傾向分數配對–淺談馬哈蘭距離之應用

副統計分析師 林怡諄

傾向分數配對(Propensity Score Matching)於流行病學與生物醫療領域是一種常見使用的方式，而在其他領域如經濟學、社會學等亦採用此方法分析某一社會經濟政策變動對於目標個體之行為影響，可見其應用範圍廣大。因此，後續有很多學者基於傳統傾向分數，提出很多增進配對成效之改良方式，使得傾向分數配對更加能夠發揮其效能。接下來，我們主要介紹其中一種改良方式–馬哈蘭距離(Mahalanobis Distance)如何應用於傾向分數配對。

一、傾向分數配對之概述

我們進行觀察性研究分析時，通常欲想得知某個我們所關心的治療方式或是影響因子，對於存活率或罹患某疾病之結果的影響程度，但分析過程之中，其結果會受到其他干擾因子的影響，使得我們無法清楚地得出結論。此時，我們可以透過配對(matching)方式，找出兩個群體，一為觀察組，另一為對照組，理論上這兩群僅有在特定因子有所差異外，其他因子均無差異，如年齡、性別、共病症均相同，就可以輕易證實其特定因子是否有其影響性。

然而，透過配對方式就可以完全排除干擾嗎？事實上，卻並非如此。因此，如何將干擾因子的影響降至最低，減少選擇偏誤(Selection Bias)或內生性問題(Endogenity)，以增進結論正確性，是諸多學者多年來努力的目標。

而其中傾向分數結合配對方式(Propensity Score Matching Method)，似乎是一種不錯的選擇。Johnson et al. (2009) 指出傾向分數配對(Propensity Score Matching)可以使得觀察組與比較組受試者之基線特性(characteristics of baseline)，與透過隨機分配所得的臨床試驗結果非常類似，並且其兩組的控制變數均呈現幾乎相同的分佈情況。

傾向分數(Propensity Score)的概念是透過一個病患自身的可觀察到變數，來衡量有這些變數特徵時，發生事件(如患病)之機率程度，其機率愈高，則分數愈高。換言之，傾向分數是個發生機率的觀念，因此，我們可知傾向分數的範圍界於 0 ~ 1 之間。

換言之，我們計算傾向分數有種逆推相似程度的概念，當兩個個體發生事件

的機率相似時，將有很高的機會兩個個體的特徵應該會類似。因此，當我們需要比較兩群個體差異時，僅需要比較兩群的傾向分數，當分數愈接近時，我們可以認定兩者的相似程度愈高，反之，相似程度愈低。套用在配對分析(Matching)上，我們只要將病例組與對照組的傾向分數算出，然後，直覺式找出傾向分數最相近者，就是我們所要配對結果，這就是傳統的傾向分數配對的想法，簡單而有效。

目前我們所常見估算傾向分數的方式以 Logistic Regression 為主，但也可應用 Simple Linear Regression、Probit Regression、discriminant analysis、classification and regression trees、neural networks (Johnson et al.，2009)。其概念就是藉由線性組合的方式，將多維度資料進行降維動作，而此一降維動作，對於日後進行配對(Matching)工作具有很大的幫助，可以大幅降低因太多共變數而無法配對的情況，在簡化配對變數數量的同時，又可增進其配對效率。

單純傳統的傾向分數配對，相較於一般 Individual Matching，理論上配對效率較高，但是仍無法滿足學者追求極致配對效率之需求，因此，有很多學者紛紛提出改良版本的傾向分數配對法，其方式有 Nearest Neighbor matching、Caliper Matching、Mahalanobis Metric Matching、Stratification Matching、Difference-in-Difference matching、Exact Matching (維基百科資料)。

由於改良種類甚多，我們挑選出一種配對效率很高的傾向分數改良版本-Mahalanobis Metric Matching 搭配 Caliper Method，進行討論與實作配對成果。

二、馬哈蘭距離之概述與應用

馬哈蘭距離(Mahalanobis Distance) 是由印度統計學大師馬哈蘭所提出，主要用於計算在多維度空間之中，點與點之距離量測的一種方式。在此，我們可以延伸多維度空間概念至多變項空間，將一個變項想像成一個維度空間，所以多個變項，則可視為多個維度。借用此一概念，多變量分析裡常見的群集分析(Cluster Analysis)，在測定兩個物體之相似程度時，就是採用「距離」概念，來判別兩物體之相似程度，進一步將其分類。

群集分析的「距離」計算方式有許多方式，如 Euclidean Distance、Minkowski Distance、City Block Distance、Mahalanobis Distance。在此，我們僅介紹 Mahalanobis Distance，其距離量測公式如下：

$$D(\vec{x}) = \sqrt{(\vec{x} - \vec{y})^{-1} \Sigma^{-1} (\vec{x} - \vec{y})}$$

其中，對照組向量為 $\vec{x} = (x_1, x_2, \ldots, x_N)^{-1}$

病例組向量為 $\vec{y} = (y_1, y_2, \ldots, y_N)^{-1}$

共變異數矩陣為

$$\Sigma = \begin{bmatrix} E[(x_1 - y_1)(x_1 - y_1)] & E[(x_1 - y_1)(x_2 - y_2)] & \ldots & E[(x_1 - y_1)(x_N - y_N)] \\ E[(x_2 - y_2)(x_1 - y_1)] & E[(x_2 - y_2)(x_2 - y_2)] & \ldots & E[(x_2 - y_2)(x_N - y_N)] \\ . & . & . & . \\ . & . & . & . \\ E[(x_N - y_N)(x_1 - y_1)] & E[(x_N - y_N)(x_2 - y_2)] & \ldots & E[(x_N - y_N)(x_N - y_N)] \end{bmatrix}$$

在此，傾向分數配對也是一種資料分群的概念，與群集分析的概念相似，我們採用馬哈蘭距離，來分析病例組樣本與對照組樣本的相似程度，並且進行配對。馬哈蘭距離傾向分數與群集分析不同的是，群集分析是將資料分群時，並無特定點作為參考，通常是藉由不斷計算與移動資料中心位置，找出最適參考點之後，再計算各點相對於參考點之距離，挑選出最近距離樣本，作為同一群體。而「馬哈蘭距離傾向分數配對」則是將病例組樣本直接作為參考點，計算其距離，即可找出最相似的對照組樣本。

讀者可能有個疑問，傾向分數透過線性組合的方式，就可組合出一個可供量測相似程度的分數標準，為何還需要馬哈蘭距離來輔助進行相似程度的量測？我們簡單從下面兩圖中，討論這個問題。

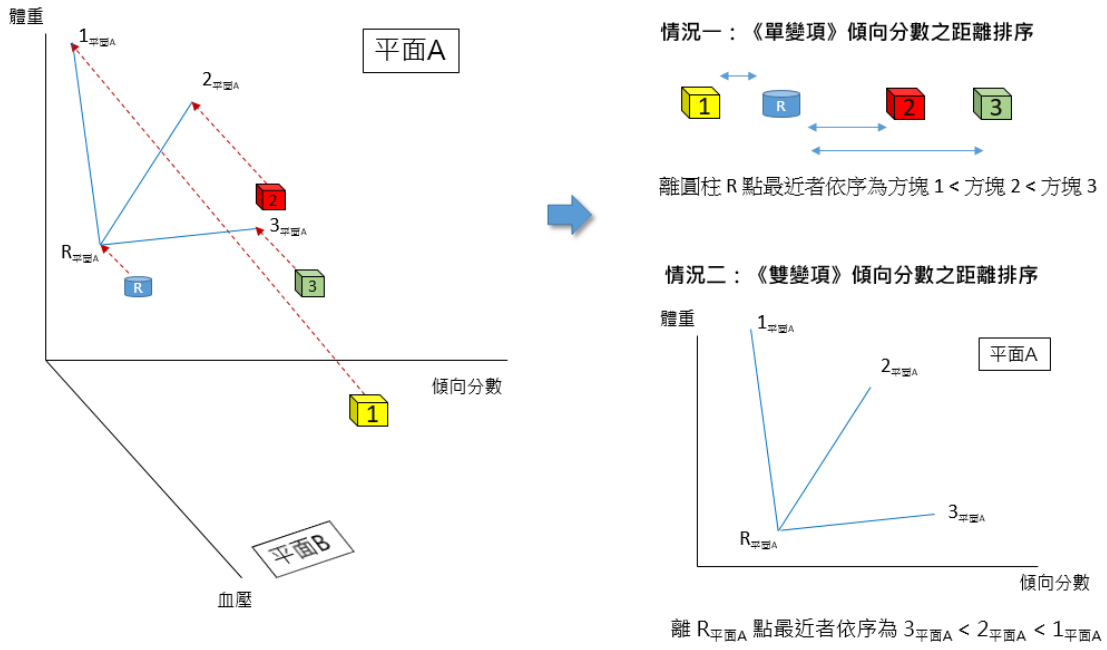
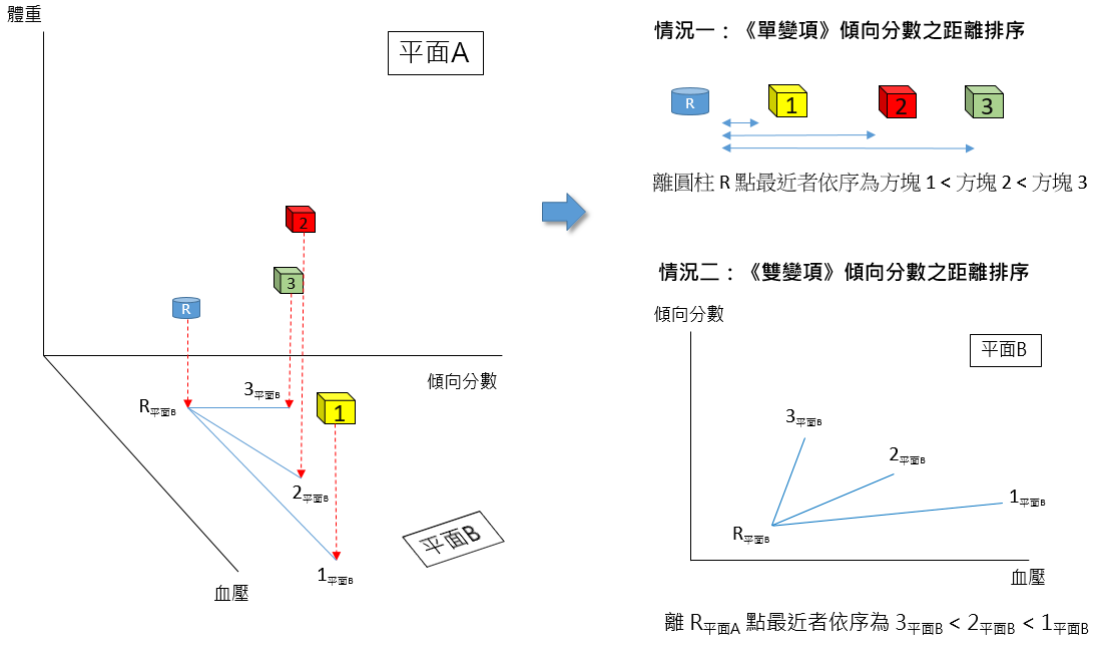


圖 1　單維度與雙維度之距離分析—從平面 A 投影分析

(Note：以藍色圓柱為病例組樣本參考點 R，其餘黃 1、紅 2、綠 3 立方體為對照組樣本點)



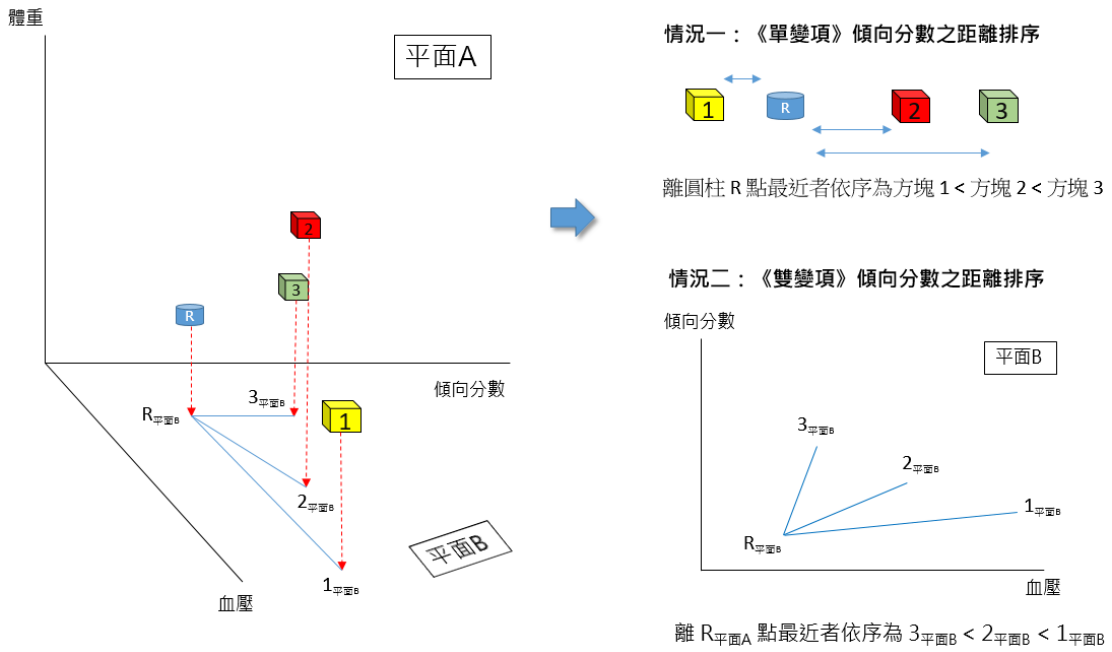


圖 2　單維度與雙維度之距離分析—從平面 B 投影分析

(Note：以藍色圓柱為病例組樣本參考點 R，其餘黃 1、紅 2、綠 3 立方體為對照組樣本點)

我們利用 3 個變數建構出一個立體空間，假設變數分別為傾向分數、體重、血壓。進一步，我們放置 4 個點，分別為病例組參考樣本(藍色圓柱 R)、對照組樣本點 1(黃色立方體 1)、對照組樣本點 1(紅色立方體 2)、對照組樣本點 1(綠色立方體 3)。

傾向分數是一種透過線性組合方式的降維過程，在這個過程中，原本資訊仍被保留在分數之中，我們可透過簡單的分數就可以比較具有多個變項的個體之差異性，就如同圖 1、2 右上角單變量圖形所示，藍色圓柱 R 為病例組樣本之傾向分數，其餘黃、紅、綠立方體為病例組樣本之傾向分數，僅以傾向分數進行比較，可以發現黃色立方體 1 距離藍色圓柱 R 最短，其次為紅色立方體 2，最後則為綠色立方體 3。

但是若是我們多了兩個變項(如體重、血壓)，我們即多出了兩個維度空間，圖形就不為直線，而是一個立體空間。簡化起見，我們分析兩個平面空間，分別為平面 A (體重與傾向分數)以及平面 B(血壓與傾向分數)，我們計算其距離，分別從圖 1 與圖 2 右下角之平面圖形中，可以發現 3 $_{平面\ A}$ 或是 3 $_{平面\ B}$ (綠色立方體之投影平面點)才是離 R $_{平面\ A}$ 或是 R $_{平面\ B}$(藍色圓柱之投影平面點)最近，其次為 2 $_{平面\ A}$ 或是 2 $_{平面\ B}$ (紅色立方體之投影平面點)，最遠為 1 $_{平面\ A}$ 或是 1 $_{平面\ B}$ (黃色

立方體之投影平面點)。

單變量所衡量的距離，在多變量空間的距離並不相同，我們必須考量多維度空間之各點距離的差異性，盡量計算考慮所有維度(變項)後之距離，才能找到最接近的兩點，即是相似程度最高的兩點。

藉由這個小小實驗可知，透過馬哈蘭距離可得知最為真實的相似程度，雖然透過傾向分數可以輕鬆地區分出個體差異，但仍無法有效估算出相似程度。在此，學者提出一種卡尺法(Caliper method)概念，先用來縮小選取範圍，其方式很簡單，就是以病例組樣本的傾向分數為主，上下加減 1/4 個傾向分數標準差，就是選取範圍，再透過馬哈蘭距離比較其範圍內所有樣本點到病例參考點之距離，即可得出最相似的對照組樣本。

由以上論述可知，若是使用「傳統傾向分數配對」可能還是會有配對不佳的情況，所以才需要考慮馬哈蘭距離，增進其配對效率。然而，事實上，情況是否如同上述實驗所示，接下來，我們就開始實務上的驗證分析。

三、傳統傾向分數配對(Original Propensity Score Matching)與卡尺馬哈蘭距離傾向分數配對(Propensity Score Matching with Mahalanobis Metric and Caliper Method)之比較分析

本文介紹 Feng et al. 於 2006 年所發展出來的 SAS 語法，進行卡尺馬哈蘭距離傾向分數配對之示範，以下會概略說明 Feng et al.對於傾向分數配對之實施步驟邏輯，並且參考 Feng et al.(2006)文中所繪製之配對步驟圖形(參見圖 3)，如下所示。

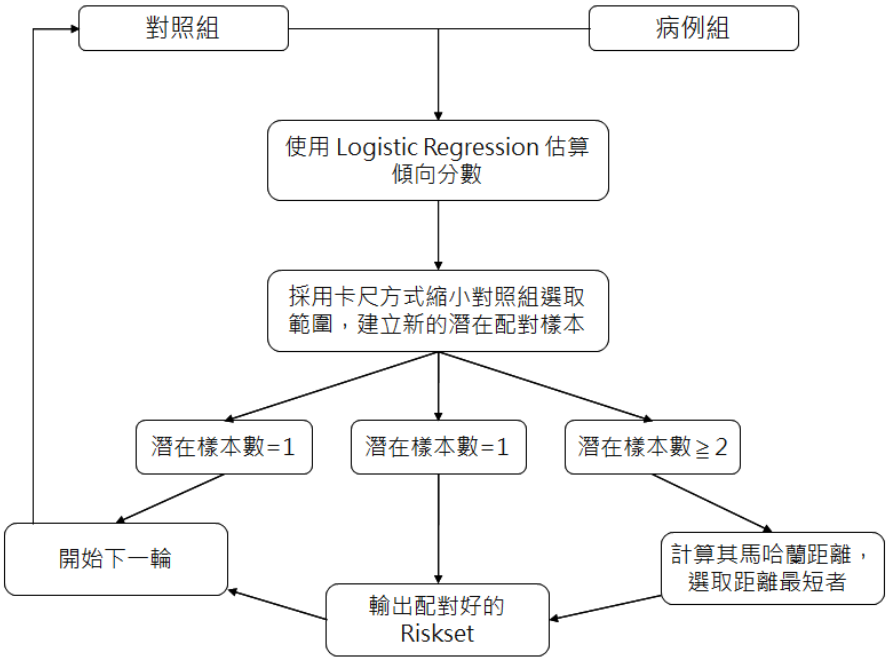


圖 3　Feng et al.(2006)卡尺馬哈蘭距離傾向分數配對之執行步驟

資料來源 : Feng et al.,2006. "A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study". P.3.

STEP 1：將範例資料庫的所有關心變數置入 Logistic Regression，估算出罹患重症肌無力之發生機率，即為傾向分數。而 Feng et al. 透過 logit function，計算其傾向分數之 logit 數值後，藉由此一數值作為比較之基準。

STEP 2：將每一個病例之 logit 數值加上 1/4 標準差後，得出選取區間上下界範圍，進而挑選出對照組 logit 數值落於此一區間之樣本，作為進階需要進行馬哈蘭距離之目標樣本。此一步驟則是卡尺法(Caliper method)之應用，用以縮小需要進一步處理的樣本範圍。

STEP 3：將其所縮小之範圍，以我們所關心的單一病例傾向分數 logit 數值為中心，計算其對照組各點到此一病例點之馬哈蘭距離，其距離之計算除了我們所關心的變數之外，還需加入先前所計算的傾向分數作為變數。計算馬哈蘭距離後，選擇距離最短之對照組樣本，即為最相似該病例點之對照組樣本。

STEP 4: 不斷重複以上 2 個步驟，直到所有病例組之樣本均有進行配對之

動作為止。在當中有可能病例樣本無法配對到合適對照組樣本，則刪除該病例樣本，若只有一個對照樣本進入縮小範圍之中，則直接採用該樣本，作為配對樣本。

Feng et al.(2006)之 SAS CODE 的連結為 http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf。其文件有詳細的使用說明，只需要簡單輸入變數、資料庫名稱即可得出配對結果。在此，本文僅就其程式設計概念與步驟進行說明，以便讀者清楚程式設計內容概念，用以修改增進其程式。

在比較分析「傳統傾向分數配對」與「卡尺馬哈蘭距離傾向分數配對」之前，我們重點式概述傳統傾向分數配對之作法，首先，計算病例組與對照組之傾向分數；其次，根據病例組樣本之傾向分數上下加減某數值(例如：0.005)，作為選取區間；再者，將選取出來的對照組樣本，取傾向分數最接近病例組樣本者，組合成一個 Riskset；最後，將所有 Riskset 集合起來，作為配對後之資料庫。

接下來，我們使用重症肌無力病患範例資料進行以下的傳統與進階傾向分數配對比較分析，此資料包含重症肌無力病患 696 位，其對照組人數則為 14,672 位。下表為重症肌無力資料之病例組與對照組之病患基線特徵，可作為配對前後之比較標準。

我們使用差異性檢定，連續型變數如年齡採用 t 檢定，而類別型變數則是採用卡方檢定。從重症肌無力病患之基線特徵表格之中，可以發現除了肝硬化之外，病例組與對照組的絕大多數變數均呈現統計上的顯著差異性。若是直接採用未配對原始資料進行臨床分析，則會導致分析結果受到干擾因子影響，而有所偏差，無法呈現其清楚的影響關係。因此，我們進一步使用傾向分數配對進行資料選取動作，減低干擾因子之衝擊。

表 1　重症肌無力病患樣本群之基線特徵分布情況

	對照組		病例組		
變數	**個數**	**%**	**個數**	**%**	**P 值**
樣本總數	14,672		696		
病患特徵					
性別					<0.0001
女	7,605	51.83%	422	60.63%	
男	7,067	48.17%	274	39.37%	

年齡	42.437±16.1045		48.731±15.5348		<0.0001
共病症					
高血壓	1,965	13.34%	183	26.29%	<0.0001
糖尿病	849	5.76%	87	12.50%	<0.0001
冠狀動脈疾病	165	1.12%	27	3.88%	<0.0001
心臟衰竭	122	0.83%	16	2.30%	<0.0001
末期腎臟疾病	84	0.57%	10	1.44%	0.0041
腦中風	42	0.29%	10	1.44%	<0.0001
慢性阻塞性肺病	129	0.88%	15	2.16%	0.0006
氣喘	338	2.29%	29	4.17%	0.0015
甲狀腺疾病	52	0.35%	18	2.59%	<0.0001
肝硬化	73	0.50%	5	0.72%	0.4179

我們先從 1:1 配對結果開始分析，下表 2 為「卡尺馬哈蘭距離傾向分數配對」分析結果，可以得知傾向分數配對相當好，兩組傾向分數均為 0.061±0.0493，檢定 P 值相當接近 1，可說是幾乎沒有差異。進一步觀看病患特徵與共病症特徵均無統計上的顯著差異。關於「傳統傾向分數配對」之結果，如表 3 所示，兩組傾向分數有些微差異，但統計上並無明顯差異性，而病患特徵與共病症特徵亦無顯著的統計性質上之差異性。

然而，比較表 2 與表 3 之兩者配對之結果，我們可採用檢定 P 值作為標準化後的統一準則，若是 P 值愈接近 1，則表示病例組與對照組兩者相同的機率愈高，亦代表兩者愈相似，反之，則表示病例組與對照組兩者的差異愈明顯。

因此，比較表 2 與表 3 的檢定 P 值結果，可知「卡尺馬哈蘭距離傾向分數配對」的結果之 P 值均較高，平均而言，在 0.8 左右，而「傳統傾向分數配對」的 P 值較低，平均而言，在 0.47 上下。可見得在 1:1 配對效率方面，「卡尺馬哈蘭距離傾向分數配對」優於「傳統傾向分數配對」。

表 2　卡尺馬哈蘭距離傾向分數配對之 1:1 配對結果

	對照組		病例組		
變數	**個數**	**%**	**個數**	**%**	**P 值**
樣本數量	696		696		
傾向分數	0.061±0.0493		0.061±0.0493		0.9938
病患特徵					
性別					0.7835
女	427	61.35%	422	60.63%	

男	269	38.65%	274	39.37%	
年齡	48.819±15.4645		48.731±15.5348		0.9157
共病症					
高血壓	186	26.72%	183	26.29%	0.8554
糖尿病	84	12.07%	87	12.50%	0.8065
冠狀動脈疾病	25	3.59%	27	3.88%	0.7774
心臟衰竭	14	2.01%	16	2.30%	0.7120
末期腎臟疾病	10	1.44%	10	1.44%	1.0000
腦中風	11	1.58%	10	1.44%	0.8260
慢性阻塞性肺病	11	1.58%	15	2.16%	0.4284
氣喘	25	3.59%	29	4.17%	0.5788
甲狀腺疾病	18	2.59%	18	2.59%	1.0000
肝硬化	5	0.72%	5	0.72%	1.0000

表 3　傳統傾向分數配對之 1:1 配對結果

	對照組		病例組		
變數	**個數**	**%**	**個數**	**%**	**P 值**
樣本數量	693		693		
傾向分數	0.058±0.0449		0.060±0.0442		0.3798
病患特徵					
性別					0.5839
女	409	59.02%	419	60.46%	
男	284	40.98%	274	39.54%	
年齡	47.659±16.9833		48.705±15.5411		0.2319
共病症					
高血壓	152	21.93%	181	26.12%	0.0683
糖尿病	87	12.55%	86	12.41%	0.9352
冠狀動脈疾病	18	2.60%	26	3.75%	0.2203
心臟衰竭	8	1.15%	15	2.16%	0.1411
末期腎臟疾病	9	1.30%	10	1.44%	0.8173
腦中風	14	2.02%	9	1.30%	0.2931
慢性阻塞性肺病	14	2.02%	14	2.02%	1.0000
氣喘	20	2.89%	28	4.04%	0.2399
甲狀腺疾病	9	1.30%	16	2.31%	0.1577
肝硬化	5	0.72%	5	0.72%	1.0000

※ Note：有完整配對到 1 個對照組樣本的 riskset 組數為 693 組，而都沒有配對到對照組樣本的 riskset 組數為 3 組。

進一步，我們分析 1:m 配對效率，當配對數量愈多時，會使得配對效果愈差，所以，我們除了比較兩者配對效果，也比較兩者因配對數量增多時，配對效

率的遞減情況。

由表 4 可知，當 m=4 時，要找到合適的對照組樣本的困難度會增加，所以，比較表 2 的 P 值時，可以發現「卡尺馬哈蘭距離傾向分數配對」的平均 P 值下降了，但是下降幅度不大。相對而言，「傳統傾向分數配對」的平均 P 值則是下降幅度較大，而其中高血壓變數已經呈現對照組與病例組統計上顯著不相同之情況。由 1:4 配對結果可知，「卡尺馬哈蘭距離傾向分數配對」亦優於「傳統傾向分數配對」甚多。

表 4　卡尺馬哈蘭距離傾向分數配對之 1:4 配對結果

	對照組		病例組		
變數	個數	%	個數	%	P 值
樣本數量	2,748		687		
傾向分數	0.058±0.0394		0.058±0.0395		0.9429
病患特徵					
性別					0.7531
女	1,674	60.92%	414	60.26%	
男	1,074	39.08%	273	39.74%	
年齡	48.767±15.600		47.459±15.511		0.8255
共病症					
高血壓	734	26.71%	179	26.06%	0.7281
糖尿病	333	12.12%	84	12.23%	0.9375
冠狀動脈疾病	100	3.64%	24	3.49%	0.8548
心臟衰竭	51	1.86%	14	2.04%	0.7542
末期腎臟疾病	33	1.20%	9	1.31%	0.8159
腦中風	35	1.27%	8	1.16%	0.8179
慢性阻塞性肺病	43	1.56%	14	2.04%	0.3853
氣喘	105	3.82%	28	4.08%	0.7569
甲狀腺疾病	38	1.38%	12	1.75%	0.4763
肝硬化	12	0.44%	5	0.73%	0.3308

※ Note：有完整配對到 4 個對照組樣本的 riskset 組數為 687 組，而有配對到 3 個對照組樣本的 riskset 組數為 3 組，有配對到 2 個對照組樣本的 riskset 組數為 1 組，有配對到 1 個對照組樣本的 riskset 組數為 1 組，而都沒有配對到對照組樣本的 riskset 組數為 4 組。

表 5　傳統傾向分數配對之 1:4 配對結果

	對照組		病例組		
變數	個數	%	個數	%	P 值
樣本數量	2,720		680		

傾向分數	0.053±0.0342		0.056±0.0333		0.1110
病患特徵					
性別					0.1599
女	1,551	57.02%	408	60.00%	
男	1,169	42.98%	272	40.00%	
年齡	47.303±17.342		48.489±15.457		0.1034
共病症					
高血壓	600	22.06%	175	25.74%	0.0409
糖尿病	282	10.37%	81	11.91%	0.2435
冠狀動脈疾病	65	2.39%	22	3.24%	0.2117
心臟衰竭	46	1.69%	13	1.91%	0.6936
末期腎臟疾病	32	1.18%	9	1.32%	0.7533
腦中風	19	0.70%	7	1.03%	0.3757
慢性阻塞性肺病	47	1.73%	13	1.91%	0.7447
氣喘	90	3.31%	27	3.97%	0.3971
甲狀腺疾病	23	0.85%	8	1.18%	0.4168
肝硬化	13	0.48%	4	0.59%	0.7153

※ Note：有完整配對到 4 個對照組樣本的 riskset 組數為 680 組，而有配對到 3 個對照組樣本的 riskset 組數為 1 組，有配對到 2 個對照組樣本的 riskset 組數為 5 組，有配對到 1 個對照組樣本的 riskset 組數為 2 組，而都沒有配對到對照組樣本的 riskset 組數為 8 組。

當然，此一比較分析結果，是基於單次的分析結果，若是要非常精確的比較結果，則需要透過多次蒙地卡羅模擬，如重複 5,000 次的結果，才可以有效判別「卡尺馬哈蘭距離傾向分數配對」優於「傳統傾向分數配對」之確切程度。但是藉由上述簡單的比較，也可以初步得知，「卡尺馬哈蘭距離傾向分數配對」的配對效率的優勢。

參考文獻

1. Baltar, V.T., C.A.D. Sousa., and M.F. Westphal. 2014. “Mahalanobis’ distance and propensity score to construct a controlled matched group in a Brazilian study of health promotion and social determinants”. SciELO Public Health.17. P.668-679.

2. d'Agostino, R.B. 1998. “Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group”. Statistics in Medicine. 17. P.2265-2281.

3. Feng, W.W., J. Yu, and R. Xu. 2006. “A Method/Macro Based on Propensity Score and

Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study". Working Paper.

4. Johnson, M.L., W. Crown, B.C. Martin, C.R. Dormuth, and U. Siebert. 2009. "Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: The ISPOR good research practices for retrospective database analysis task force report—Part III". Value Health. 12. P.1062-1073.

5. King, G., R. Nielsen, C. Coberley, and J.E. Pope. 2011. "Comparative Effectiveness of Matching Methods for Causal Inference". Working Paper.

6. Marcelo Coca-Perraillon. 2007. "Local and Global Optimal Propensity Score Matching". SAS Global Forum.

7. Steiner, P.M., D. Cook. 2013. "Matching and Propensity Scores". The Oxford Handbook of Quantitative Methods. Oxford University Press.

8. Wikipedia."Propensity Score Matching". https://en.wikipedia.org/wiki/Propensity_score_matching