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

《統計軟體R與SAS在統計分析之應用》

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

- 在開始分析之前
- 基礎統計分析方法
  - 降血壓藥物隨機分派試驗
  - 感興趣的結果變數（連續 / 數值 / 類別 / 名義）
- 模擬試驗存活分析
  - 抗凝血藥物選擇與未來糖尿病併發症風險
  - 使用傾向分數配對處理干擾因子

資料的產生：真實世界

- 2月18日深夜

- 一名8歲男性兒童由父母帶入急診

- CRIES分數為6分

- 經診斷為急性闌尾炎（acute appendicitis）

CRIES, Children Revised Impact of Event Scale

資料的儲存：樣態與編碼

- 結構化資料表（data table）
- 譯碼簿（codebook）

欄 / column / 變項 / variable

列 / row / 觀察值 / observation

id	date	male	age	pain	diagnosis
S1911	02-18	1	8	6	K35
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

變項名稱	中文意義	資料類型	編碼方式
id	身分證號	文字	S+四位數字
date	就醫日期	日期	mm-dd
male	男性	數值	1 = 男性；0 = 女性
age	年齡	數值	單位：歲
pain	疼痛指數	數值	CRIES量表分數
diagnosis	主診斷	文字	ICD-10-CM編碼

降血壓藥物隨機分派試驗

- 對象 Population
  - 高血壓病人
- 介入 Intervention
  - 新開發藥物
- 對照 Control
  - 現行指引最佳藥物
- 結果 Outcome
  - 血壓下降
- 設計 Study
  - RCT

模擬資料編碼說明(1)基本資料、分組

	caseid	txgp	male	age	agegp
1	1	1	0	65	2
2	2	0	0	62	2
3	3	1	1	63	2
4	4	1	0	64	2
5	5	1	1	61	2
6	6	0	0	63	2

變項	中文意義	類型	編碼方式
caseid	收案流水號	數字	隨機亂數流水號
txgp	分派組別	數值	1 = 新藥；0 = 現行最佳治療
male	男性	數值	1 = 男性；0 = 女性
age	年齡	數值	單位：歲
agegp	年齡分組	數值	1 = 60以下；2 = 60-69； 3 = 70以上

模擬資料編碼說明(2)前後測結果

	caseid	sbp_pre	dbp_pre	qol_scale_pre	qol_bad_pre	sbp_post	dbp_post	qol_scale_post	qol_bad_post
1	1	130	86	4	0	115	80	2	0
2	2	131	86	3	0	130	88	2	0
3	3	132	84	3	0	125	81	2	0
4	4	129	84	3	0	116	78	1	0
5	5	129	86	4	0	116	86	2	0
6	6	129	84	3	0	130	86	3	0

變項	中文意義	類型	編碼方式
sbp_pre / post	收縮壓前測 / 後測	數值	單位：mm-Hg
dbp_pre / post	舒張壓前測 / 後測	數值	單位：mm-Hg
qol_scale_pre / post	生活品質量表	數值	單位：1~10分，越大越不好
qol_bad_pre / post	生活品質量表測量為不佳（5分以上）	數值	1 = 是；0 = 否

模擬資料編碼說明(3)後續不良反應監測

	caseid	sae_ft	sae_occur	sae_ft_5y	sae_count
1	1	365	0	765	0
2	2	365	0	1825	0
3	3	365	0	1825	0
4	4	365	0	1825	0
5	5	365	0	1825	0
6	6	365	0	1825	0

變項	中文意義	類型	編碼方式
sae_ft	一年內追蹤SAE時間	數值	單位：天
sae_occur	一年內追蹤SAE是否發生	數值	1 = 是；0 = 否
sae_ft_5y	五年內追蹤SAE時間	數值	單位：天
sae_count	五年內追蹤SAE發生次數	數值	單位：次數

依據你最感興趣的變數（Y）分為

- 數值型態
  - Mean、SD、Person's *r*、Box-plot
  - *t*-test、ANOVA
  - Linear regression
- 類別型態
  - N、Percent
  - 交叉表、χ^2test、Fisher exact test
  - Logistic regression

AF & DM的病人使用口服抗凝血劑對未來併發症有影響？

Ann Intern Med. 2022 Apr;175(4):490-498. doi: 10.7326/M21-3498. Epub 2022 Feb 15.

Diabetes-Related Complications and Mortality in Patients With Atrial Fibrillation Receiving Different Oral Anticoagulants : A Nationwide Analysis

Huei-Kai Huang [1], Peter Pin-Sung Liu [2], Shu-Man Lin [3], Jin-Yi Hsu [4], Jih-I Yeh [5], Edward Chia-Cheng Lai [6], Carol Chiung-Hui Peng [7], Kashif M Munir [8], Ching-Hui Loh [4], Yu-Kang Tu [9]

Affiliations + expand

PMID: 35157495 DOI: 10.7326/M21-3498

定義PICOS

- 對象 Population
  - Patients with AF & DM
- 介入 Intervention
  - NOAC
- 對照 Control
  - Warfarin
- 結果 Outcome
  - DM complications
- 設計 Study
  - Cohort study

Ann Intern Med (2022) doi: 10.7326/M21-3498

模擬資料編碼說明(1)基本資料、分組與指標日期

	id	male	age	oacs	noac	index_date
1	S00010	0	94	noac	1	2012-05-08
2	S00014	1	74	warf	0	2012-11-14
3	S00017	1	78	noac	1	2012-08-30
4	S00045	0	59	warf	0	2012-12-06
5	S00046	1	77	warf	0	2012-06-09

變項	中文意義	類型	編碼方式
id	身分證號	文字	S+5位數字
male	男性	數值	1 = 男性；0 = 女性
age	年齡	數值	單位：歲
oacs	使用抗凝血劑類型	文字	noac = NOAC warf = warfarin
noac	使用NOAC (虛擬變數)	數值	1 = 是；0 = 否
index_date	指標日期	日期	YYYY-MM-DD

模擬資料編碼說明(2)事件發生的日期與追蹤時間

	id	index_date	event_occur	event_date	event_ft
1	S00010	2012-05-08	1	2016-12-15	4.605065024
2	S00014	2012-11-14	0	2019-11-14	6.997946612
3	S00017	2012-08-30	0	2019-08-30	6.997946612
4	S00045	2012-12-06	0	2019-12-06	6.997946612
5	S00046	2012-06-09	1	2014-11-07	2.412046543

變項	中文意義	類型	編碼方式
event_occur	觀察期間內發生事件	數值	1 = 是；0 = 否
event_date	觀察期間內發生事件之日期	日期	YYYY-MM-DD
event_ft	觀察期間內追蹤時間	數值	單位：年

模擬資料編碼說明(3)指標日期分組與虛擬變數

	id	index_year_gp	year_2012_2013	year_2014_2015	year_2016_2017
1	S00010	2012_2013	1	0	0
2	S00014	2012_2013	1	0	0
3	S00017	2012_2013	1	0	0
4	S00045	2012_2013	1	0	0
5	S00046	2012_2013	1	0	0

變項	中文意義	類型	編碼方式
index_year_gp	指標年份分組	數值	2012_2013 = 2012 - 2013 2014_2015 = 2014 - 2015 2016_2017 = 2016 - 2017
year_2012_2013	指標年份2012 - 2013 (虛擬變數)	數值	1 = 是；0 = 否
year_2014_2015	指標年份2014 - 2015 (虛擬變數)	數值	1 = 是；0 = 否
year_2016_2017	指標年份2016 - 2017 (虛擬變數)	數值	1 = 是；0 = 否

模擬資料編碼說明(4)其他干擾因子

	id	c2vs	hyperlipidemia	ckd	cancer
1	S00010	6	0	1	1
2	S00014	0	0	0	0
3	S00017	0	0	0	0
4	S00045	2	1	1	0
5	S00046	2	0	0	0

變項	中文意義	類型	編碼方式
c2vs	CHA_2DS_2-VASc Score	數值	單位：分數，範圍：0 - 9分
hyperlipidemia	高血脂病史	數值	1 = 是；0 = 否
ckd	慢性腎臟病病史	數值	1 = 是；0 = 否
cancer	癌症病史	數值	1 = 是；0 = 否

背景特質比較

Table 1. Baseline characteristics of patients

Variable	NOAC N = 6,916		Warfarin N = 3,335		SMD
	N	(%)	N	(%)	
Male	3,602	52.1	1,848	55.4	0.067
Age*	73.61	9.73	68.56	11.36	0.477
Index year group					0.846
2012-2013	849	12.3	1,504	45.1	
2014-2015	2,444	35.3	1,063	31.9	
2016-2017	3,623	52.4	768	23.0	
CHA_2DS_2-VASc Score	2.02	1.42	2.02	1.41	0.001
Hyperlipidemia	2,435	35.2	1,164	34.9	0.006
CKD	815	11.8	492	14.8	0.088
Cancer	365	5.3	176	5.3	0.001

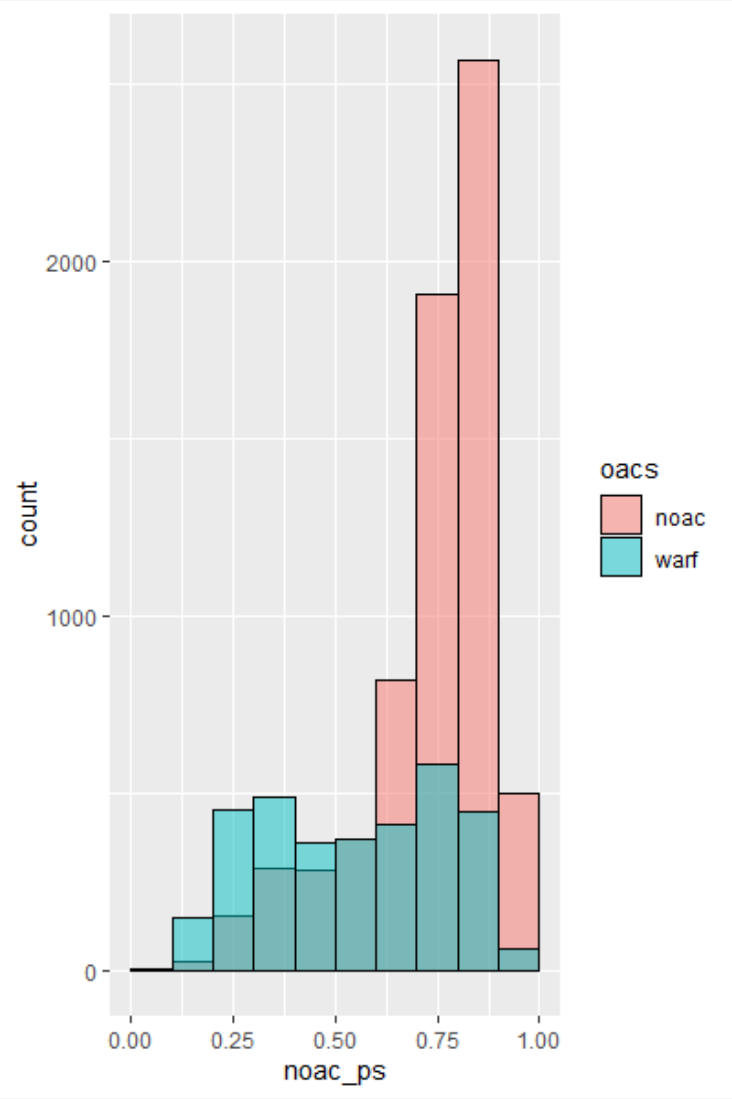
* Expressed as mean and SD.

Abbreviations: n, number; SD, standard deviation; CKD, chronic kidney disease

計算與評估傾向分數（Propensity score, PS）

- Logistic regression model 使用藥物 ~ 性別+年齡
  - $ln(\frac{P(NOAC=1)}{P(NOAC=0)}) =$

    $-3.87 + (-0.10 * male) + (0.04 * age)$
- Propensity score
  - 65歲男性
  - $\hat{p} = \frac{1}{1+e^{(-(-3.87+\beta_{male}X_{male}+\beta_{age}X_{age}))}}$
  - $\hat{p} = \frac{1}{1+e^{(-(-3.87+(-0.10*1)+(0.04*65)))}} = 0.2026$



應用PS使樣本背景特質相近（干擾因子與分組獨立）

Table 1. Baseline characteristics of patients (original population)

Variable	NOAC N = 6,916		Warfarin N = 3,335		SMD
	N	(%)	N	(%)	
Male	3,602	52.1	1,848	55.4	0.067
Age*	73.61	9.73	68.56	11.36	0.477
Index year group					0.846
2012-2013	849	12.3	1,504	45.1	
2014-2015	2,444	35.3	1,063	31.9	
2016-2017	3,623	52.4	768	23.0	
CHA_2DS_2-VASc Score	2.02	1.42	2.02	1.41	0.001
Hyperlipidemia	2,435	35.2	1,164	34.9	0.006
CKD	815	11.8	492	14.8	0.088
Cancer	365	5.3	176	5.3	0.001

* Expressed as mean and SD.

Abbreviations: n, number; SD, standard deviation; CKD, chronic kidney disease

Table 1. Baseline characteristics of patients (matched-population)

Variable	NOAC N = 2,728		Warfarin N = 2,728		SMD
	N	(%)	N	(%)	
Male	1,297	47.5	1,448	53.1	0.111
Age*	72.41	11.65	70.29	11.17	0.185
Index year group					0.160
2012-2013	842	30.9	907	33.2	
2014-2015	918	33.7	1,053	38.6	
2016-2017	968	35.5	768	28.2	
CHA_2DS_2-VASc Score	2.34	1.51	2.01	1.41	0.222
Hyperlipidemia	1,162	42.6	953	34.9	0.158
CKD	540	19.8	356	13	0.183
Cancer	241	8.8	142	5.2	0.142

* Expressed as mean and SD.

Abbreviations: n, number; SD, standard deviation; CKD, chronic kidney disease

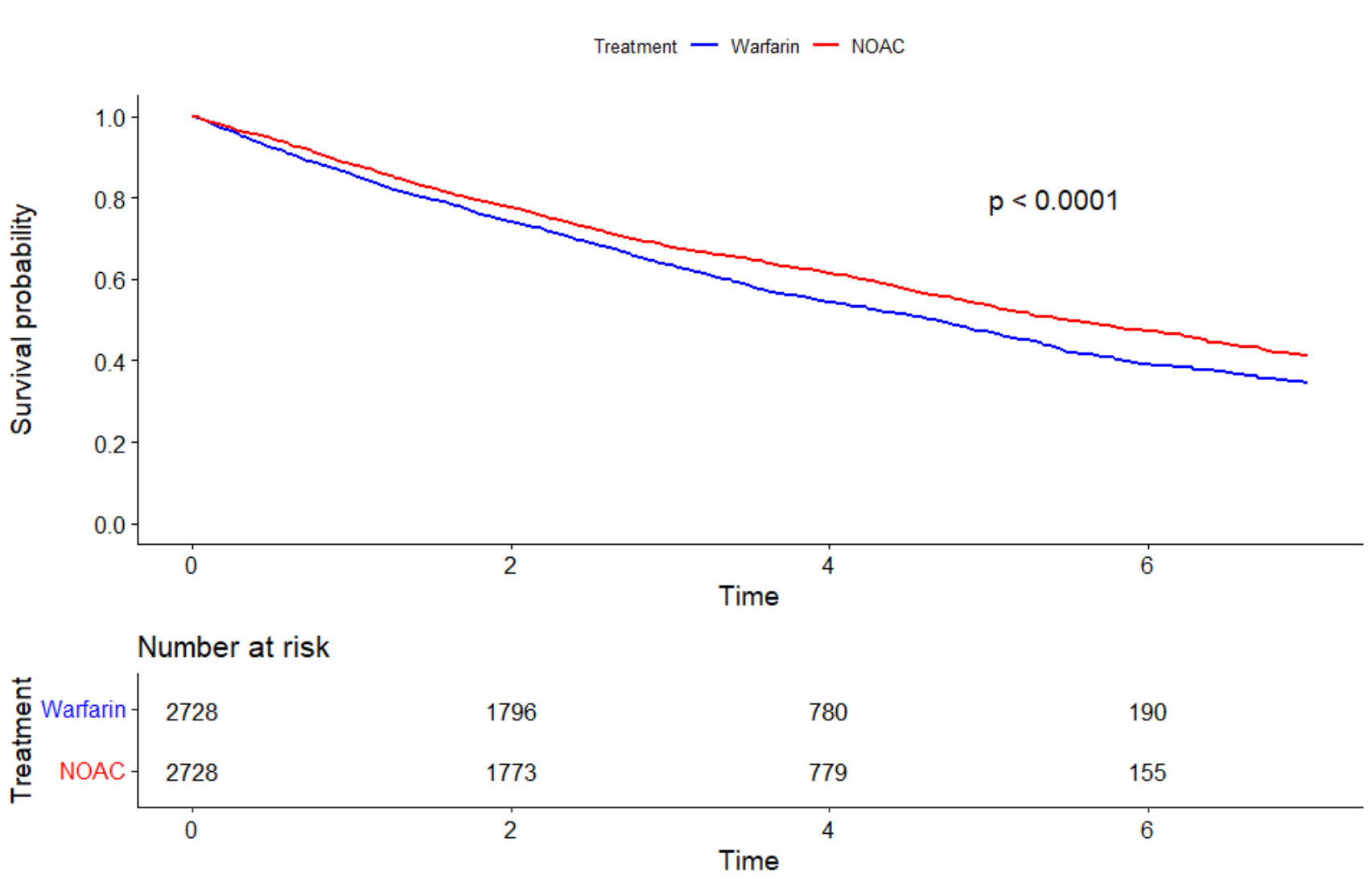
觀察期間中風事件之發生率比較與治療效果

Table 2. Risk of macrovascular complication

	N	Events	FU	IR	aHR (95% CI)	p value
NOAC	2,728	1,125	8,929	125.9	0.82 (0.75-0.90)	<.0001
Warfarin	2,728	1,342	8,785	152.7	1.00 (reference)	

Abbreviations: n, number; FU, follow-up time (years); IR, incidence rate per 1,000 person years;
aHR, adjusted hazard ratio; CI, confidence intervals.

觀察期間中風事件之KM curves比較



實作時間

- SAS軟體
- R軟體

- 有問題隨時舉手！
- 有問題隨時舉手！
- 有問題隨時舉手！

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/