家常便飯
家腸病患

腸江後浪推前浪

710533113 統碩一 賀慶軒
710533114 統碩一 陳彥霖
710633101 統碩零 鄭琪珍

大綱

1 緒論

(1) 研究動機
(2) 研究目的
(3) 研究流程

2 資料相關說明

(1) 資料來源
(2) 資料變數
(3) 輔助資料

3 資料處理

(1) 變數摘要
(2) 探索性資料分析

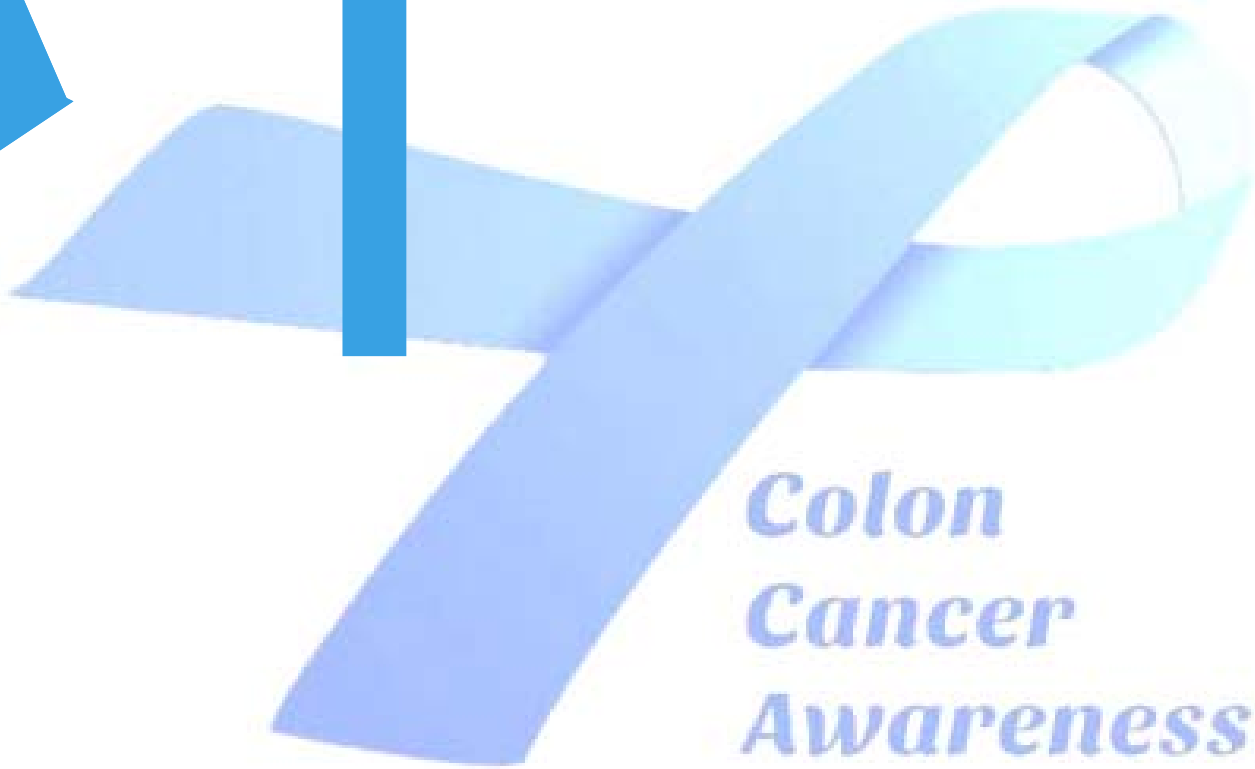
4 資料分析

(1) 統計方法
(2) 模型配適與選擇
(3) 分析解釋

5 結論

(1) 結果與貢獻
(2) 分析的限制

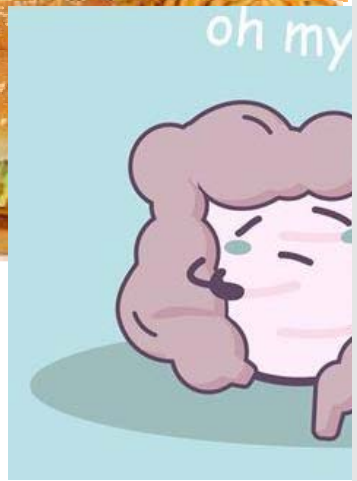
緒論 1

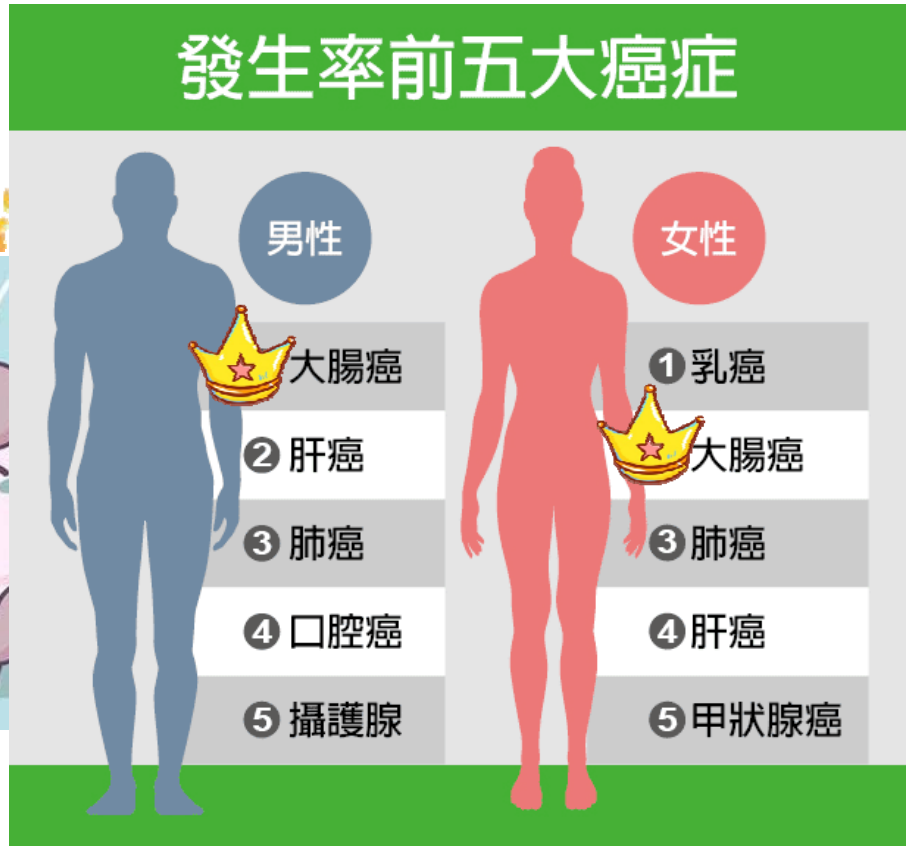


1

緒論

研究動機


oh my

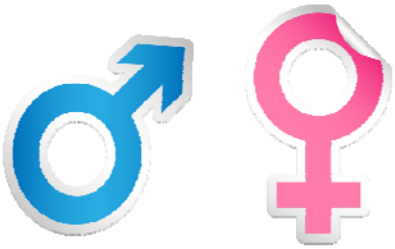

發生率前五大癌症
男性
大腸癌
❷ 肝癌
❸ 肺癌
❹ 口腔癌
❺ 攝護腺
女性
❶ 乳癌
大腸癌
❸ 肺癌
❹ 肝癌
❺ 甲狀腺癌

1

緒論

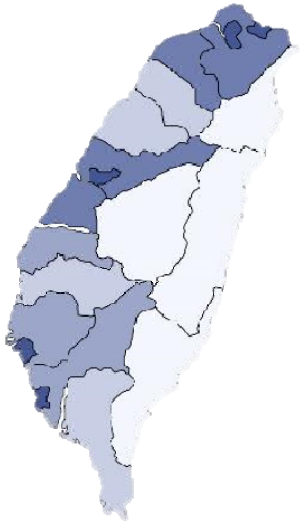
研究目的

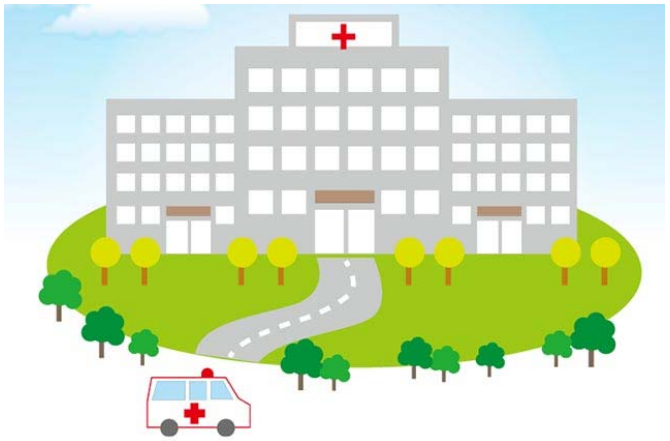
性別



年齡

地區



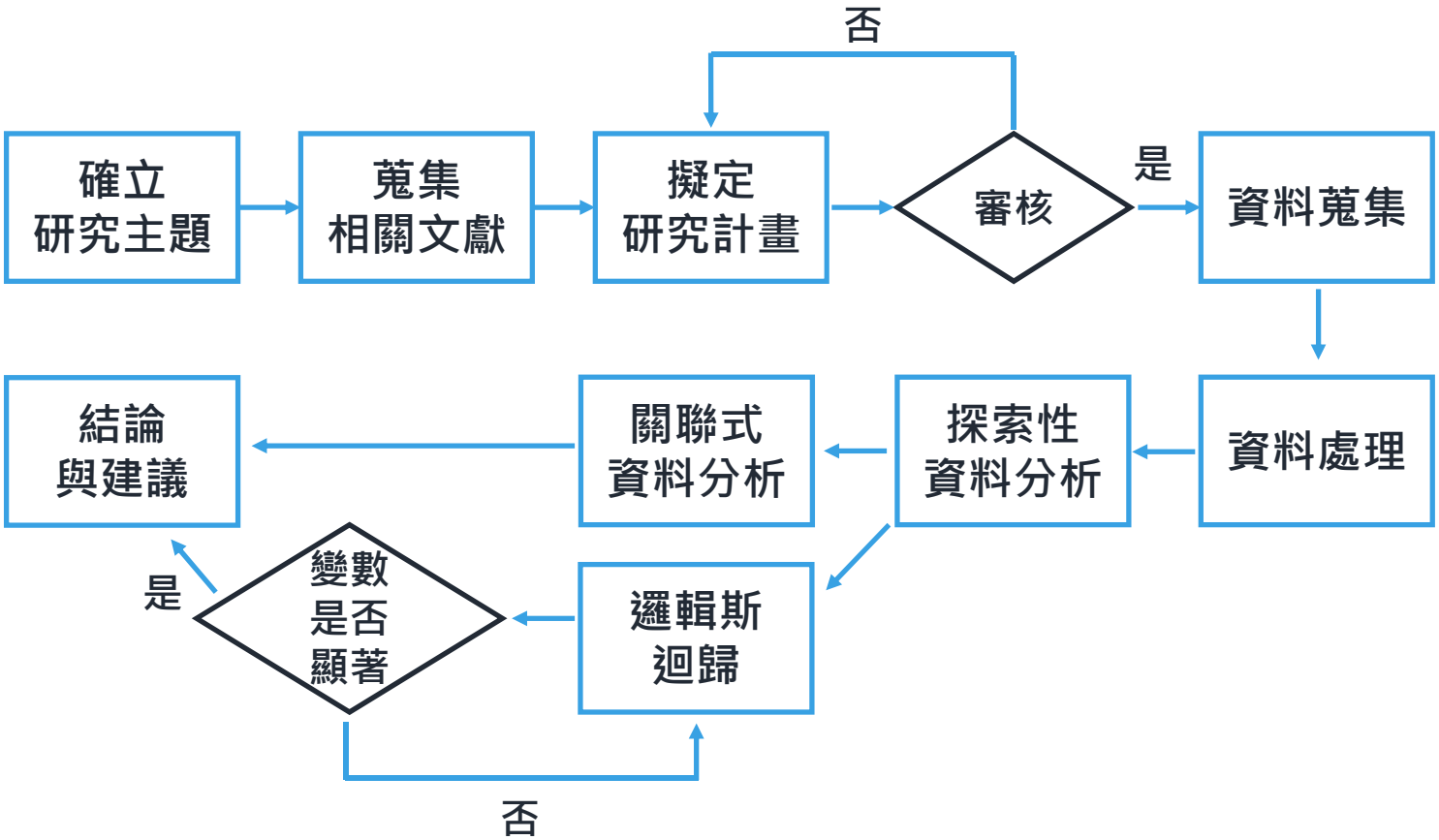


就醫科別

1

緒論

研究流程


確立
研究主題
蒐集
相關文獻
擬定
研究計畫
審核
否
是
資料蒐集
資料處理
探索性
資料分析
關聯式
資料分析
結論
與建議
邏輯斯
迴歸
變數
是否
顯著
是
否

資料相關說明

2

2

資料相關說明

資料來源

2013 年度台灣全民健保門診處方及治療明細檔之百萬歸人檔。

	費用年月	特定治療項目代號（一）	特			身份證統一編號
1	201301	09				a2215f0dec8359037a62d68
2	201301				02	60f4a112a2215f0dec8359037a62d68
3	201301				02	60f4a112a2215f0dec8359037a62d68
4	201301			D0	06	16c64fc8cc35738f76c713b7d0e8f91f
5	201301				06	16c64fc8cc35738f76c713b7d0e8f91f
6	201301			D0	06	16c64fc8cc35738f76c713b7d0e8f91f
7	201301			D0	06	16c64fc8cc35738f76c713b7d0e8f91f
8	201301	02			01	16c64fc8cc35738f76c713b7d0e8f91f
9	201301	66			08	16c64fc8cc35738f76c713b7d0e8f91f
10	201301				01	0362b02b6dd60c67c085e7d2003a32
11	201301					
12	201301	A6				
13	201301					
14	201301					
15	201301					
16	201301					
17	201301					
18	201301	35				
19	201301					
20	201301	09	02	47	04	d5cdb14f41d008d63ed564ac4f0dc18
21	201301	11			10	d5cdb14f41d008d63ed564ac4f0dc18
22	201301	15			07	17f437aae0d82cf1bdf3cc8a69296b98
23	201301				09	17f437aae0d82cf1bdf3cc8a69296b98
24	201301				01	d287666d4490ff9b7c2342c38ab337d
25	201301				01	d287666d4490ff9b7c2342c38ab337d
26	201301				01	d287666d4490ff9b7c2342c38ab337d
27	201301				01	d287666d4490ff9b7c2342c38ab337d
28	201301				01	d287666d4490ff9b7c2342c38ab337d
29	201301	47			13	afdf3c3c7967bf08befe1be50b0d0696
30	201301				01	afdf3c3c7967bf08befe1be50b0d0696

資料筆數為 15,831,525 筆
共有 904,404 人。

2

資料相關說明

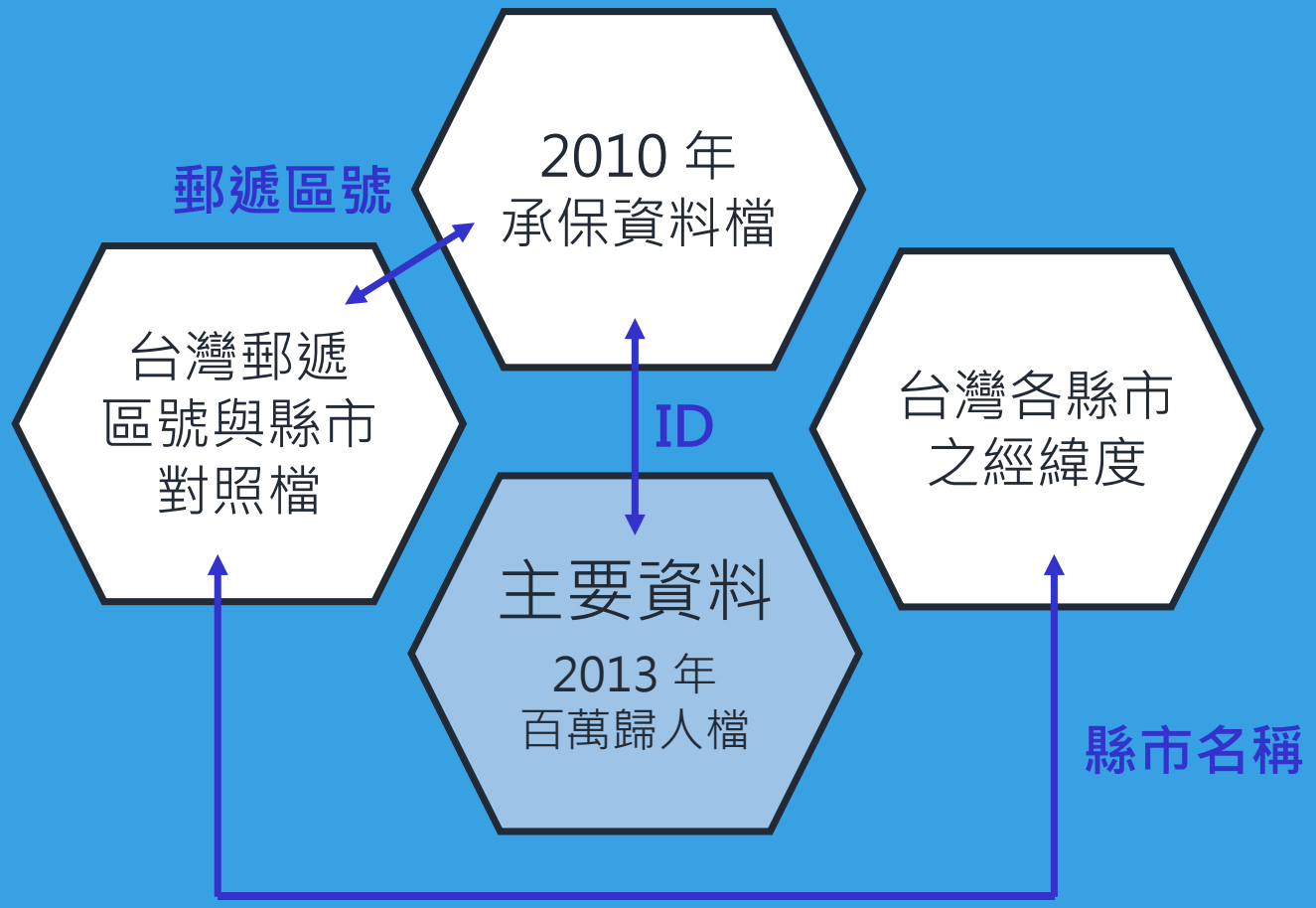
資料變數

英文欄位	中文欄位	形態
FEE_YM	費用年月	int
CURE_ITEM	特定治療項目代號	chr
FUNC_TYPE	就醫科別	chr
FUNC_DATE	就醫日期	int
ID_BIRTHDAY	出生日期	int
ID	身分證字號	chr
ACODE_ICD9	國際疾病分類號	chr
T_AMT	合計點數	int
ID_SEX	性別	chr

2

資料相關說明

輔助資料



資料處理

3

3

資料處理

資料處理

NA

遺失值處理

特定治療項目代號、國際疾病分類號、主手術代碼可為空白值，其餘的變數皆未遺失。

X

刪除異常

1. 同一身分證編號卻對應到多個出生年月或縣市。
2. 年齡小於3 歲的資料。

+

新增變數

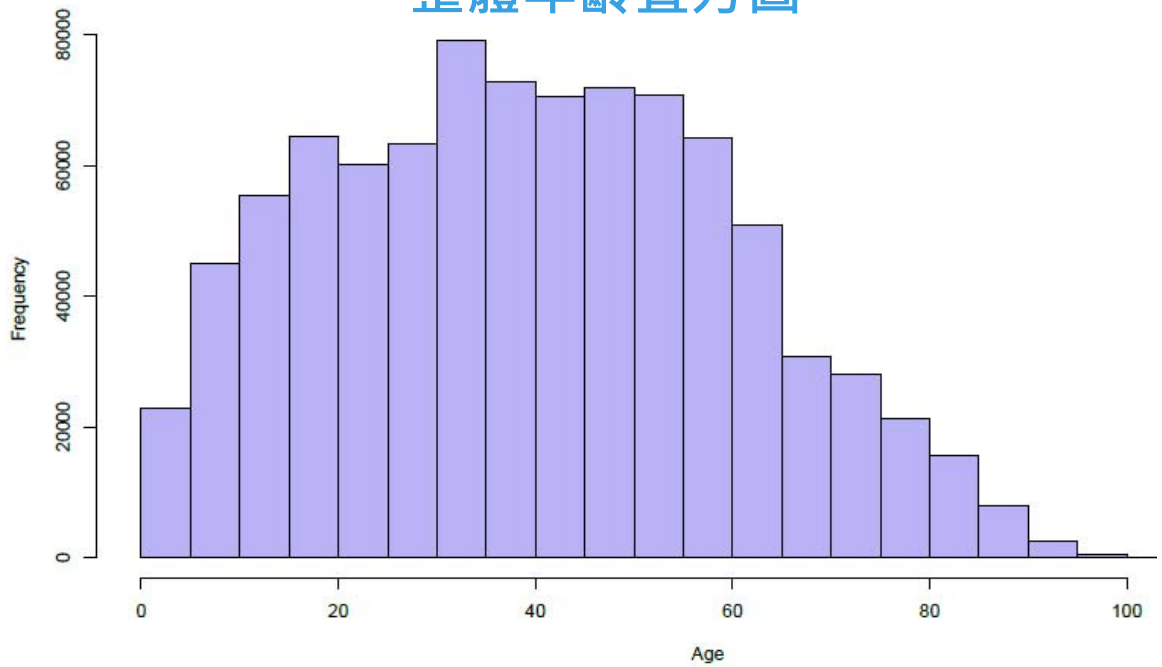
新增變數年齡、區域、是否罹患大腸直腸癌的指標、總點數。

3

資料處理

探索性分析 (EDA)

整體年齡直方圖



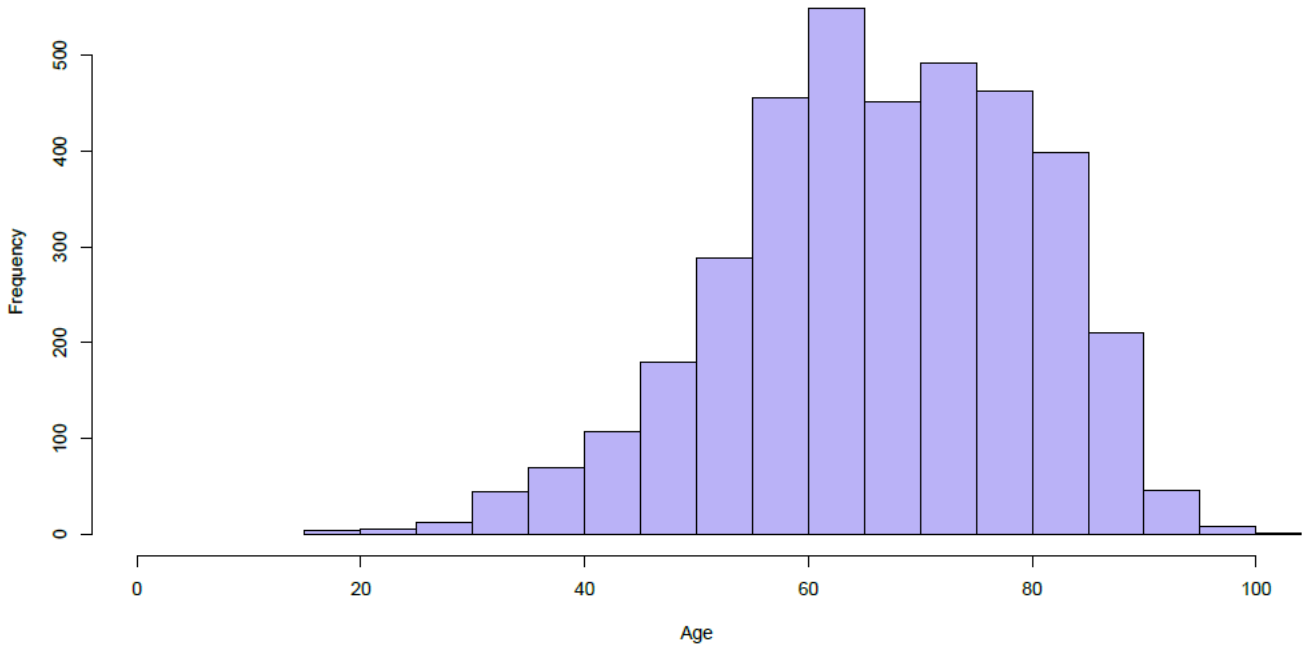
最小值	Q1	中位數	Q3	最大值	平均數	標準差
3	24	39	55	113	40.21	20.67

3

資料處理

探索性分析 (EDA)

罹患大腸直腸癌的患者之年齡直方圖

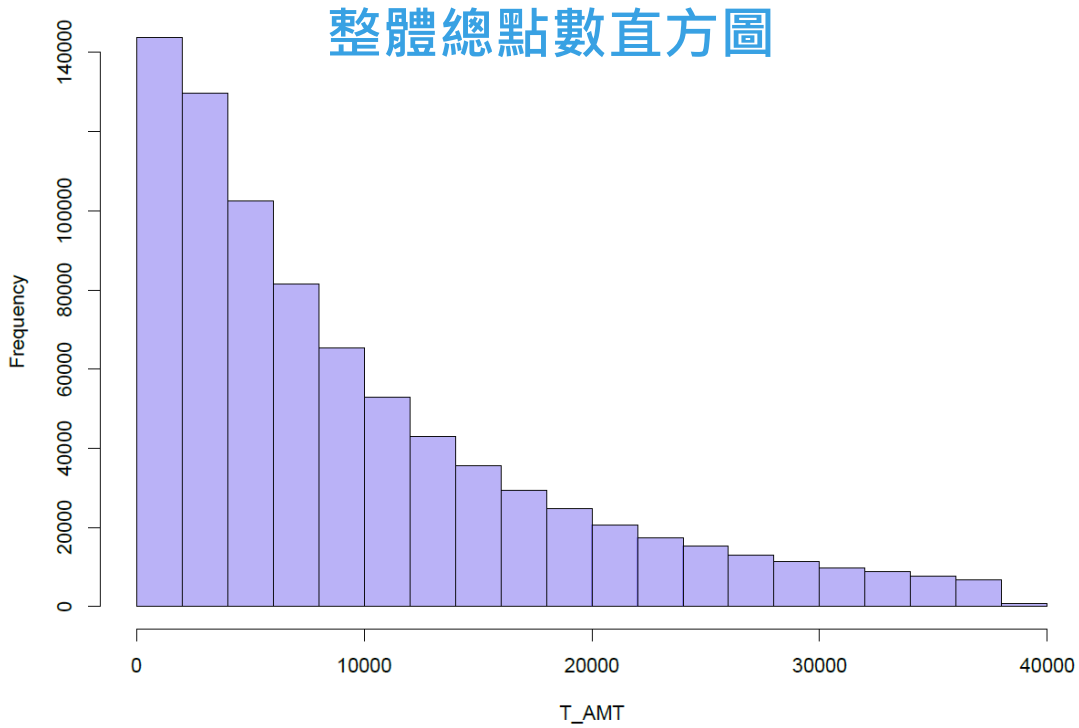


最小值	Q1	中位數	Q3	最大值	平均數	標準差
16	58	67	77	104	66.98	13.47

3

資料處理

探索性分析 (EDA)



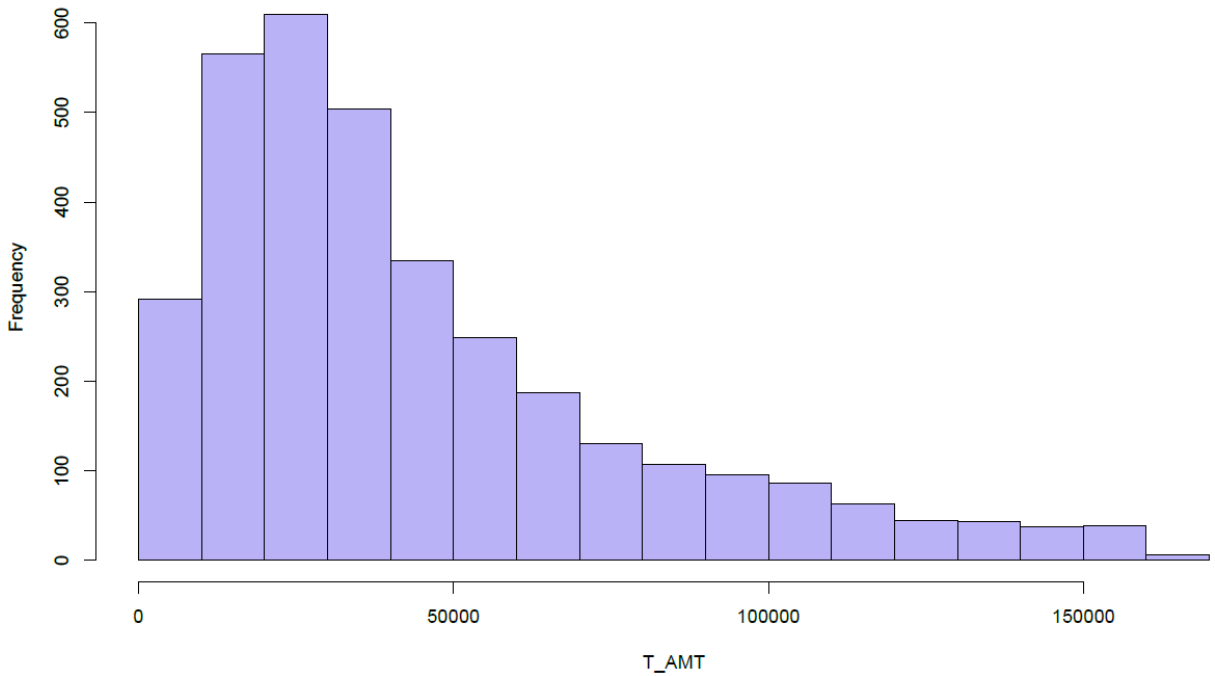
最小值	Q1	中位數	Q3	最大值	平均數	標準差
0	3,184	7,753	17,211	3,854,9181	18,024	81,267.38

3

資料處理

探索性分析 (EDA)

罹患大腸直腸癌的患者之總點數直方圖



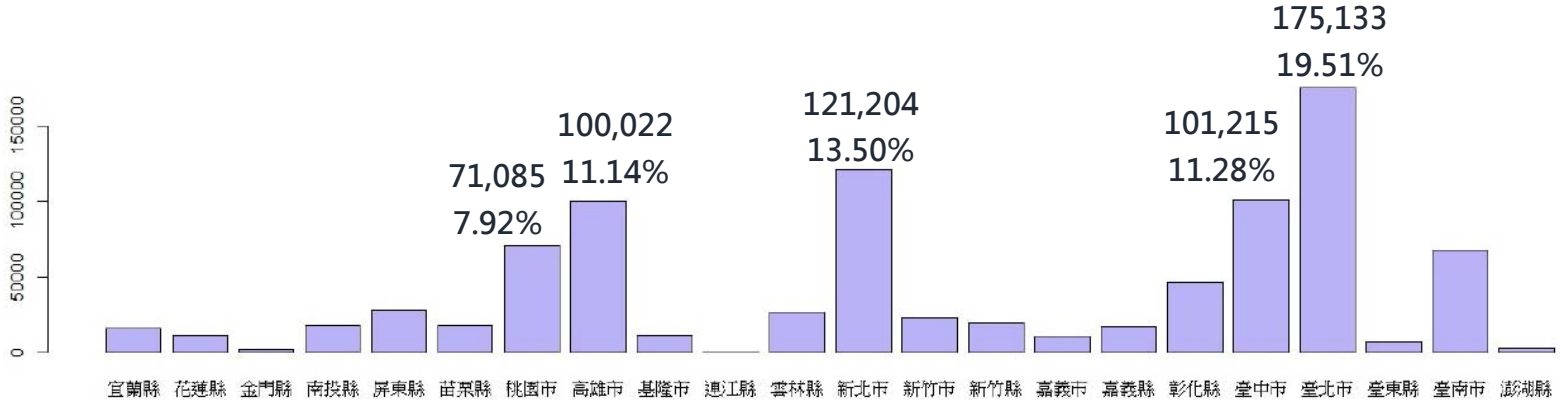
最小值	Q1	中位數	Q3	最大值	平均數	標準差
391	21,350	38,430	77,800	1,385,000	81,090	132,585.4

3

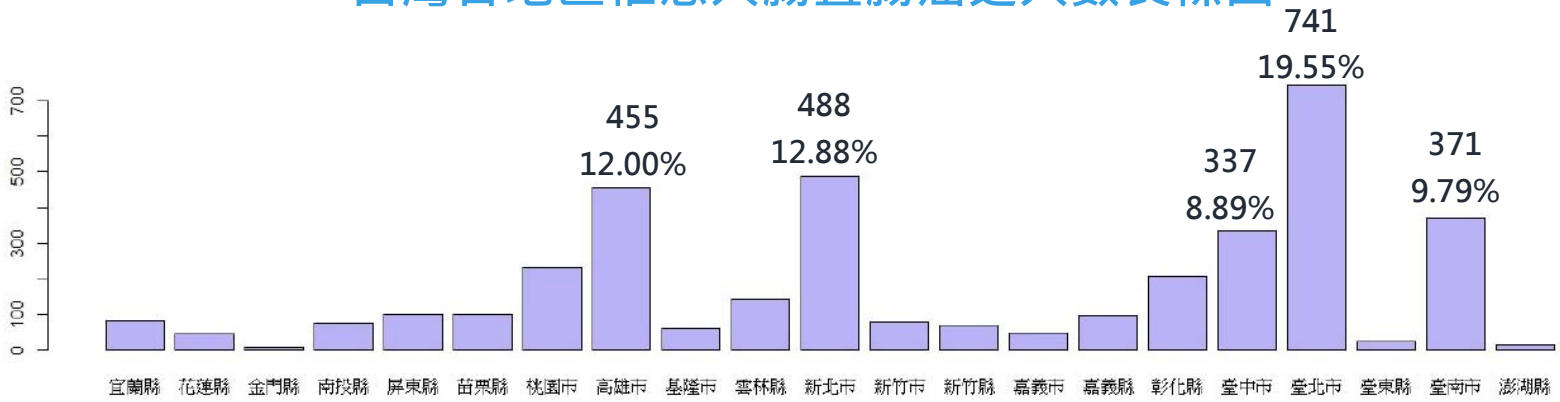
資料處理

探索性分析 (EDA)

台灣各地區總人數長條圖



台灣各地區罹患大腸直腸癌之人數長條圖

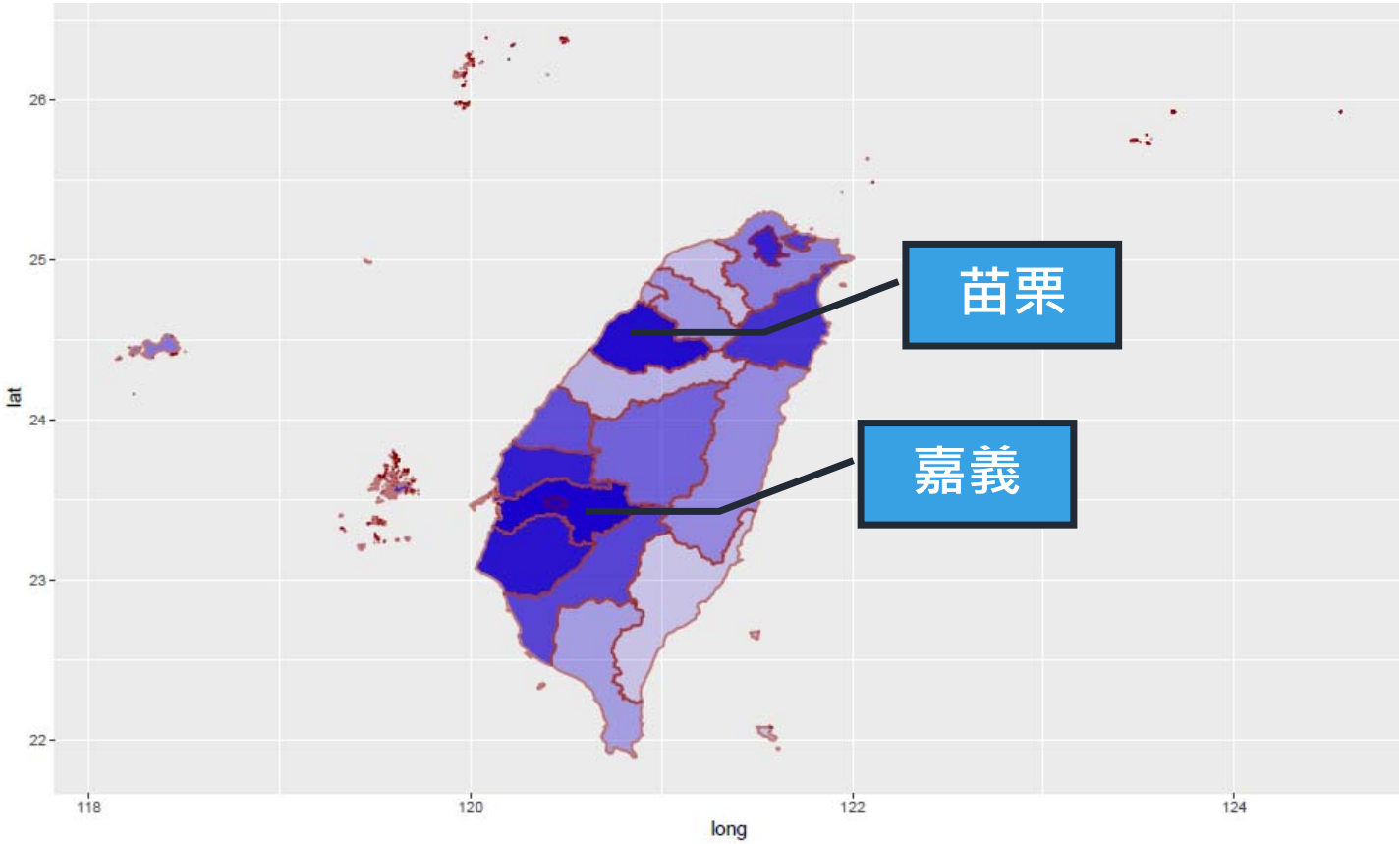


3

資料處理

探索性分析 (EDA)

台灣各地區罹患大腸直腸癌的比例



3

資料處理

探索性分析 (EDA)

整體特定治療項目代號

3

資料處理

探索性分析 (EDA)

罹患大腸直腸癌的患者
罹病前之特定治療項目代號

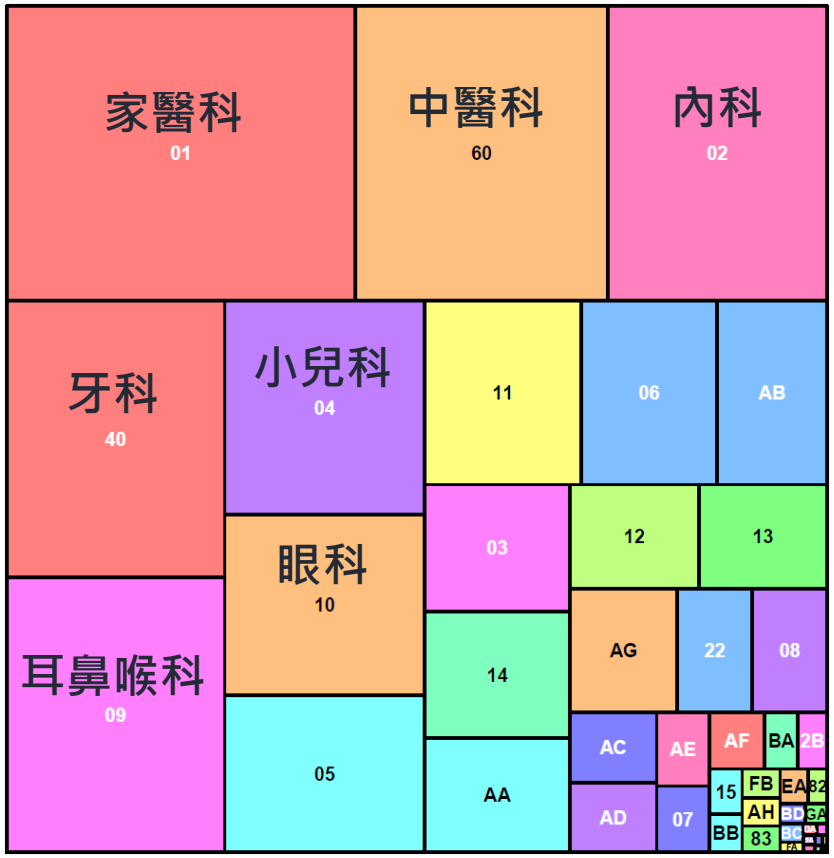
罹患大腸直腸癌的患者
罹病後之特定治療項目代號

3

資料處理

探索性分析 (EDA)

整體就醫科別



3

資料處理

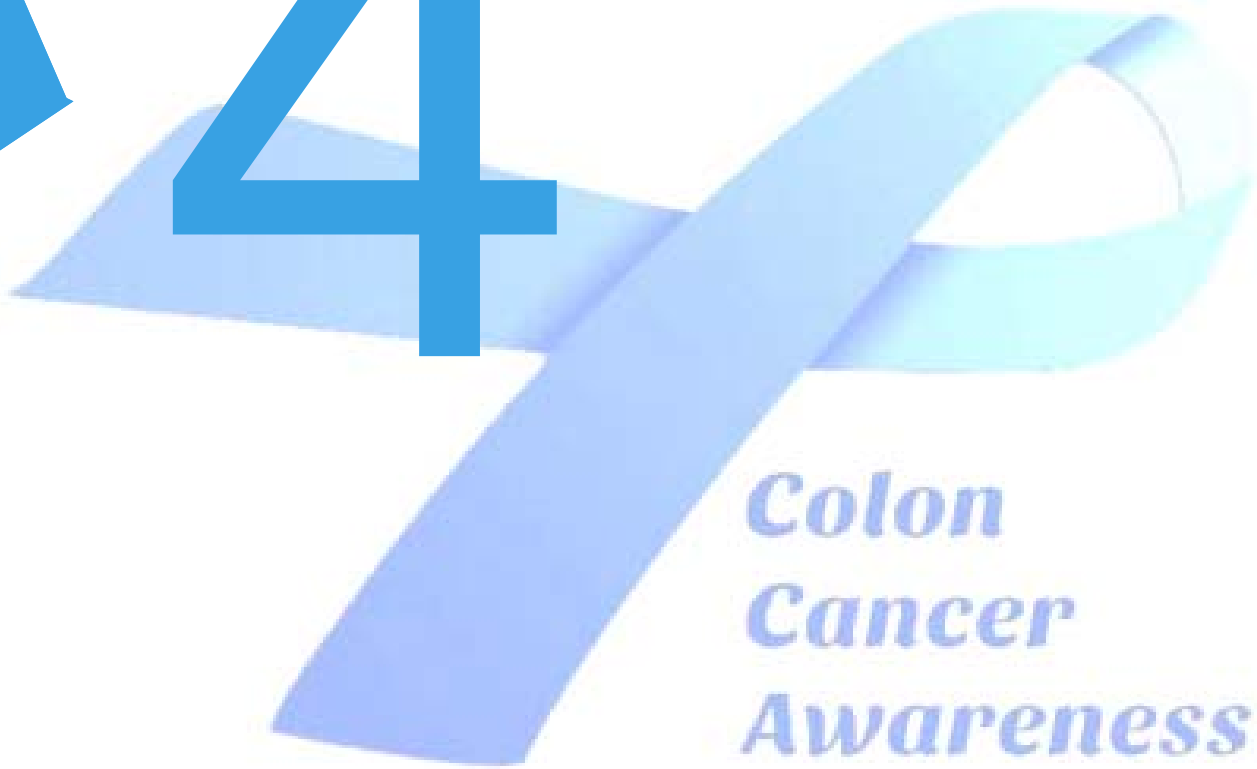
探索性分析 (EDA)

罹患大腸直腸癌的患者
罹病前之就醫科別

罹患大腸直腸癌的患者
罹病後之就醫科別

資料分析

4



4

資料分析

統計方法

關聯式分析

	lhs 前項		rhs 後項	support	confidence	lift
[1]	{}	=>	{colon=1}	0.008552	0.0086	1.0000
[2]	{01}	=>	{colon=1}	0.001044	0.0071	0.8245
[3]	{BA}	=>	{colon=1}	0.000826	0.3066	35.8525
[4]	{02}	=>	{colon=1}	0.000748	0.0080	0.9386
[5]	{AF}	=>	{colon=1}	0.000519	0.1150	13.4475
[6]	{03}	=>	{colon=1}	0.000456	0.0172	2.0060
[7]	{AA}	=>	{colon=1}	0.000393	0.0165	1.9241
[8]	{AB}	=>	{colon=1}	0.000381	0.0130	1.5239
[9]	{10}	=>	{colon=1}	0.000376	0.0072	0.8451

直腸外科

外科

心臟血管科

4

資料分析

統計方法

卡方獨立性檢定

與罹患大腸直腸癌是否獨立		
	χ^2	P 值
性別	51.46	<0.0001
縣市名稱	104.25	< 0.0001
地區	32.85	< 0.0001

T 檢定

變數平均與罹患大腸直腸癌是否有關				
	罹患	未罹患	T 值	P 值
年齡	66.98	40.10	-122.26	<0.0001

4

資料分析

統計方法

邏輯斯迴歸

完整模型：

$P_{colon=1}$

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} + \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \beta_{12} X_{12i} + \beta_{13} X_{13i}$$

$i = 1,2,\ldots,897489$

AIC ：42186

4

資料分析

統計方法

邏輯斯迴歸

向後選取法：

完整模型 → 簡化模型1 → 簡化模型2 → 簡化模型3

刪除耳鼻喉科看診次數
AIC ：42187

>

刪除心臟血管內科看診次數
AIC ：42186

>

刪除內科看診次數
AIC ：42185

4

資料分析

統計方法

邏輯斯迴歸

最終模型：

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_9 X_{9i} + \beta_{12} X_{12i} + \beta_{13} X_{13i}$$

$i = 1,2,\ldots,897489$

4

資料分析

統計方法

邏輯斯迴歸

變數	參數	參數估計	標準誤	Z值	P值
截距項	β_0	-9.444	0.07		
性別(男性)	β_1	0.268	0.03		
年齡	β_2	0.068	0.00		
地區(北區)	β_3	0.101	0.04		
地區(離島)	β_4	-0.089	0.20		
地區(東區)	β_5	-0.282	0.12		
地區(南區)	β_6	0.125	0.04		
家醫科看診次數	β_7	-0.004	0.00		
腸胃內科看診次數	β_9	0.065	0.00		
眼科看診次數	β_{12}	-0.011	0.00		
中醫科看診次數	β_{13}	0.008	0.00		

中醫科看診每增加一次，罹病機率增加 $\exp(0.008) = 1.01$ 倍

結論 5

5

結論

結果與貢獻

大部分的患者在罹癌前較常就醫的科別有家醫科、腸胃內科等等,可提供高風險的患者做為依據。

性別、年齡、地區等等可能是影響是否罹患大腸直腸癌的重要因素。

根據探索性分析也可以初步的發現, 罹患大腸直腸癌的患者在健保點數上的花費較高。

5

結論

分析的限制

未記錄該患者的身高、體重和重要的生活習慣, 例如: 抽菸、運動、飲食等。

國際疾病分類碼只能看出該患者是否有罹患某種疾病, 但無法看出癌症的分期。

若要進行預測可能需要三年的資料, 然而受限於電腦設備無法匯入如此龐大的資料。

謝謝
ThankYou
Merci
Salamat
Hvala
Kop
Tack
kun
Xie
Gra
Diolch
Kiitos
Sheun
umesc
Kasih
Todah
Mamnoon
Shnorhakalutiun
Shokriya
Spaas
Mul
Ači
Gamsahapnida
Ngiyabonga
Dziękuję
Shokrun
Dank
Takk
Cam
Dankie
Kruthagnathalu
Arigatou
Dhonnobaad
Dhanyavad
Khopjai
Gomapsupnida
Terima
Enkosi
Danke
dank
Euxaristo
Kun
Shukriya
Or
ederim
Hain
Dhan
Asante
daa
Falemindirit
Grazie
Go
Dhanyavaad
Dhanyavaadaalu
Waad
Daw
Dakujem
krap
Grazzi
raibh
Gracias
Nandree
Blagodariya
Fyrir
Te ekkür
Dekuju/Dekujeme
Gamsahapnida
Or
al