

健康資料管理與研究實務

衛生福利資料的研究設計與資料管理

《世代追蹤研究 Cohort study》

劉品崧 統計諮詢分析師 / 組長

花蓮慈濟醫院高齡暨社區醫學部

112年度資料管理與研究實務（下半年）

• 課程列表

日期	時間	地點	主題	軟體
10/06(五)	13：30 - 16：30	臺北醫學大學信義校區	統計軟體R與SAS在資料管理與統計分析之應用	SAS+R
10/16(一)	09：00 - 12：00	臺北醫學大學雙和校區	衛福資料庫之研究設計與統計分析：病例對照研究	SAS
10/16(一)	13：30 - 16：30	臺北醫學大學雙和校區	衛福資料庫之研究設計與統計分析：病例對照研究	R
10/20(五)	09：00 - 12：00	國家衛生研究院（苗栗）	衛福資料庫之研究設計與統計分析：世代追蹤研究	SAS
10/20(五)	13：30 - 16：30	國家衛生研究院（苗栗）	衛福資料庫之研究設計與統計分析：世代追蹤研究	R
10/28(六)	09：00 - 16：30	慈濟大學（花蓮）	衛福資料庫之研究設計與統計分析：病例對照研究	SAS
11/03(五)	09：00 - 12：00	高雄醫學大學	統計軟體R與SAS在資料管理與統計分析之應用	SAS+R
11/06(一)	09：00 - 12：00	國立成功大學	衛福資料庫之研究設計與統計分析：世代追蹤研究	SAS
11/06(一)	13：30 - 16：30	國立成功大學	衛福資料庫之研究設計與統計分析：世代追蹤研究	R

112年度資料管理與研究實務（下半年）

- 課前具備基礎
 - 軟體操作（R / SAS）、流行病學、研究設計、生物統計
- 課程設計理念
 - 思考研究設計、實際資料管理、完成統計分析
- 學習目標重點
 - 追求邏輯貫通、分享實戰經驗

課程注意事項

- 兩個承諾
 - 每50分鐘休息10分鐘，讓各位intake / output
 - 過程當中隨時可以打斷我，問題留給我，收穫你帶走
- 兩個不可以
 - 課程練習資料為模擬資料檔，不可以直接用於實際研究用途
 - 操作定義僅供教學演練使用，不可以直接用於實際研究用途

課程大綱

- 世代追蹤研究
- 研究範例說明
- 實作練習時間

世代追蹤研究 Cohort study

- 目標族群的介入組 (intervention)

Population + exposure

- 目標族群的對照組 (control)

Population + non-exposure

- 追蹤兩組未來的事件發生狀況 (outcome)

Interested event

- 探討過去暴露與事件發生的相關性 (association)

Statistical analysis

肺炎住院病人的DM病史與未來死亡事件之相關性

- 目標族群的介入組 (intervention)

肺炎患者 DM病史(+)

- 目標族群的對照組 (control)

肺炎患者 DM病史(-)

- 追蹤兩組未來的事件發生狀況 (outcome)

死亡事件

- 探討過去暴露與事件發生的相關性 (association)

發生率、KM plot、Cox regression model

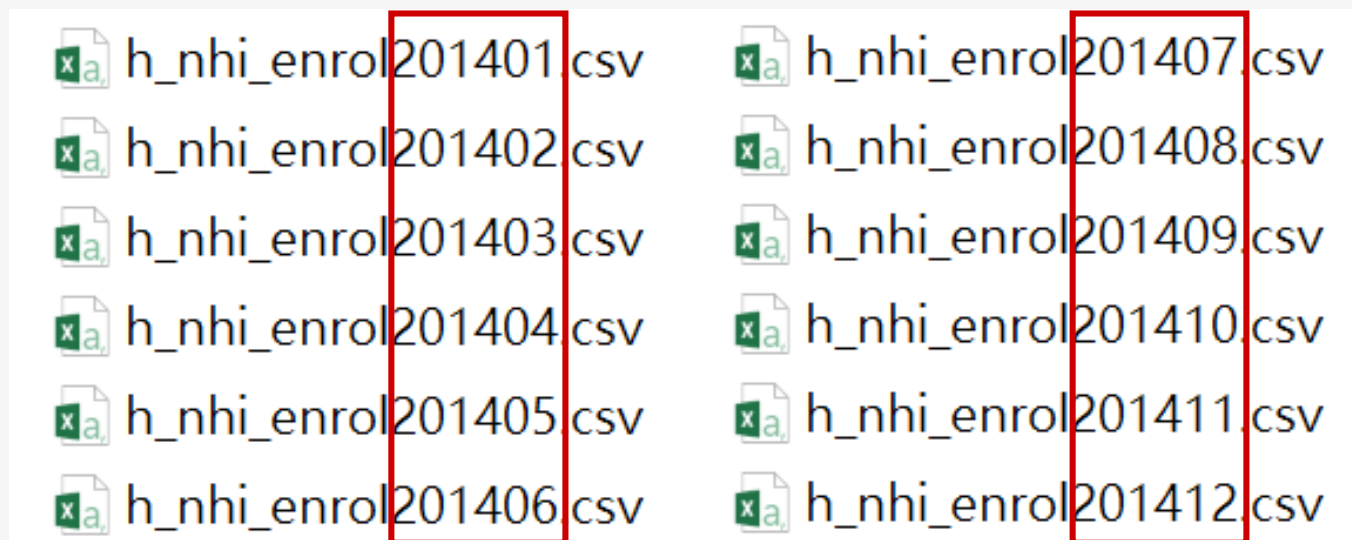
研究材料：衛生福利資料庫模擬資料檔

- 10萬人模擬資料檔，資料年度為2014年1月～12月
- 主要用途
 - 認識資料輪廓
 - 瞭解資料關聯
 - 樣本數量估算
 - 測試程式結果
- 今日主軸五大常用資料庫
 - 門診、住院、藥局、承保、死因

研究材料：衛生福利資料庫模擬資料檔

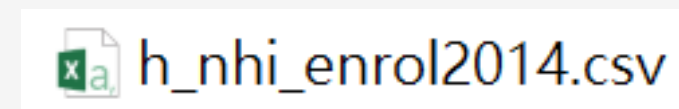
- 原始型態

- 不同月份資料分開儲存
- 實際模樣
- 需要迴圈處理



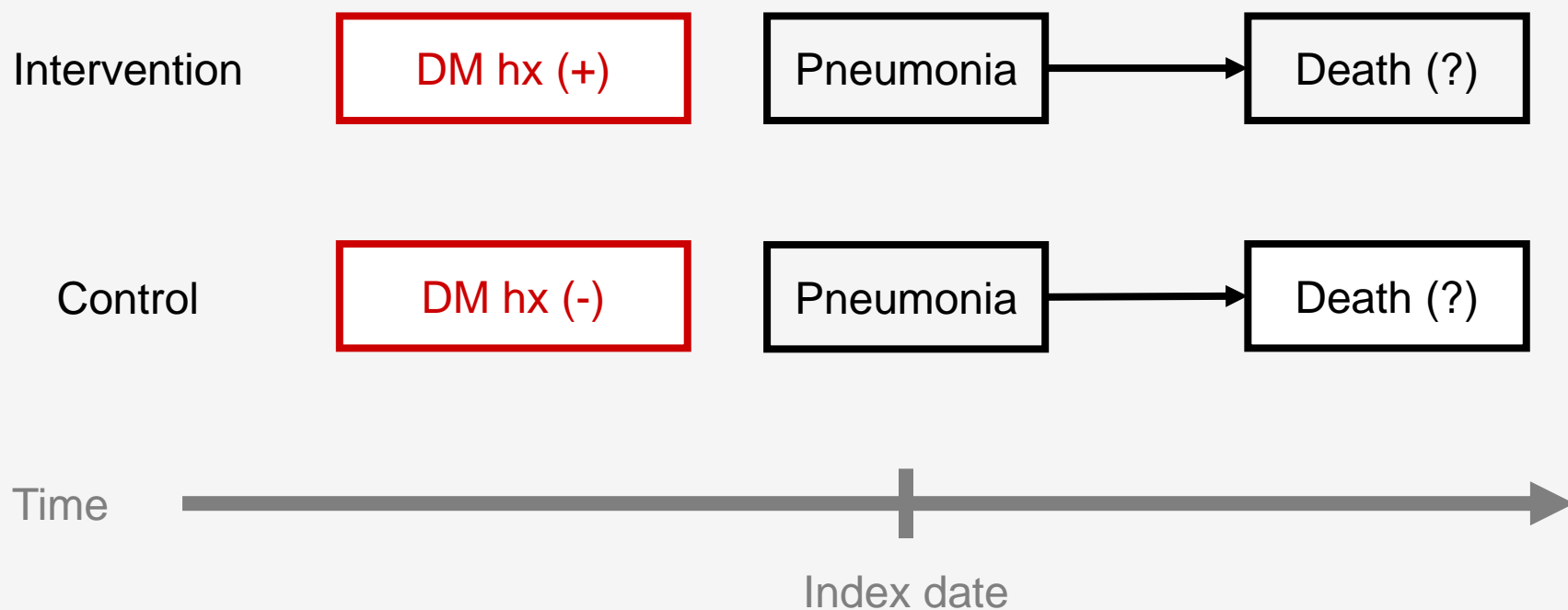
- 堆疊型態 (本日使用)

- 整個年度一起儲存
- 教學使用
- 不要執著技術層面



將核心理念進行圖像化

- 肺炎住院病人的DM病史與未來死亡事件之相關性

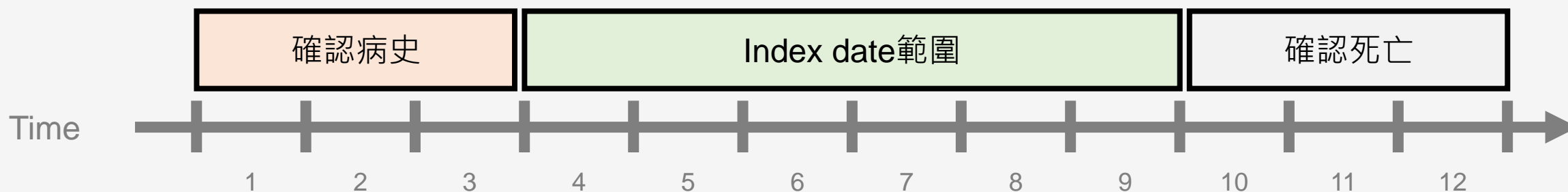


將核心理念進行拆解區塊工作項目



1. 找出研究樣本
2. 連結死亡登記，確認住院後死亡事件的發生時間
3. 撈取健保承保資料，確認研究樣本基本資料
4. 撈取健保承保資料，將投保金額級距進行分類
5. 觀察過去有無門診糖尿病藥物處方紀錄
6. 把整理好的資料儲存下來，並且合併資料
7. 統計分析

研究樣本：肺炎住院病人

- 住院費用檔有研究樣本診斷碼
 - 任一診斷碼欄位：icd9cm_1、icd9cm_2、icd9cm_3、icd9cm_4、icd9cm_5
 - ICD-9-CM：480 — 486
- 以日期排序後取出每人在2014年的首次住院
- 首次住院的日期僅限於4 — 9月
 - 過去至少有3個月可以找出病史，未來至少有3個月可以確認死亡



研究樣本：肺炎住院病人

- 使用正規表達式尋找疾病診斷編碼
- HTN
 - 原始描述 480 – 486
 - 實際樣貌 480、481、482、483、484、485、486
 - 模式開頭標記 48[0-6]
 - 表示字串模式開頭↑ ↑表示48後面的字元可以是0, 1, 2, 3, 4, 5, 6
(pattern)

終點事件：入院後死亡事件

- 從全國的死亡登記檔找出研究樣本的資料
 - 身分證號：**id**
 - 死亡日期：**d_date**
- 將 d_date (文字型態) 轉換為日期型態
- 與 index_date 相減得到追蹤時間 (event_ft)
- 判斷 index_date 起到資料觀察終點 (2014.12.31) 為止
有無發生事件 (event_occur)

基本資料：健保投保的出生年與性別

- 從全國的健保承保檔找出研究樣本的資料
 - 基本資料：`id, id_s, id_birth_y`
 - 投保月份：`prem_ym`
- 年齡：出生年度 - 住院年度，不可以 <0 歲或是 >100 歲
- 性別編碼只能是1（男性）和2（女性）
- 取投保月份最早的一筆

社經狀況：健保投保的保費

- 從全國的健保承保檔找出研究樣本的資料
 - 投保級距：**id1_amt**
 - 投保月份：**prem_ym**
 - 原始格式為yyyymm，把投保日假定為每月1日，轉換為日期格式 yyyymmdd
- 找出距離住院日期（不含）前最近一次的投保紀錄
- 投保金額級距切分
 - (1) 15,840以下
 - (2) 15,840 — 29,999
 - (3) 30,000以上

DM病史 (1 / 2)

- DM藥物健保代碼清單
 - 資料夾use的code_dmdrug.csv，所有ATC code為A10開頭的藥品清單
 - 醫令鍵值：drug_no
- 門診醫令檔 (h_nhi_opdto2014)
 - 醫令鍵值：drug_no
 - 申報串檔：hosp_id, fee_ym, appl_date, appl_type, case_type, seq_no
- 門診費用檔 (h_nhi_opdte2014)
 - 個人鍵值：id, func_date
 - 申報串檔：hosp_id, fee_ym, appl_date, appl_type, case_type, seq_no

DM病史 (2 / 2)

- 找出糖尿病藥物處方 (申報資料 & 藥物代碼的交集)
 - DM藥物健保代碼清單 ← **drug_no** → 門診醫令檔
- 取出目標樣本的就醫資料，保留住院日 (不含) 以前的就醫紀錄
 - 個人住院資訊 ← **id** → 門診費用檔， **func_date < index_date**
- 合併申報及醫令
 - 醫令 ← **hosp_id, fee_ym, appl_date, appl_type, case_type, seq_no** → 費用
- 個人歸戶
 - 過去只要有任何DM藥物都算是DM病史

資訊合併

- 將整理好的眾多資料依照病人身分證號 (id) 進行合併
 - 研究族群：pop
 - 基本資料：idfile
 - 健保承保：nhi
 - DM病史：dm
- 如果有遺漏值，記得要補 0 或是對應的虛擬值 !!!
 - 純粹遺漏未補的話整筆觀察值會被排除於統計分析之外

樣本納入與排除條件處理

- 排除條件
 - 基本資料不齊全者
 - 死亡日期早於住院日
 - 住院時年齡 <0 歲 或是 >100 歲

描述性統計分析：Table 1

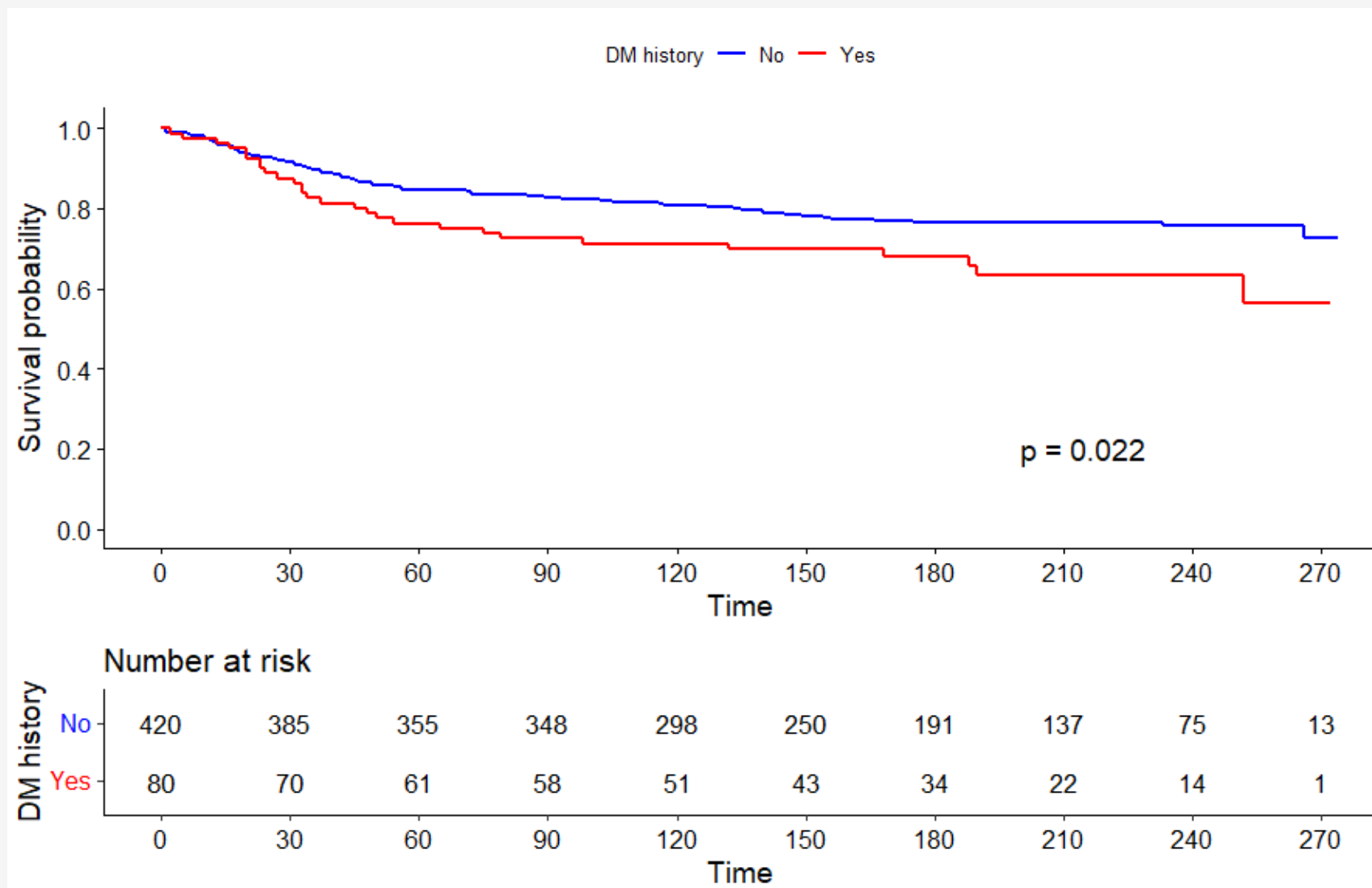
- 以DM病史（dm）為分組
- 比較兩組的特性差異
 - 性別、年齡、投保金額類別
- 使用Standardized mean differences（SMD）量化差異的大小
 - 通常使用 $SMD > 0.1$ 判斷兩組之間有差異存在
 - SMD在大數據時代較不容易因為N值擴大而產生統計顯著

描述性統計分析：Table 1

Variable	DM history				SMD
	No		Yes		
	N = 420		N = 80		
	N	%	N	%	
Age*	49.8	34.2	74.1	12.9	0.939
Sex					
Male	245	58.3	43	53.8	0.091
Female	175	41.7	37	46.2	0.091
NHI premium range					
Dependent	119	28.3	29	36.2	0.170
15,840–29,999	191	45.5	38	47.5	0.040
Above 30,000	110	26.2	13	16.2	0.247

* Expressed in mean and standard deviation.

存活狀態：KM Curves

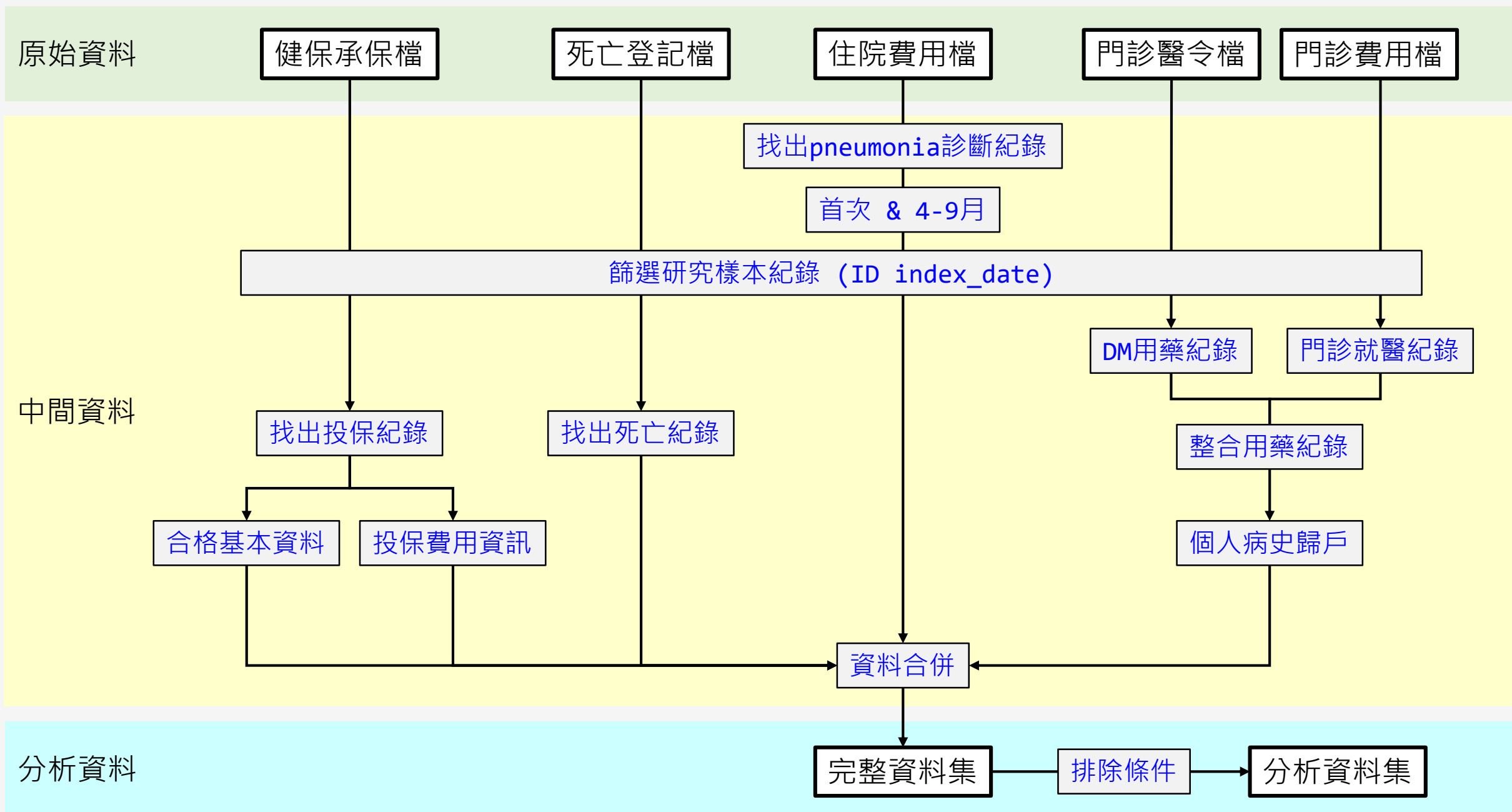


推論性統計分析：Cox regression model

- Model
 - Crude (univariable) model
 - Y = 死亡事件發生
 - X = DM病史
 - Adjusted (multivariable) model
 - Y = 死亡事件發生
 - X = DM病史 + Age + Male + NHI
- Hazard ratio & 95% confidence intervals of **DM病史**

推論性統計分析：Cox regression model

Variable	Crude model				Adjusted model			
	HR	95% CI		<i>p</i>	aHR	95% CI		<i>p</i>
DM history	1.63	1.07	2.48	0.023	1.03	0.67	1.57	0.895
Age					1.04	1.03	1.05	<.001
Male					1.31	0.90	1.90	0.156
NHI premium range								
Dependent					1.00	(Reference)		
15,840—29,999					0.94	0.63	1.40	0.765
Above 30,000					1.00	0.60	1.68	0.994



統計軟體實際操作練習

- 檔案資料夾
 - 原始模擬資料：raw
 - 健保承保檔：h_nhi_enrol2014
 - 住院費用檔：h_nhi_ipdte2014
 - 門診費用檔：h_nhi_opdte2014
 - 門診醫令檔：h_nhi_opdto2014
 - 死亡登記檔：h_ost_death2014
 - 完成整理資料：use
- 將程式碼挖空的部分填入正確的內容
- R程式碼以UTF-8編碼儲存，先調整RStudio設定，開啟才不會是亂碼

Summary

- 核心理念
- 工作心流
- 技術實踐
- 知識獲取問ChatGPT
- 系統訓練找小劉老師
- 開放提問時間
- 劉品崧
- Peter Pin-Sung Liu
- psliu520@gmail.com
- <https://github.com/PSLiu/>



109年度R基礎課程-劉品崧老師

