

113年度衛生福利資料科學中心統計軟體推廣課程

2024.10.01 (二) 09:00-12:00 @ 台北醫學大學 信義校區杏春樓 電腦暨語文教室

# 健康資料研究與R軟體

## 《資料管理進階班》

劉品崧 統計諮詢暨分析師

花蓮慈濟醫院健康長壽中心

# 課程大綱

- 課程目標
- 議題設定
  - 研究目的
  - 執行計畫
- 進階資料管理技術實作
  - 函數設計
  - 迴圈撰寫
  - 報表製作
  - 綜合應用
  - 實作練習

# 研究工作者的技術提升

- 課程受眾
  - 以健保資料庫為研究材料 / 日常工作者
  - 已經有使用過R軟體經驗 ( `data.table`, `tableone`, `glm` )
  - 已經有研究方法和生物統計背景者
- 學習目標
  - 將常用功能自訂函數 ( `function` )
  - 用迴圈大量操作資料與運用函數
  - 從複雜的報表中將需要的統計結果整理成想要的樣子
  - 綜合應用

# 議題設定

- 研究目的
- 執行計畫

# 研究議題背景參考文獻

## Original Investigation

November 20, 2023

# Risk of Bleeding Following Non-Vitamin K Antagonist Oral Anticoagulant Use in Patients With Acute Ischemic Stroke Treated With Alteplase

Tou-Yuan Tsai, MD<sup>1,2,3</sup>; Yu-Chang Liu, MD<sup>4,5</sup>; Wan-Ting Huang, MS<sup>6</sup>; Yu-Kang Tu, PhD<sup>3,7</sup>; Shang-Quan Qiu, MD<sup>8</sup>; Sameer Noor, BS<sup>9</sup>; Yong-Chen Huang, MS<sup>3</sup>; Eric H. Chou, MD<sup>10</sup>; Edward Chia-Cheng Lai, PhD<sup>11</sup>; Huei-Kai Huang, MD<sup>2,3,12,13</sup>

» Author Affiliations

*JAMA Intern Med.* 2024;184(1):37-45. doi:10.1001/jamainternmed.2023.6160


# 研究議題設定（今日課程使用）


- 探討缺血性中風住院患者之院內出血事件與過去服用抗凝血劑相關性
  - 研究設計：病例對照研究（case-control study）
  - 納入條件：2014年住院檔主診斷（[icd9cm\\_1](#)）為缺血性中風（433.x、434.x）的患者
  - 排除條件：非今年首次住院、2014 / 04 / 01（不含）以前的住院、承保基本資料缺失
  - 結果事件：指標住院同時在其他疾病診斷代碼欄位（[icd9cm\\_2-5](#)）有出血性疾病診斷
    - 疾病診斷代碼參考Tsai（2024）*JAMA Intern Med* eTable 1
  - 暴露因子：住院日（不含）前是否在門診有任何抗凝血劑或抗血小板藥物處方
    - Warfarin（B01AA03）、NOAC（B01AE07、B01AF01、B01AF02、B01AF03）
    - Antiplatelet（B01AC04、B01AC06、B01AC30），參考[temp/druglist.csv](#)
  - 干擾因子：年齡、性別
  - 統計分析：Logistic regression，另外依據性別（男、女）、年齡（65歲上下）分層分析


# 衛生福利模擬資料檔 ( 虛擬10萬人2014年資料 )


- 原始型態 ( 本日使用 ) base資料夾
  - 不同月份資料分開儲存
  - 實際模樣
  - 需要迴圈處理


**Warning !!!!!!!**  
**教學使用虛擬資料**  
**不得作為研究素材**


 h\_nhi\_enrol201401.csv


 h\_nhi\_enrol201402.csv


 h\_nhi\_enrol201403.csv


 h\_nhi\_enrol201404.csv


 h\_nhi\_enrol201405.csv


 h\_nhi\_enrol201406.csv


 h\_nhi\_enrol201407.csv

 h\_nhi\_enrol201408.csv

 h\_nhi\_enrol201409.csv

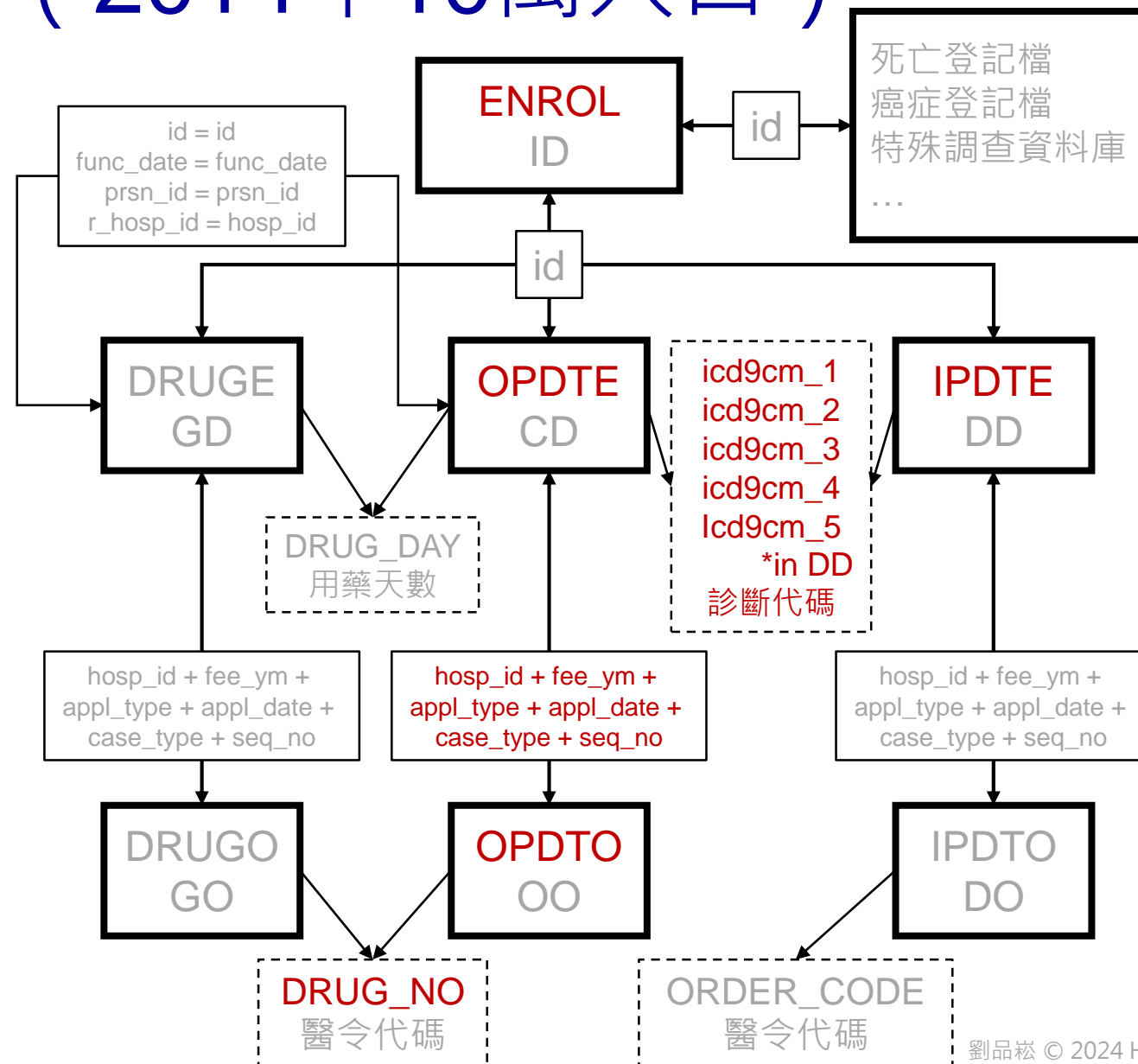
 h\_nhi\_enrol201410.csv

 h\_nhi\_enrol201411.csv

 h\_nhi\_enrol201412.csv

# 衛生福利資料模擬檔 ( 2014年10萬人口 )

- 預設各位已經熟悉檔名與架構
- 實際使用範圍
  - 門診費用檔 ( OPDTE )
  - 門診醫令檔 ( OPDTO )
  - 住院費用檔 ( IPDTE )
  - 住院醫令檔 ( IPDTO )
  - 藥局費用檔 ( DRUGE )
  - 藥局醫令檔 ( DRUGO )
  - 健保承保檔 ( ENROL )
  - 死因統計檔 ( DEATH )





# 進階資料管理技術實作

- 函數設計
- 迴圈撰寫
- 報表製作
- 綜合應用
- 實作練習

# 函數 ( function ) 動手練習

# 建立函數

```
upper_name_fn <- function(invar){
```

```
  # 未來放入函數的物件代稱為invar
```

```
  outvar <- toupper(invar)
```

```
  # 選定要將哪一個函數中的產物輸出
```

```
  return(outvar)
```

```
}
```

# 使用函數

```
new_char <- upper_name_fn("liu")
```

```
print(new_char)
```

```
"LIU"
```

# 函數 ( function ) 動手練習

# 建立清單

```
in_name_list <- list("taiwan", "japan", "korea")
```

# 輸入清單給函數操作

```
new_char_list <- lapply(in_name_list, upper_name_fn)  
print(new_char_list)
```

```
[[1]]           # 清單項目的第1項  
[1] "TAIWAN"     # 清單項目的第1項內容的第一個物件/第一行資料 ... etc
```

```
[[2]]           # 清單項目的第2項  
[1] "JAPAN"       # 清單項目的第2項內容的第一個物件/第一行資料 ... etc
```

```
[[3]]           # 清單項目的第3項  
[1] "KOREA"       # 清單項目的第3項內容的第一個物件/第一行資料 ... etc
```

# 迴圈 ( for loop ) 動手練習 & 進入正題

# 將迴圈參數做為內容

```
name_country <- c("taiwan", "japan", "korea")
```

```
for (ii in name_country) {  
  print(ii)  
}
```

```
"taiwan"    "japan"    "korea"
```

# 將迴圈參數作為指標

```
name_city <- c("taipei", "tokyo", "seoul")
```

```
for (ii in 1:3) {  
  print(paste(name_country[ii], name_city[ii]))  
}
```

```
"taiwan taipei"    "japan tokyo"    "korea seoul"
```

# 迴歸模型報表架構與資料整理

## # 迴歸分析

```
tab_2_model <- glm(bleeding ~ drug_use + age + male, data = masterfile,  
                   family = binomial("logit"))
```

## # 報表摘要

```
summary(tab_2_model)
```

## # 迴歸係數(beta)與標準誤(SE)

### Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.35117	1.88593	-1.777	0.0756 .
drug_use	-0.69613	0.60583	-1.149	0.2505
age	0.02444	0.02392	1.022	0.3070
male	-0.58562	0.61068	-0.959	0.3376

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 迴歸模型報表架構與資料整理

# 確認報表物件的架構

```
str(summary(tab_2_model))
```

```
List of 17
```

```
$ call : glm(formula = bleeding ~ drug_use + age + male,  
             family = binomial("logit"), data = masterfile)
```

```
.  
. .
```

```
$ coefficients : num [1:4, 1:4] -3.3512 -0.6961 0.0244 -0.5856 1.8859 ...
```

```
..- attr(*, "dimnames")=List of 2
```

```
.. ..$ : chr [1:4] "(Intercept)" "drug_use" "age" "male"
```

```
.. ..$ : chr [1:4] "Estimate" "Std. Error" "z value" "Pr(>|z|)"
```

# 迴歸模型報表架構與資料整理

# 擷取係數、標準誤和p值

```
tab_2_coefs <- summary(tab_2_model)$coefficients  
print(tab_2_coefs)
```

# 擷取變項名稱

```
tab_2_names <- rownames(tab_2_coefs)  
print(tab_2_names)
```

# 組合成為報表

```
tab_2 <- data.table(  
  vars = tab_2_names,  
  or_point = exp(tab_2_coefs[, "Estimate"]),  
  or_lower = exp(tab_2_coefs[, "Estimate"] - 1.96 * tab_2_coefs[, "Std. Error"]),  
  or_upper = exp(tab_2_coefs[, "Estimate"] + 1.96 * tab_2_coefs[, "Std. Error"]),  
  p_value = tab_2_coefs[, "Pr(>|z|)"]  
)
```

# 分層分析報表統整

# 自訂分析function

```
tab_3_fn <- function (indt) {  
  ...  
}
```

# 將分層分析所需要的資料以清單的方式疊在一起

```
tab_3_list <- list(  
  data[male == 1], data[male == 0], data[age < 65], data[age >= 65]  
)
```

# 將分層分析結果以清單的方式疊在一起進行批次清單處理

```
tab_3 <- lapply(tab_3_list, tab_3_fn)
```

# 將分析結果整理成為清單當中的data.table並使用rbindlist進行堆疊

```
tab_3 <- rbindlist(tab_3, use.names = T, fill = T)
```



# 實作練習說明

1. 請參考[data-management-technical-programming-ST.R](#)進行練習
2. 講義因排版需求有簡化指令，以老師的程式碼為主
3. 將所有挖空的部分修改為正確的指令  
大量動手的過程，一定要專心
4. 如果有落後可以隨時喊停  
如果有能力可以隨時超前

# Summary

- 議題設定
  - 研究目的
  - 執行計畫
- 進階資料管理技術實作
  - 函數設計
  - 迴圈撰寫
  - 報表製作
  - 綜合應用
  - 實作練習
- 開放提問時間
- FB : 劉品崧
- PubMed : Peter Pin-Sung Liu
- Mail : [psliu520@gmail.com](mailto:psliu520@gmail.com)
- Open source : <https://github.com/PSLiu/>