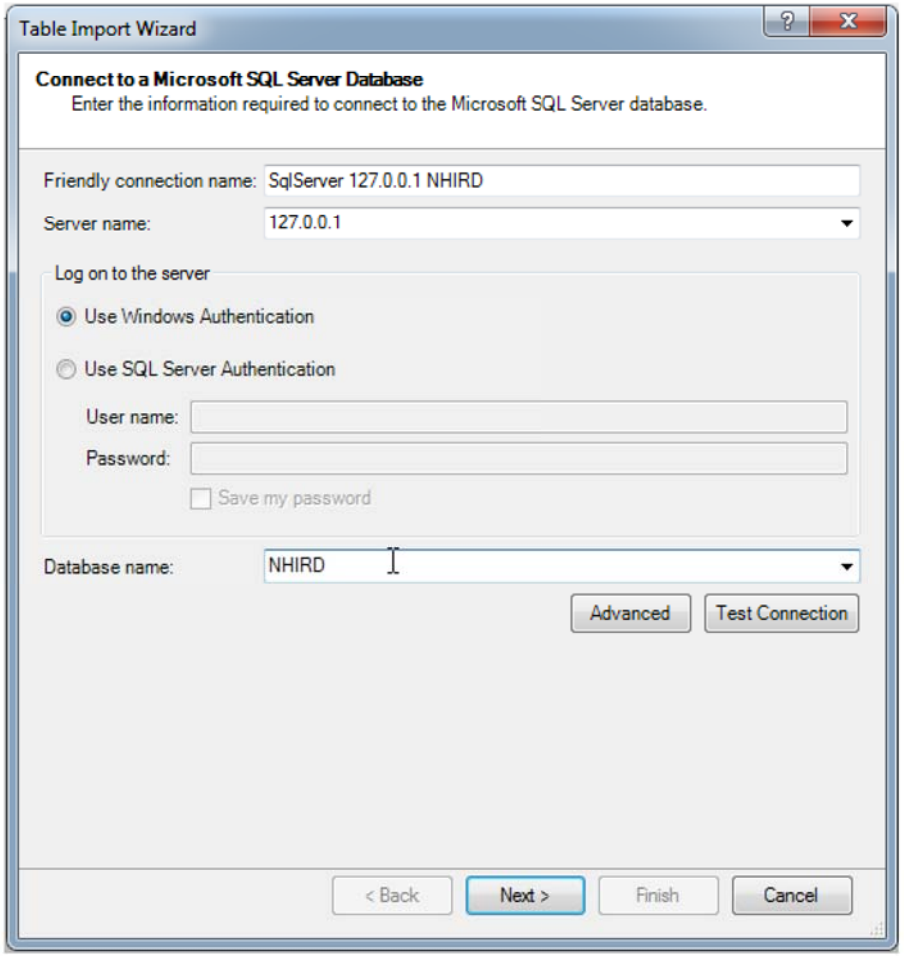
❶

❷

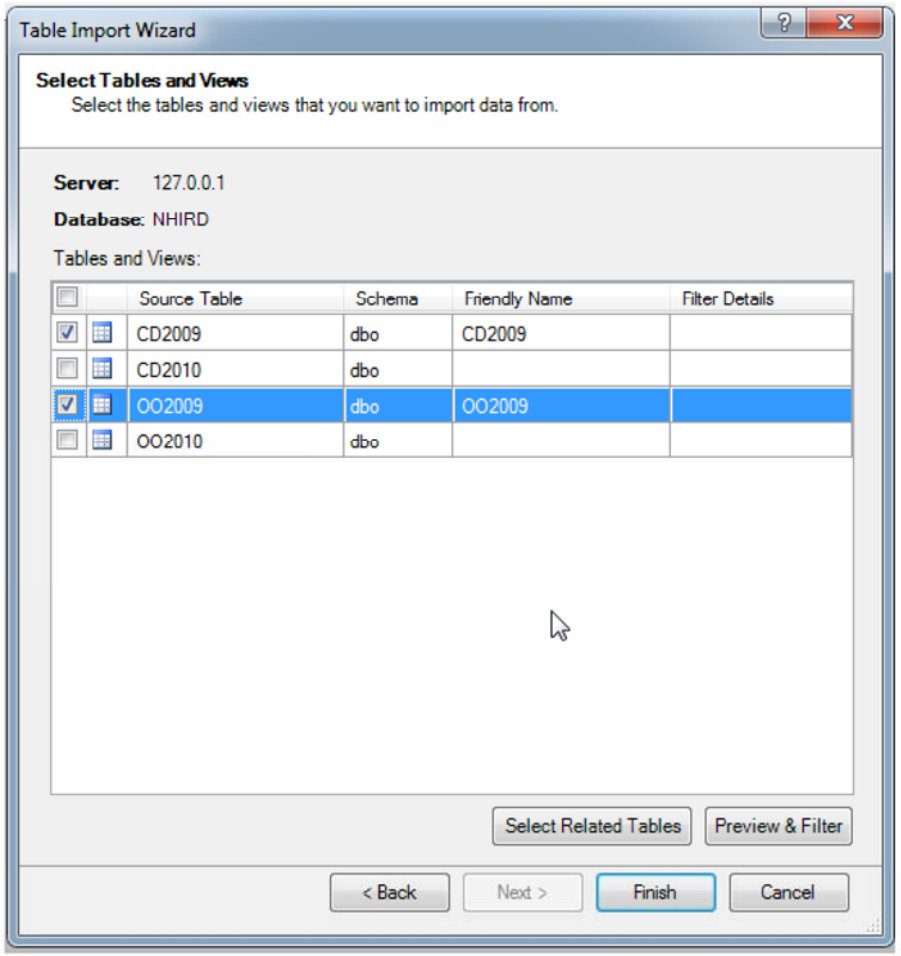
Power Pivot視窗



3 連結 NHIRD 資料庫



4 連結資料表



5 資料表關聯

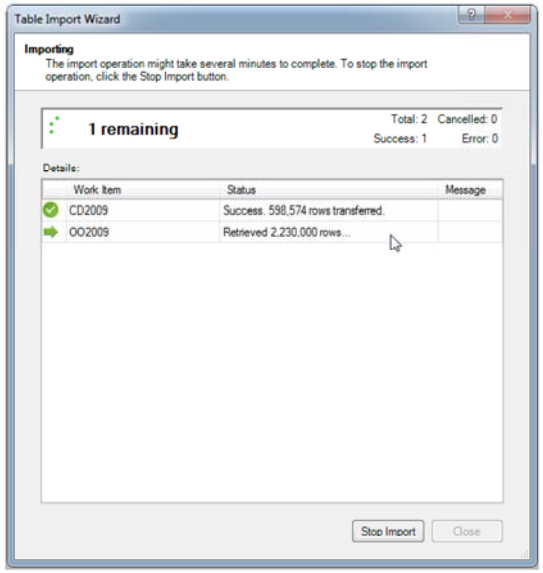
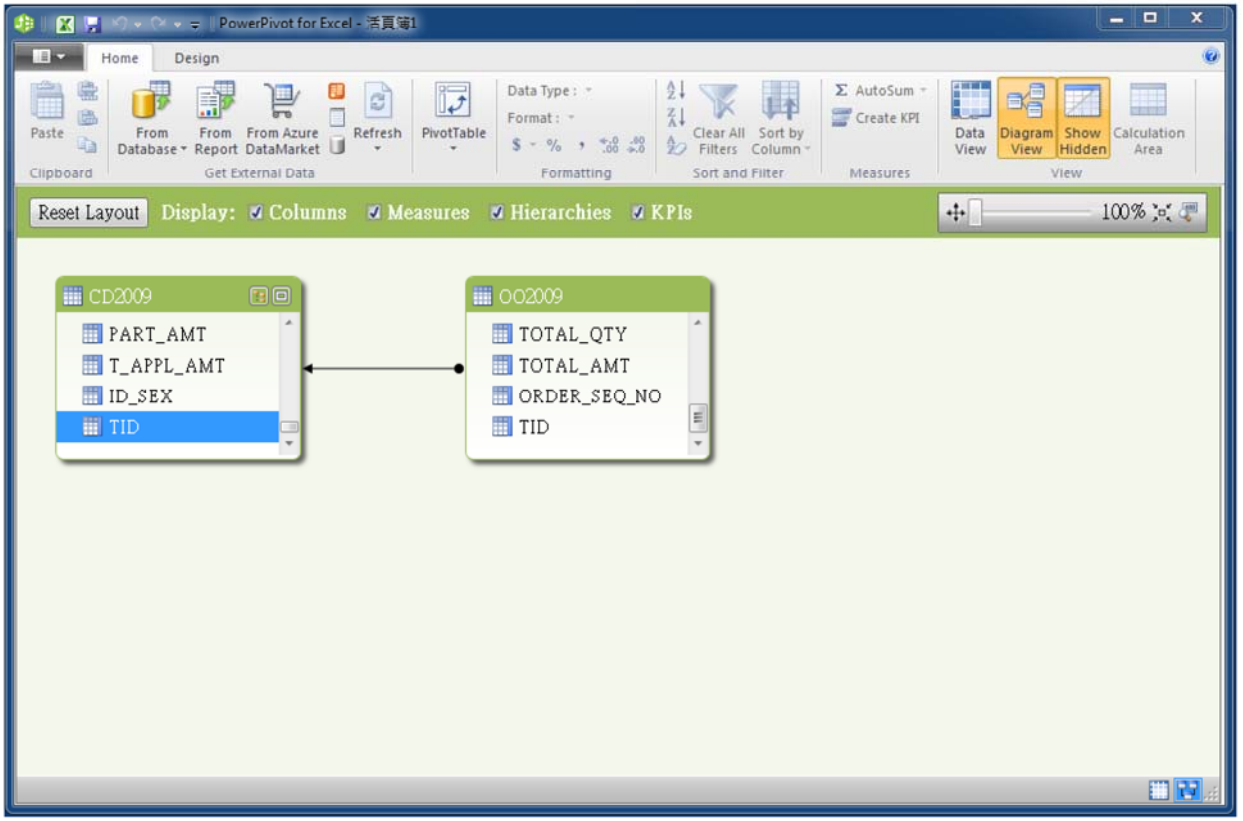


Table Import Wizard
Importing
The import operation might take several minutes to complete. To stop the import operation, click the Stop Import button.
1 remaining
Total: 2 Cancelled: 0
Success: 1 Error: 0
Details:
Work Item
Status
Message
CD2009
Success. 598,574 rows transferred.
OO2009
Retrieved 2,230,000 rows...
Stop Import
Close


PowerPivot for Excel - 活頁簿1
Home
Design
Paste
From Database
From Report
From Azure DataMarket
Refresh
PivotTable
Data Type :
Format :
Clear All Filters
Sort by Column
AutoSum
Create KPI
Data View
Diagram View
Show Hidden
Calculation Area
Clipboard
Get External Data
Formatting
Sort and Filter
Measures
View
Reset Layout
Display:
Columns
Measures
Hierarchies
KPIs
100%
CD2009
PART_AMT
T_APPL_AMT
ID_SEX
TID
OO2009
TOTAL_QTY
TOTAL_AMT
ORDER_SEQ_NO
TID

6 EXCEL 輕鬆操作大量健保資料