美兆資料研究

目錄

壹、 資料說明

美兆資料庫一共有 210 個，觀測值 62321 筆。其中基本人口變量 20 個，功能性醫學變量 70 個，以及健康問卷變量 140 個。

一、 測驗次數統計

測驗次數	3	4	5	總計
人數	12954	4111	1403	18468

二、 測驗年份統計

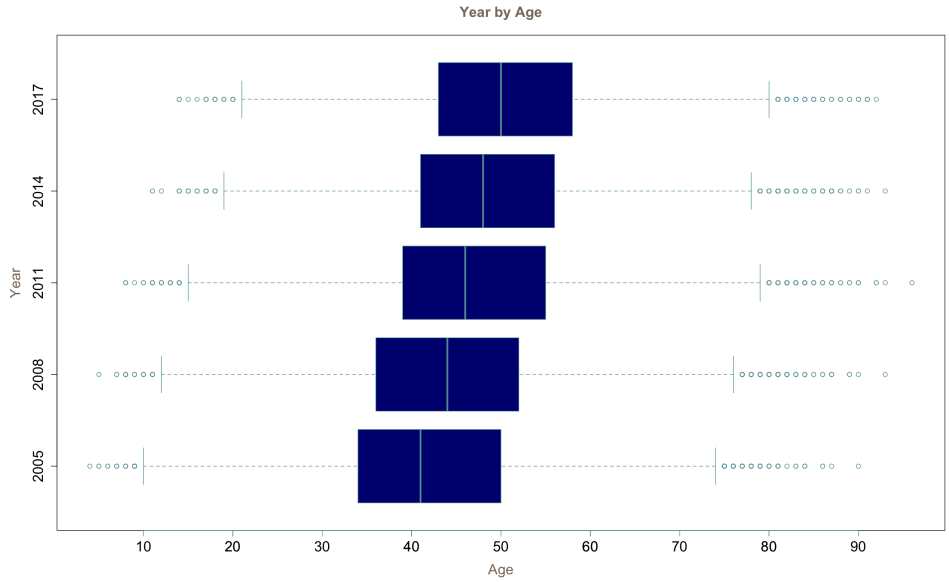
測驗年份	2005	2008	2011	2014	2017
人數	13576	16122	16467	9868	6288

三、 變數介紹

變數名稱	變數個數
人口變項	18
問卷資料	
工作狀況	5
生活習慣	12
飲食習慣	21
運動習慣	15
睡眠品質與壓力	9
近況	19
個人及家族病史	41
健檢資料	
一般檢查	15
血液常規檢查	5
白血球五項分類	5
血糖檢查	2
肝膽功能檢查	9
腎功能檢查	3
尿酸檢查	1
血中脂肪檢查	4
鈣、磷、血清鐵檢查	
B 型肝炎檢查	4
腫瘤標記檢查	2
甲狀腺功能篩檢	1
組織發炎篩檢	1
血型檢查	1

尿液常規篩檢	9
大腸直腸癌篩檢	1
X 光檢查	1
肺功能檢查	3
婦科檢查	1
骨質密度篩檢	2

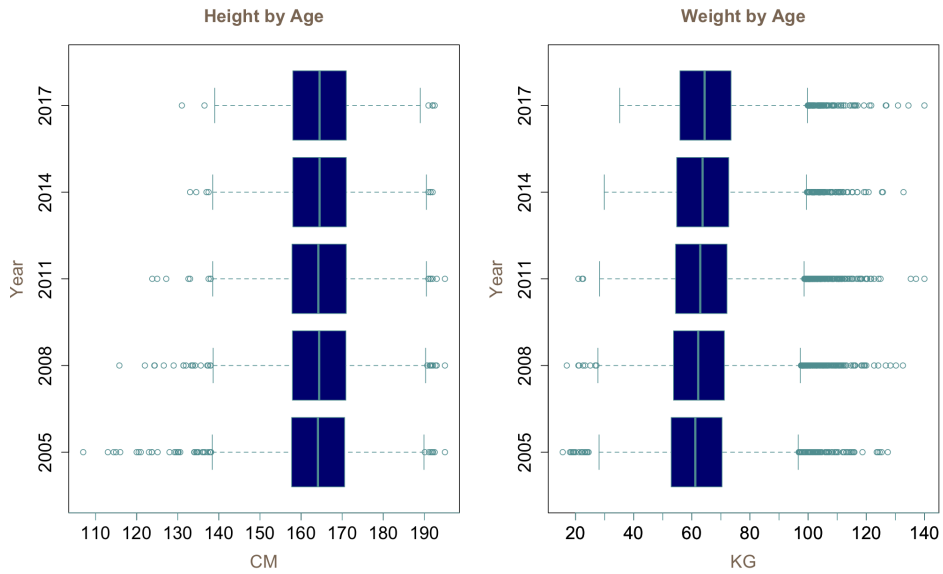
四、 敘述統計



(一) 年紀

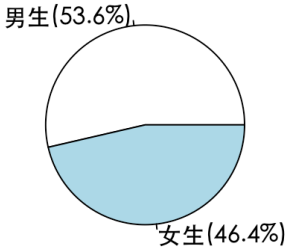
從 2005 年至 2017 年，可以看到受測者年齡層分布逐漸往上。由 2005 的資料皆為第一次受測，受測人數 13576 人，而 2008 年第一次受測的人數剩下 3791 人；又此份資料中每人至少受測三次，所以 2014 及 2017 年皆無第一次測驗的人，各年度受測年齡層上升是必然的結果。

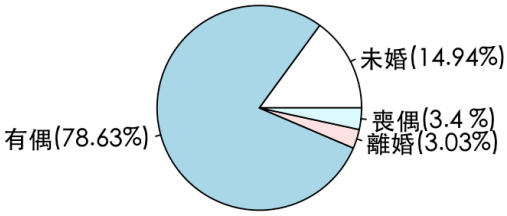
(二) 身高與體重

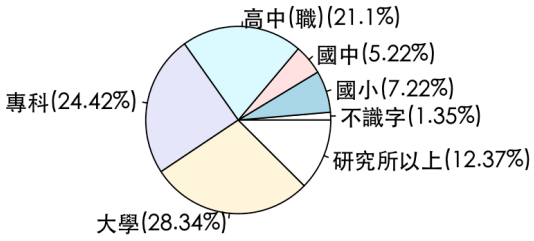


各年度的受測者身高並無明顯差異，體重則隨年度增加又略微上升趨勢，但整體變化幅度不大，透過 T 檢定發現，2005 與 2008 年間身高皆具有顯著差異(p 值皆小於 0.0001)，而 2011、2014 與 2017 年間則不具有顯著差異，體重方面各年度皆具有顯著差異，將離群值移除後，則體重在 2011、2014 與 2017 年間不具有顯著差異。

(三) 性別、婚姻狀況與教育程度

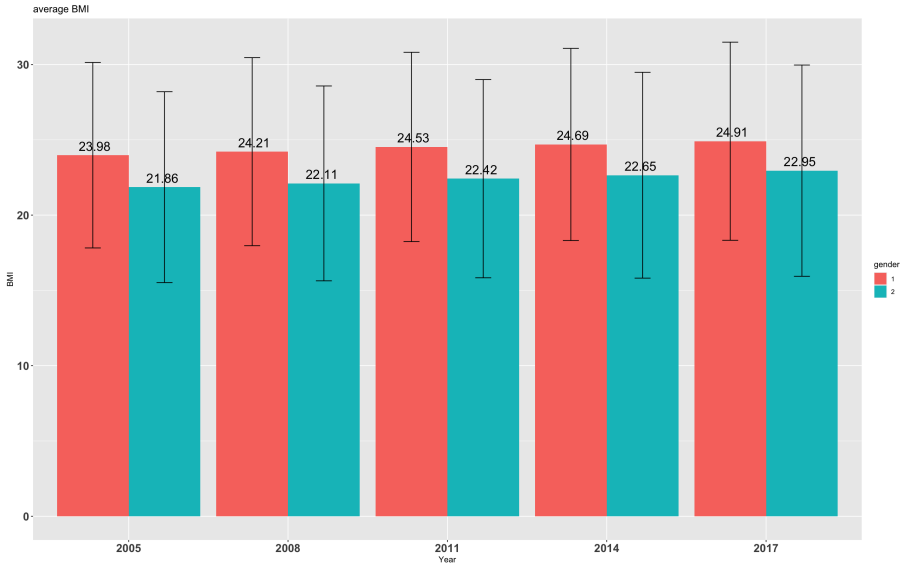






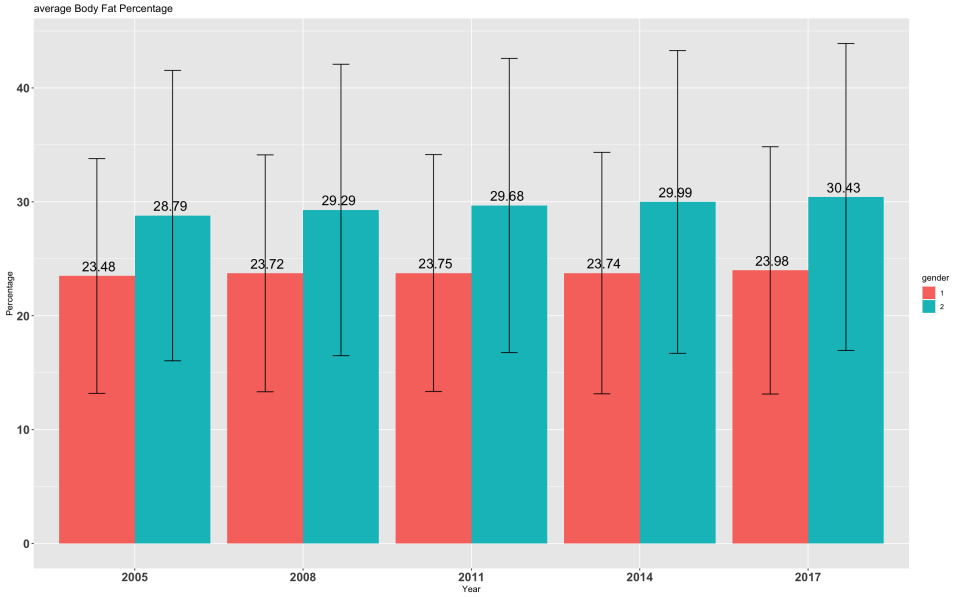
性別方面，受測者男生多於女生，相差 7 個百分點；婚姻狀況方面，以有偶及未婚為主，其餘婚姻狀況佔約 6.43%；教育程度方面，佔比前三名依序為大學，專科及高中職，而研究所人數則約為專科的一半。

(四) 身體質量指數(1:男生,2:女生)



各年度受測者男女生平均 BMI 有上升趨勢，男生約在 24-25 之間，女生平均 BMI 則在 22-23 之間，男生或女生的平均值皆逐年往上提升。

(五) 體脂肪率(1:男生,2:女生)



男生平均體脂肪率沒有明顯變化，在 23-24 之間，女生在 28-30 左右，且平均值有逐年上升的趨勢，誤差區間也有加大。

貳、 代謝症候群

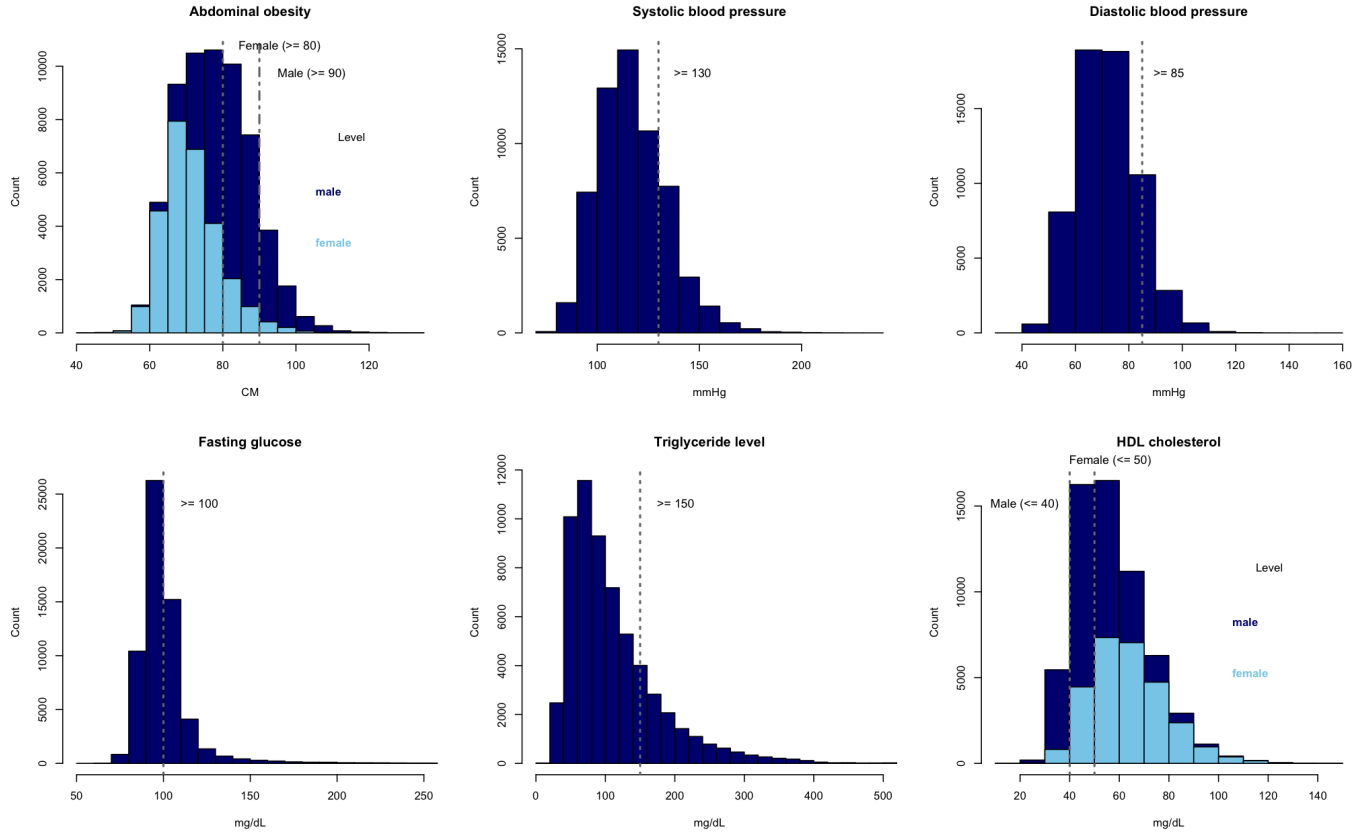
一、 代謝症候群判定標準

1. 腹部肥胖:男性的腰圍≧90cm(35 吋)、女性腰圍≧80cm(31 吋)。
2. 血壓偏高:收縮壓≧130mmHg 或舒張壓≧85mmHg，或是服用醫師處方高血壓治療藥物。
3. 空腹血糖偏高：空腹血糖值≧100mg/dL，或是服用醫師處方治療糖尿病藥物。
4. 空腹三酸甘油酯偏高：≧150mg/dL，或是服用醫師處方降三酸甘油酯藥物。
5. 高密度脂蛋白膽固醇偏低：男 性<40mg/dL、女性<50mg/dL。

以上五項組成因子，符合三項(含)以上即可判定為代謝症候群 。

資料來源：衛生署 https://www.hpa.gov.tw/Pages/List.aspx?nodeid=221

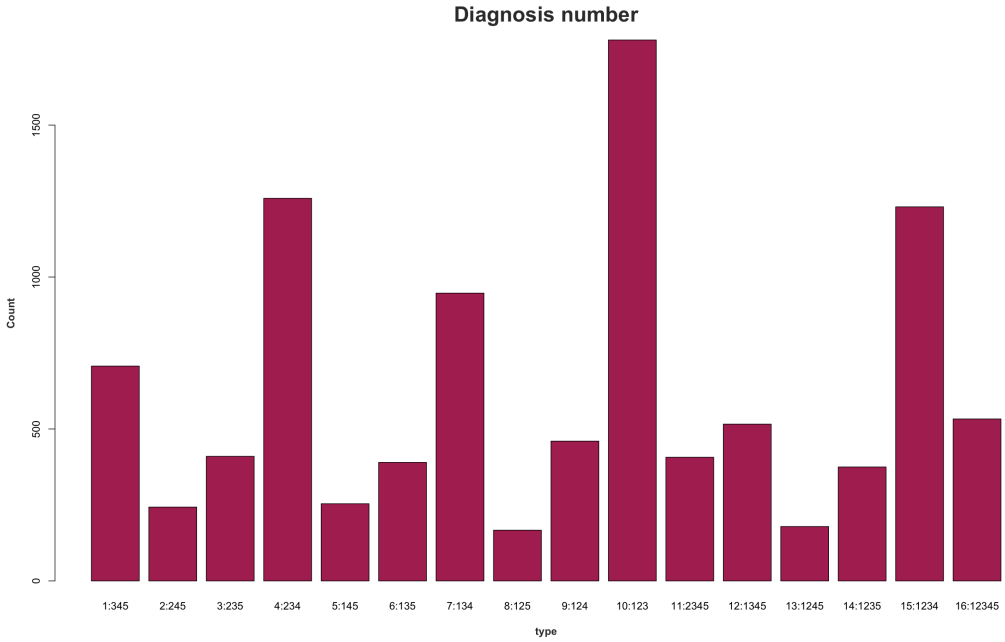
二、 代謝症候群狀態



三、 代謝症候群比例

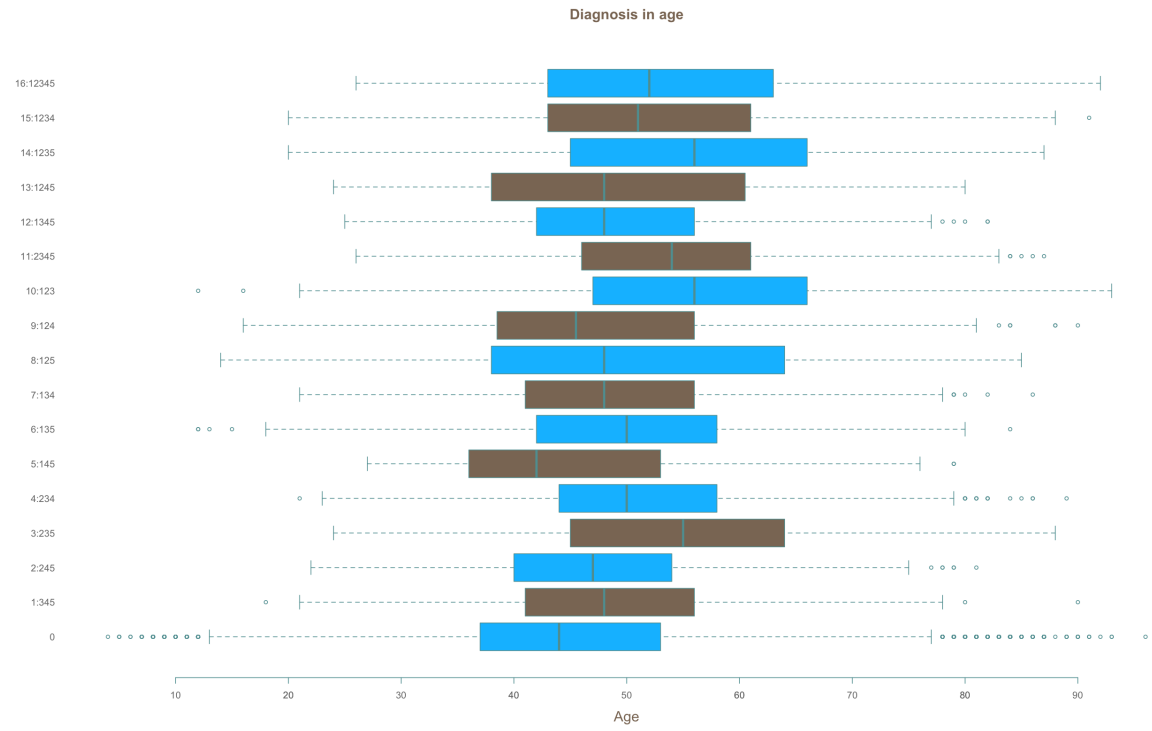
由於代謝症候群共有 16 種情況，將症狀依序標為 1-5，以利後續視覺化描述。

症狀	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
(1)腹部肥胖	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1
(2)血壓偏高	0	1	1	1	0	0	0	1	1	1	1	0	1	1	1	1
(3)空腹血糖偏高	1	0	1	1	0	1	1	0	0	1	1	1	0	1	1	1
(4)空腹三酸甘油酯偏高	1	1	0	1	1	0	1	0	1	0	1	1	1	0	1	1
(5)高密度脂蛋白膽固醇偏低	1	1	1	0	1	1	0	1	0	0	1	1	1	1	0	1
總計	707	243	410	1259	254	390	947	167	460	1780	407	516	179	375	1231	533
比例	1.17	0.4	0.68	2.08	0.42	0.64	1.56	0.28	0.76	2.94	0.67	0.85	0.3	0.62	2.03	0.88



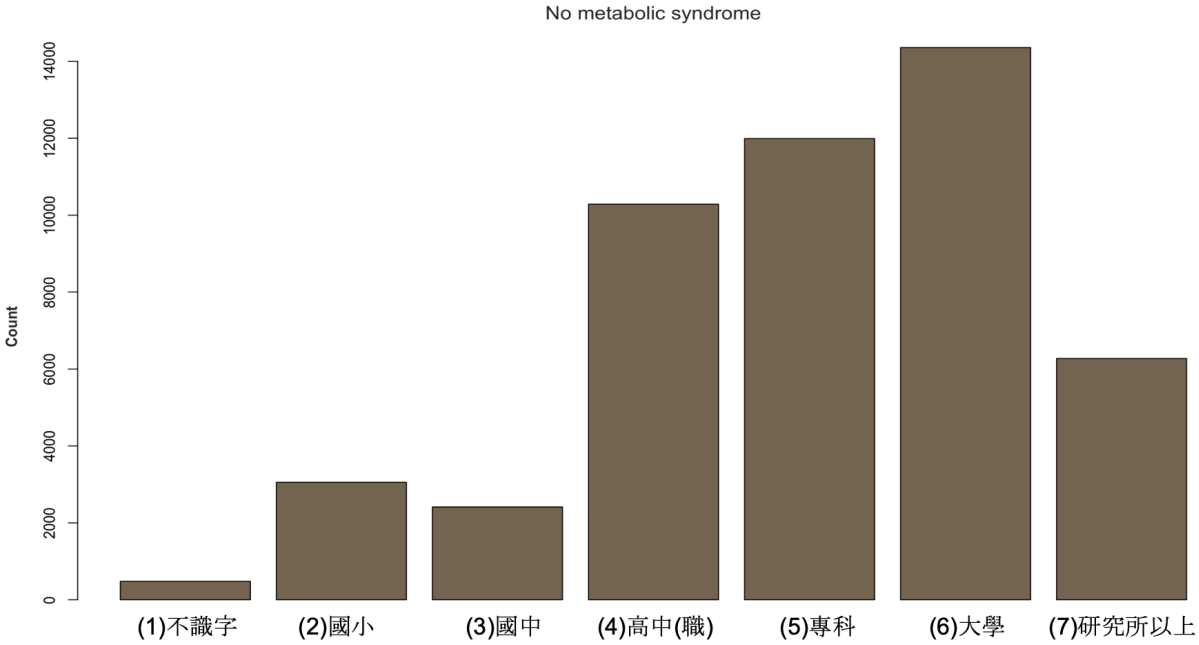
四、 代謝症候群敘述統計

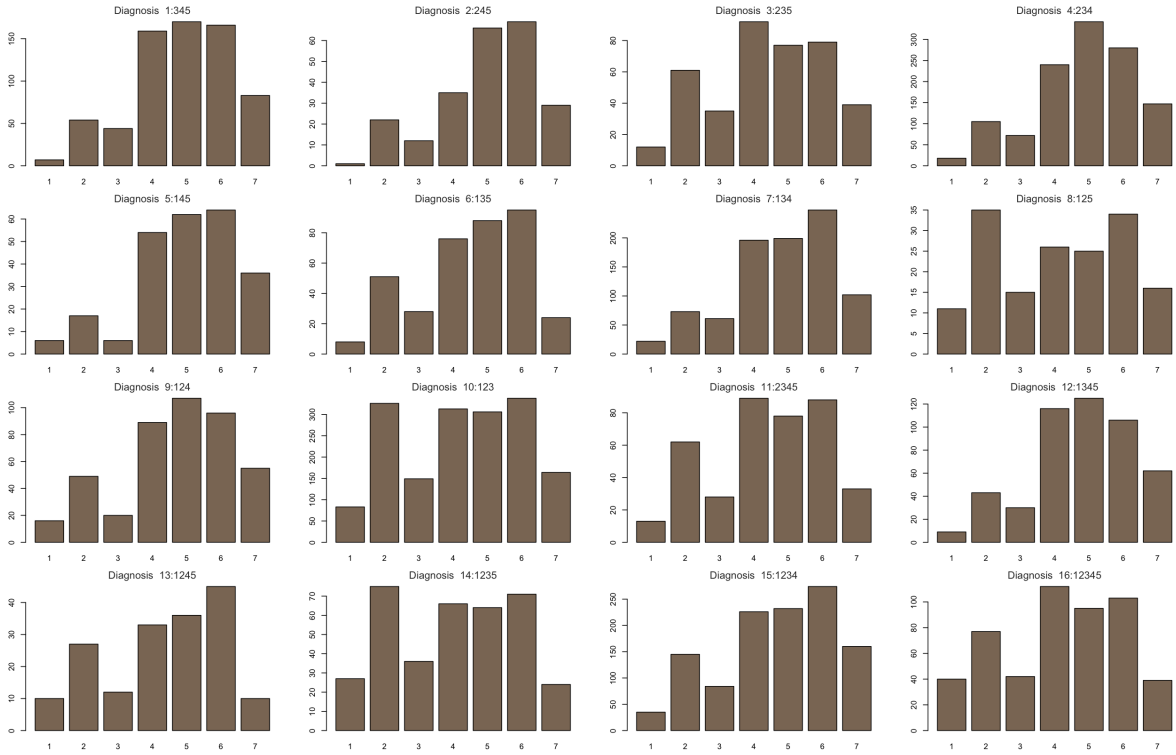
(一) 年紀



與無代謝症候群得受測者相比，除了第五類(腹部肥胖、三酸甘油酯偏高，高密度脂蛋白膽固醇偏低)的人之外，其餘患者普遍年紀較大。

(二) 教育程度

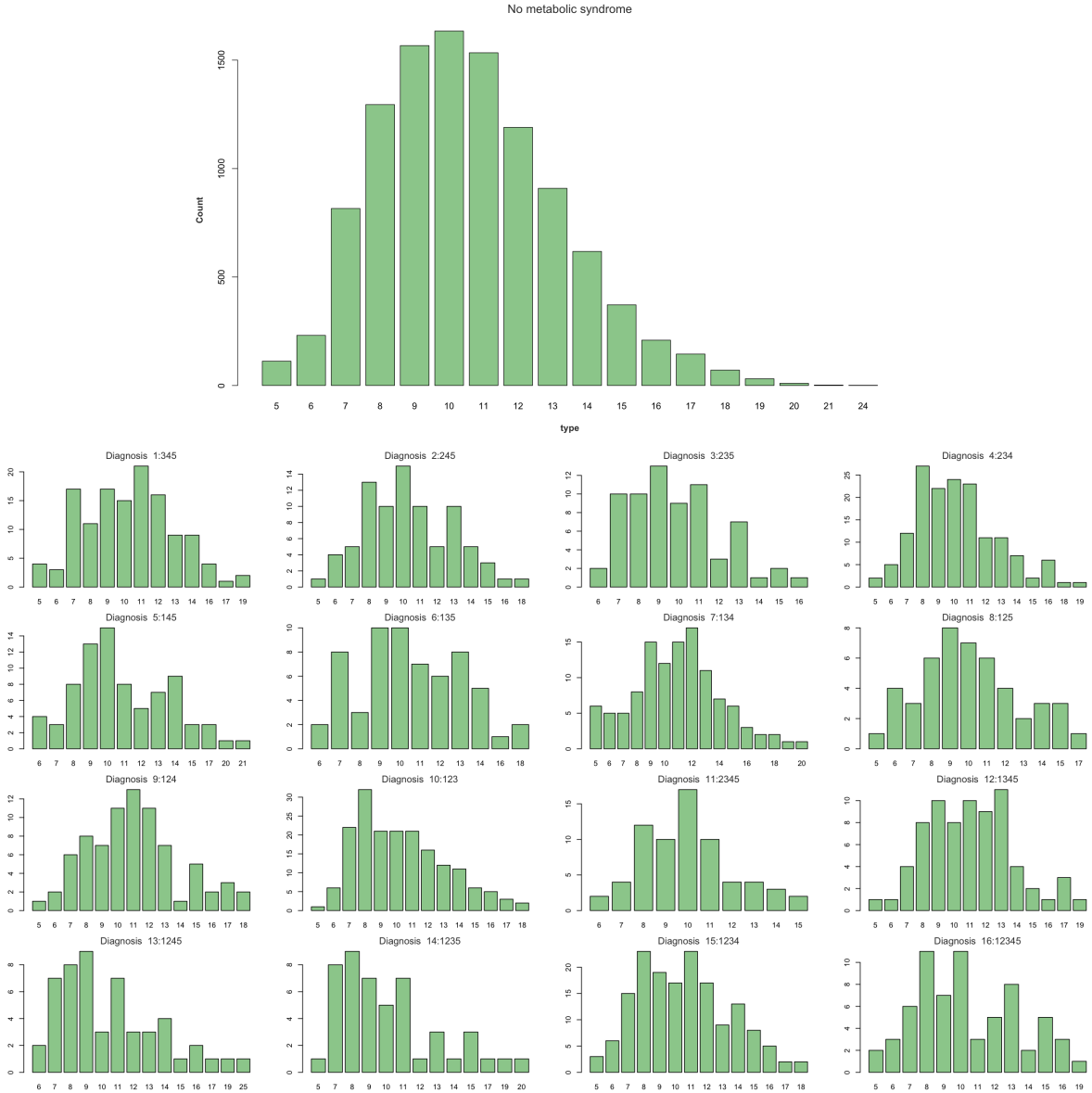




與無代謝症候群相比，教育程度為國小的受測者在第 8、10、14 組佔有比例皆有明顯上升，共同特徵皆為腹部肥胖與血壓偏高，而教育程度為國小者平均年齡為 56 歲，標準差 14，屬於年紀偏高的一群，研判教育程度僅國小多數為老人，造成代謝症候群且腹部肥胖、血壓偏高的比例較高。

(三) 飲食習慣(以 2005 年做分析)

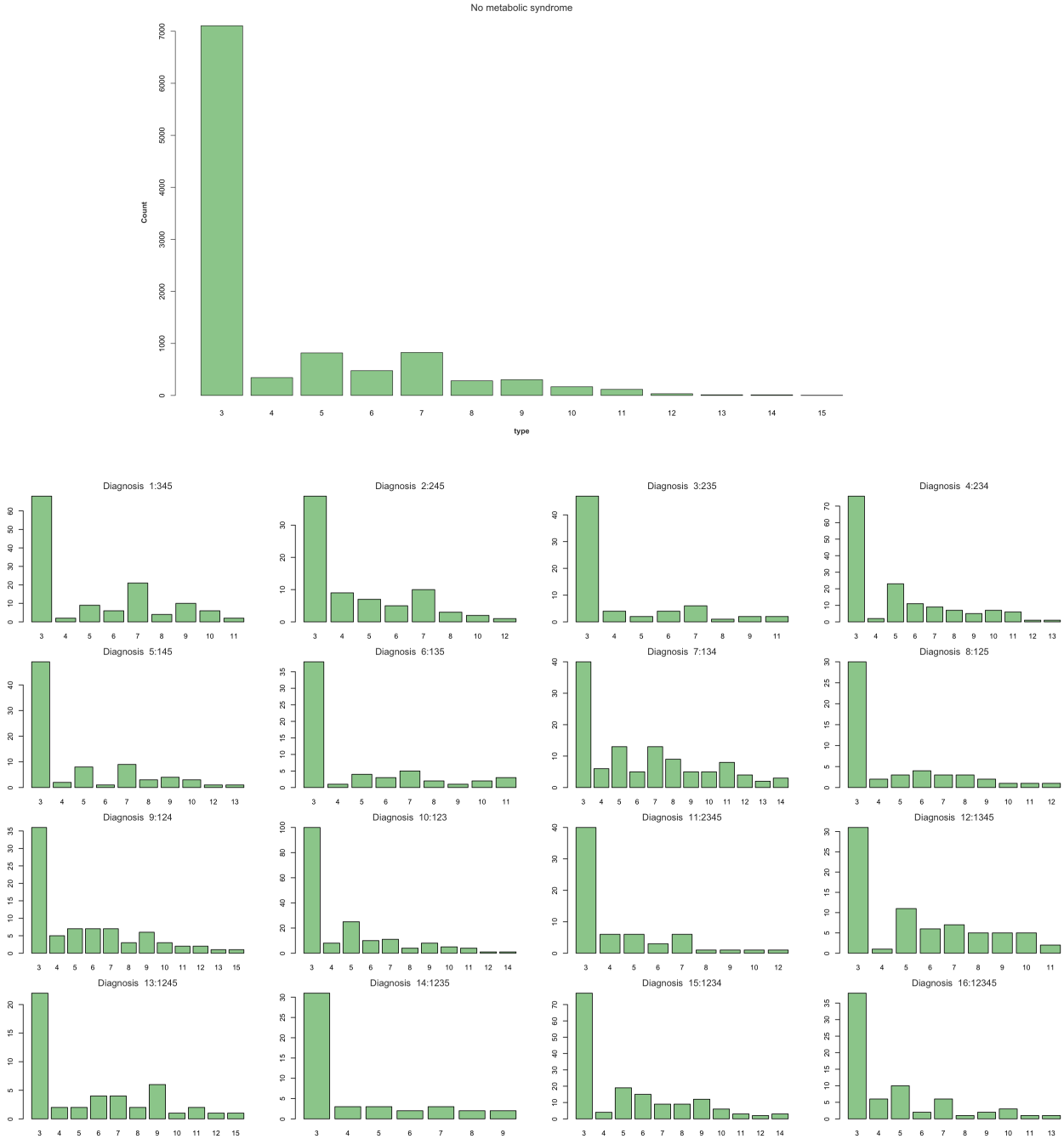
變數篩選		
food06_98	肉類(含豬、雞、鴨、羊、牛肉)吃多少? (1 盤約為豬牛排 1 片<約手掌大小厚 1 公分>或棒棒腿 1 隻或漢堡肉 1 塊、或其他瘦肉約 4 湯匙)	(1)不吃或每週少於 1 份 (2)每週吃 1-3 份 (3)每週吃 4-6 份 (4)每天吃 1 份 (5)每天吃 2 份或以上
food07_98	水產類吃多少? (1 份相當於中型秋刀魚 1 尾、或生魚片 4 片、或魚肉 4 湯匙、或草蝦 4 尾、蚵 16 粒)	(1)不吃或每週少於 1 份 (2)每週吃 1-3 份 (3)每週吃 4-6 份 (4)每天吃 1 份 (5)每天吃 2 份或以上
food19_98	有沒有吃加果醬或蜂蜜的食物? (1 份相當於果醬或蜂蜜約 2 茶匙)	(1)不吃或每週少於 1 份 (2)每週吃 1-3 份 (3)每週吃 4-6 份 (4)每天 1 份 (5)每天吃 2 份或以上
food20_98	有沒有喝加糖的咖啡、可可、茶、果汁或飲料? (如汽水、綠豆湯)(240C.C.為 1 杯)	(1)不喝或每週少於 1 杯 (2)每週喝 1-3 杯 (3)每週喝 4-6 杯 (4)每天喝 1 杯 (5)每天喝 2 杯或以上
food21_98	平常除主食外，其他(14 增加)用油炸的、或用油煎的食物吃多少?(1 份以半碗計)	(1)不吃或每週少於 1 份 (2)每週吃 1-3 份 (3)每週吃 4-6 份 (4)每天吃 1 份 (5)每天吃 2 份或以上



可以看到不管有無代謝症候群，在飲食上並沒有一定的趨勢，即使是第 16 組(代謝症候群五個症狀皆有的受測者)，飲食習慣上也不會過度攝取油炸、糖類肉類等食物。也可能是因為知道自己有代謝症候群，因此飲食上開始節制，這點無從得知。

(四) 生活習慣(以 2005 年做分析)

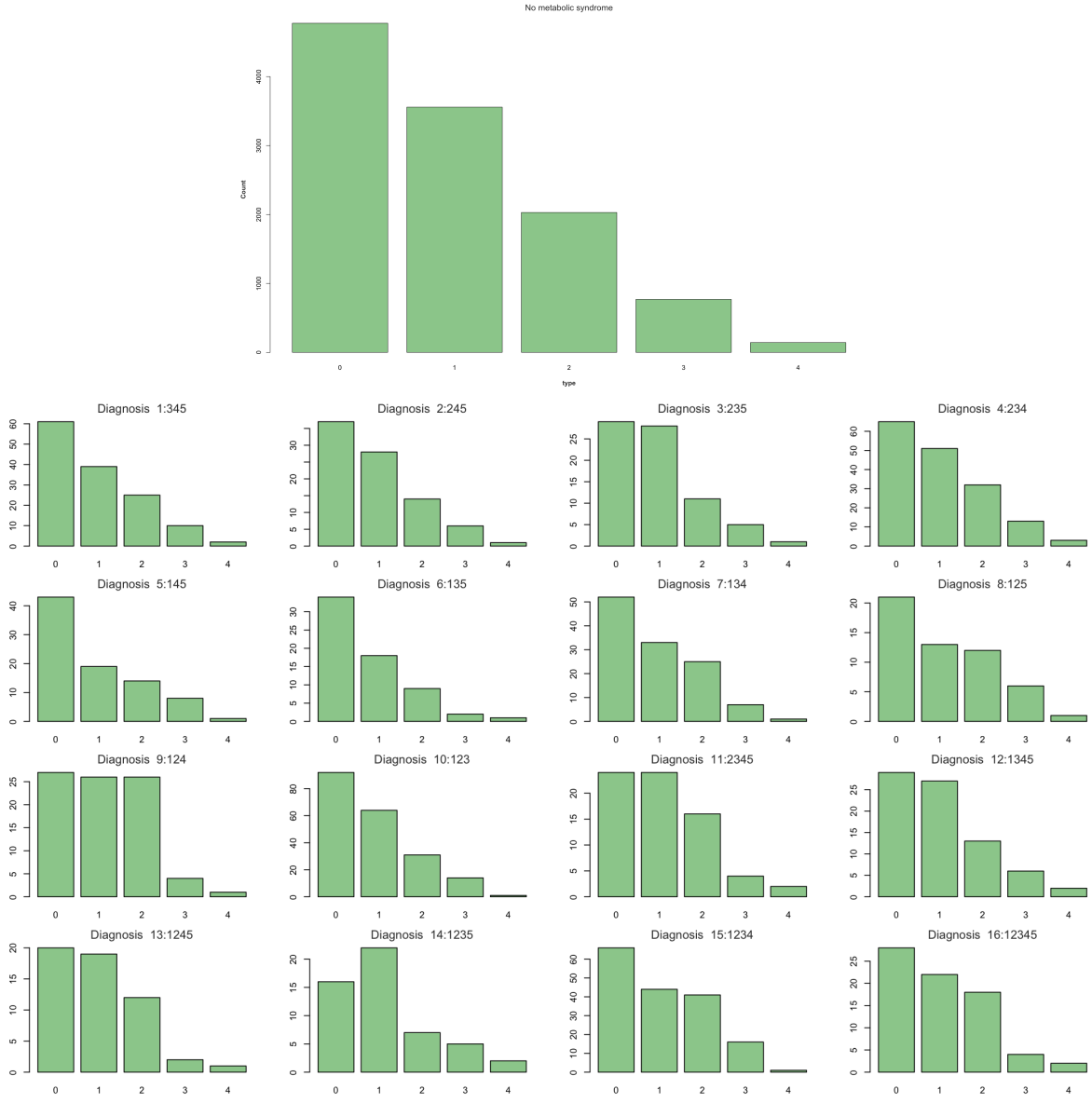
變數篩選		
smokeornot_03	您抽煙嗎?	(1)不抽(選擇不抽者跳答下一大題) (2)不抽，但經常吸二手煙 (3)以前抽，現已戒煙 (4)偶爾抽 (5)每天抽
cocohabit_98	您是否嚼檳榔	(1)不嚼 (2)以前嚼，現已戒 (3)每週 1~3 次 (4)每週 4~5 次 (5)每天嚼
drinkornot_98	您喝酒嗎?	(1)不喝或每週少於 1 次(2)以前喝，現已戒酒 (3)每週 1~2 次 (4)每週 3~4 次 (5)每天喝



可以看到有代謝症候群的受測者，在生活習慣上，抽煙喝酒嚼檳榔的比例明顯高於無代謝症候群的受測者，且當有較高比例這種生活習慣時並有代謝症候群的受測者，大多具有第四種症狀：空腹三酸甘油酯偏高。

(五) 家族病史(以 2005 年做分析)

變數篩選		
rsick09	親屬(祖父母、父母、兄弟姊妹及子女)疾病?高血壓	(0)否 (1)是
rsick10	親屬(祖父母、父母、兄弟姊妹及子女)疾病?高血糖(糖尿病)	(0)否 (1)是
rsick11	親屬(祖父母、父母、兄弟姊妹及子女)疾病?腦血管疾病	(0)否 (1)是
rsick12	親屬(祖父母、父母、兄弟姊妹及子女)疾病?心臟血管疾病	(0)否 (1)是



可以看出有代謝症候群的受測者，在被診斷出某些症狀時，與親屬是否罹患疾病有一定關係。可以看到第 14 組（1235）甚至 1 的人數大於 0 的人數，另外在第 3、9、11 組，親屬是否有特定疾病的比例也比沒有代謝症候群的人明顯來得高。

另外，在長期服用藥物方面，有代謝症候群者明顯多於無代謝症候群者，運動方面則是有代謝症候群者運動量較大，也可能是因為知道自己有代謝症候群才開始運動，並無從得知。

參、 模型預測(以 2005 年做分析)

一、 資料前處理

利用 logistic regression 模型觀察變數顯著與否，篩選變數如下表：

人口變項	
education	教育程度
occupation	職業
fincome	家庭年所得
生活習慣	
smokeornot_03	你抽煙嗎
cocohabit_98	你嚼檳榔嗎
drinkornot_98	你喝酒嗎
飲食習慣	
nutrino	您平常額外補充何種營養品或保健食品
Food20_98	有沒有喝加糖的咖啡、可可、茶、果汁或飲料?(如汽水、綠豆湯)(240C.C.為 1 杯)
Food07_98	水產類吃多少?(1 份相當於中型秋刀魚 1 尾、或生魚片 4 片、或魚肉 4 湯匙、或草蝦 4 尾、蚵 16 粒)
Food06_98	肉類(含豬、雞、鴨、羊、牛肉)吃多少?(1 盤約為豬牛排 1 片<約手掌大小厚 1 公分>或棒棒腿 1 隻或漢堡肉 1 塊、或其他瘦肉約 4 湯匙)
運動習慣	
sportornot_98	固定做運動的時間
近況	
allergydrug	您對藥物過敏嗎?
個人及家族病史	
rsick09	親屬(祖父母、父母、兄弟姊妹及子女)疾病?高血壓
rsick10	親屬(祖父母、父母、兄弟姊妹及子女)疾病?高血糖(糖尿病)
rsick11	親屬(祖父母、父母、兄弟姊妹及子女)疾病?腦血管疾病
rsick12	親屬(祖父母、父母、兄弟姊妹及子女)疾病?心臟血管疾病
一般檢查	
g_bmi	身體質量指數
g_pul	脈搏
g_fat	體脂肪率

二、 機器學習

(一) 模型一（有無代謝症候群）

首先將上述變數放入不同模型中，利用性別及年齡做 1:1 的 Match 後，再透過 10-fold CV 交叉驗證，希望透過上述變數判斷受測者是否有代謝症候群，模型的效能指標如下表。

模型效能指標

模型	Accuracy	Precision	Recall
羅吉斯迴歸	0.7649	0.7725	0.7580
隨機森林	0.7335	0.7371	0.7299
SVM	0.7319	0.7327	0.7333

(二) 模型二（單一代謝症候群情況+其他）

將代謝症候群 16 種情況分開，輪流將每種情況視為有代謝症候群，利用模型預測受測者有無此種情況下的代謝症候群，使用性別及年齡變數做 1:1 的 Match，以下為邏輯斯回歸與隨機森林，做 10-fold Cross-validation 後，模型的效能指標：

情況	邏輯斯回歸			隨機森林		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
1:345	0.7585	0.7790	0.7362	0.7462	0.7479	0.7395
2:245	0.7558	0.7458	0.7777	0.7419	0.7450	0.7354
3:235	0.7550	0.7477	0.7773	0.7415	0.7436	0.7374
4:234	0.7585	0.7768	0.7419	0.7410	0.7435	0.7371
5:145	0.7570	0.7549	0.7686	0.7391	0.7395	0.7368
6:135	0.7594	0.7493	0.7796	0.7395	0.7420	0.7318
7:134	0.7558	0.7606	0.7519	0.7434	0.7445	0.7436
8:125	0.7582	0.7614	0.7610	0.7399	0.7434	0.7346
9:124	0.7566	0.7722	0.7346	0.7414	0.7412	0.7419
10:123	0.7605	0.7538	0.7700	0.7406	0.7422	0.7367
11:2345	0.7621	0.7591	0.7687	0.7450	0.7478	0.7409
12:1345	0.7562	0.7600	0.7620	0.7431	0.7481	0.7363
13:1245	0.7546	0.7592	0.7457	0.7443	0.7446	0.7404
14:1235	0.7586	0.7647	0.7612	0.7418	0.7430	0.7392
15:1234	0.7593	0.7411	0.8030	0.7442	0.7479	0.7373
16:12345	0.7589	0.7405	0.8000	0.7426	0.7451	0.7407

由上表可以發現，當預測是否有代謝症候群，改為預測是否有代謝症候群的 16 種狀況其中一種時， 不管是邏輯斯回歸或隨機森林，各效能指標與模型一（預測整體是否有代謝症候群）相比，沒有明顯的提升。

(三) 模型三（單一代謝症候群情況＋無代謝症候群）

將代謝症候群 16 種情況輪流與無代謝症候群的受測者合併，並利用性別及年齡做 1:1 的 Match，使用性別及年齡變數做 1:1 的 Match，以下為邏輯斯回歸與隨機森林，做 10-fold Cross-validation 後，模型的效能指標：

情況	邏輯斯回歸			隨機森林		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
1:345	0.7101	0.7908	0.6322	0.5838	0.5787	0.5990
2:245	0.6819	0.6320	0.8167	0.6333	0.6355	0.6630
3:235	0.7263	0.6897	0.8306	0.4705	0.4754	0.4501
4:234	0.6944	0.6763	0.8656	0.5822	0.5786	0.5972
5:145	0.8489	0.7985	0.9407	0.7879	0.7576	0.8664
6:135	0.8402	0.8313	0.9157	0.8667	0.8585	0.9050
7:134	0.8905	0.8958	0.8881	0.8143	0.8076	0.8067
8:125	0.7695	0.7317	0.8083	0.8736	0.8405	0.9267
9:124	0.9095	0.8966	0.9095	0.8748	0.8756	0.8700
10:123	0.8527	0.8211	0.9120	0.7980	0.7718	0.8478
11:2345	0.8091	0.7974	0.8282	0.6455	0.6281	0.6811
12:1345	0.8616	0.8184	0.9472	0.7904	0.7610	0.8849
13:1245	0.7500	0.7286	0.8833	0.8119	0.6667	0.7333
14:1235	0.8982	0.9207	0.9350	0.8964	0.8900	0.9383
15:1234	0.8828	0.8425	0.9467	0.8205	0.8147	0.8444
16:12345	0.8526	0.8208	0.9083	0.8603	0.8468	0.9056

由上表可以發現，模型預測單一種代謝症候群情況與無代謝症候群受試者資料時，在有腹部肥胖的條件下判斷為代謝症候群患者的預測效果似乎較佳。於是將無腹部肥胖問題卻有代謝症候群的受測者移除後，重新預測模型對於判斷受測者是否有代謝症候群的能力。

以下是透過邏輯斯回歸判斷有腰部肥胖的問題，且有代謝症候群的受測者，與模型一相比，各效能指標並沒有提升，代表模型無法在受測者有腹部肥胖問題時，較有效預測受測者是否有代謝症候群。

模型	Accuracy	Precision	Recall
羅吉斯迴歸	0.6159	0.6443	0.5751

肆、 結論

從美兆的資料中發現幾點：

1. 年紀的增長確實與代謝症候群有相關。
2. 代謝症候群中，高密度脂蛋白膽固醇偏低較難從模型中被判斷出是否有此一條件，而當受測者有腰部肥胖，高血糖或高三酸甘油酯問題時，較容易透過模型預測是否有代謝症候群問題。
3. 從飲食上並無一定趨勢可以判斷代謝症候群與常吃甜點油炸類食物間的關聯。
4. 從資料上來看，抽煙喝酒嚼檳榔，以及親屬是否罹患相關疾病，都會增加有代謝症候群的機率。