

## 二、研究計畫內容（以 10 頁為限）：

### （一）摘要

根據衛生福利部歷年來的十大死因統計，心臟病經常名列第二名，又冠狀動脈心臟病為最常見，為猝死的主因，近年來更發現患者有年輕化的趨勢，所以若能預測罹患冠狀動脈心臟病的風險機率，則可提早預防。目前最常使用 Framingham Risk Score<sup>[1]</sup>來預測罹患冠狀動脈心臟病的風險機率，然而此方法只使用到年齡、總膽固醇、糖尿病等傳統危險因子，但有越多研究指出脂肪厚度，也就是俗稱的心包油和血管中鈣的沉積的程度，稱為鈣化分數(Calcium Score)有助於預測罹患冠狀動脈心臟病的風險機率。本研究先將心包油和鈣化分數加入 Cox PH 模型去預測風險機率；然而在我們從成大醫院拿到近十年的心臟血管電腦斷層資料可以發現，有部分的患者在研究期間為已治癒的狀況，用先前的 Cox PH 模型可能不適合，所以我們採用混合治癒模型，將患者分為已治癒和未治癒，各別去預測風險機率。最後，在沒有任何統計假設的前提下，使用機器學習方法中的決策樹，用資料本身的特性去預測風險機率，然後再透過模擬，分析並比較三種模型，希望可以找出在不同資料結構下適用的模型，以便於去更好提升預測罹患冠狀動脈心臟病的風險機率。

總結本研究的四大目的，分別為使用 Cox PH 模型、混合治癒模型和決策樹去預測罹患冠狀動脈心臟病的風險機率，再模擬資料並比較不同資料結構下哪種模型較適用。

**關鍵詞：**預測冠狀動脈心臟病風險機率、鈣化分數、心包油、Cox PH 模型、混合治癒模型、決策樹

### （二）研究動機與研究問題

#### 1. 研究動機

在衛福部公布的最新台灣十大死因中，心臟病已經長期位居第二名。在 108 年，便有將近二十萬人死於心臟病，其死亡率甚至高達每十萬人就有 84.2 人死亡。其中冠狀動脈心臟病最為常見之一。冠狀動脈心臟病(Coronary Artery Disease, CAD,此簡稱冠心病)是當冠狀動脈的脂肪堆積到一定的程度，發生粥樣硬化導致動脈失去彈性，管壁變厚、變硬，內腔漸漸變窄甚至堵塞，造成血液不易流通，使心臟肌肉因缺血而壞死。心絞痛、心肌梗塞是常見的冠心病，更是形成猝死的主要原因。

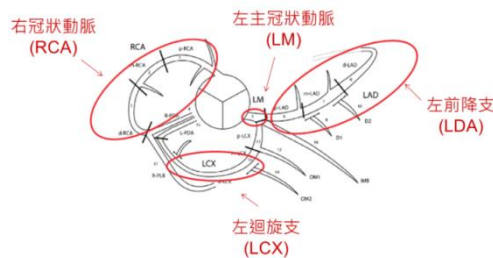
根據衛福部統計，在台灣每年約有一萬多人曾有急性心肌梗塞發作，且心肌梗塞已不再是老年人專屬的疾病，近年來發現患者有越來越年輕化的趨勢。現代人生活習慣改變，還有其他因素如家庭病例史、糖尿病、高血壓、高血脂等問題，皆會影響冠狀動脈硬化，進而增加罹患冠心病的機率。所以冠心病的預防、早期診斷和治療，已受到相當的重視。

而我爸爸也是潛在的冠心病患者，他的爸爸和媽媽都是死於急性心

心肌梗塞。我爸爸現在的狀況是有高血壓、高血糖和高血脂，曾經也有出現胸悶，還有因肥胖導致的呼吸中止。除了我爸爸現在有較高機率罹患心血管疾病，我也可能因遺傳基因而有一定的機率在未來罹患，且已知冠心病有年輕化的趨勢，所以除了想知道普羅大眾都知道的一些生活上的預防，還想更深入探討冠心病。

## 2. 研究問題

在全球，心血管疾病是主要的死亡原因，而冠心病佔所有心血管疾病死亡的一半<sup>[2]</sup>。有約 25% 發生非致命性急性心肌梗塞或猝死的患者室沒有先前的症狀的，所以若能預測罹患冠心病的風險機率則可提早預防。目前最常用來預測冠心病的是 Framingham Risk Score(FSR)，但它只是用了傳統危險因子(包含年齡、性別、種族、高密度脂蛋白、降血脂藥、收縮壓、降血壓藥、糖尿病、抽菸和心臟疾病家族史<sup>[1]</sup>)來預測風險機率，而台灣心臟科權威吳銘庭、林興中和邱寬饒醫師曾研究出傳統危險因子加上左心房室溝脂肪厚度和鈣化分數能有效預測冠心病的風險機率<sup>[3]</sup>，故我們在成大醫院從 2010 年到 2019 年蒐集的冠心病患者資料中，挑出和脂肪厚度與鈣化分數有相關的四大動脈，主要包含三條主動脈，分別為左主冠狀動脈(LM)及其分支左前降支(LAD)和左迴旋支(LCX)，和右側的右冠狀動脈(RCA)。(圖一)為冠狀動脈和心臟的平面示意圖。



(圖一) 冠狀動脈和心臟的平面示意圖

又在此資料中只有不到 1.8% 的死亡，很難聯想到冠心病會是位居前十大死亡的第二名，所以猜測可能原因是此研究的患者有大部分早已痊癒。故本研究有以下四個目的：

### 目的一：

加入傳統因子以外的心包油和鈣化分數，然後使用 Cox PH 風險迴歸模型(Cox Proportional-hazards Regression model)去預測存活機率。

### 目的二：

我們嘗試用混合治癒模型(Mixture cure model)，將患者分為已治癒和未治癒，各別去預測存活機率。

### 目的三：

最後我們會使用機器學習的方法—決策樹，透過資料本身的特性去

預測存活機率。

**目的四：**

希望藉由模擬研究來探討不同的資料結構適用 Cox PH 迴歸分析、混合治癒模型或決策樹分析

**(三) 文獻回顧與探討**

**1. Cox PH 迴歸模型(Cox Proportional-hazards Regression model)**

目前最廣泛被使用的生存分析方法是 Cox PH 迴歸模型(Cox Proportional-hazards Regression model)<sup>[4]</sup>，模型一般可寫成<sup>[5]</sup>

$$h(t|X) = h_0(t)e^{\sum_{i=1}^p \beta_i x_i}$$

其中 $X = (x_1, x_2, \dots, x_p)$ 為解釋變數， $e^{\beta_i x_i}$ 是風險比率(Hazard ratio)， $h_0(t)$ 為基礎危險函數(Baseline hazard function)，也就是所有解釋變數都為 0 的時候，而這個也是此模型的一個重要前提假設—解釋變數對風險效應的風險比率是不隨時間改變的(Time-independent X's)。1972 年 D.R.Cox 提出在此模型中不需對 $h_0(t)$ 做任何估計，但是可以預測解釋變數對風險的效應 $\beta$ 進行參數估計，所以此模型又被稱為半參數化模型(Semiparametric model)。這裡 Cox 使用部分概似法(Partial likelihood)在不估計 $h_0(t)$ 的情況下估計 $\beta$ ，接下來，把所有發生特定事件的機率相乘，就得到了部分概似(Partial likelihood)，用 $L_p$ 表示為

$$L_p = \prod_j \frac{\exp(\beta^T x_{ij})}{\sum_{k \in R_j} \exp(\beta^T x_k)}$$

將 $L_p$ 取對數用 $l_p$ 表示為

$$l_p = \sum_j \beta^T x_{ij} - \sum_j \log(\sum_{k \in R_j} \exp(\beta^T x_k))$$

使用最大概似估計(Maximum likelihood)去估計 $\beta$ ，且已知 $\hat{\beta}$ 的 MLE 可以用 Fisher information matrix 的逆矩陣(Inverse of matrix)計算出來<sup>[6]</sup>。

**2. 混合治癒模型(Mixture Cure model, MCMs)**

上述的 Cox PH 迴歸模型背後的假設是在觀察追蹤時間夠長的情況下，所有患者皆會發生特定事件。但在很多研究中，隨著追蹤時間越長，會有越來越多患者治癒，故在此情況下不適用一般的 Cox PH 模型，更注重治癒率(Cure fraction)，而混合治癒模型(Mixture Cure model, MCMs)目前為最廣泛被使用來看治癒率的模型。1952 年，Berkson & Gage 首次提出了指數分配和固定治癒率的混合治癒模型<sup>[7]</sup>，並假設已痊癒的患者可以活到無窮大，故已痊癒患者的存活函數為 1，而模型可表示為

$$S_{pop}(t) = \pi S(t) + (1 - \pi) \quad (1)$$

$S_{pop}(t)$ 為存活函數， $(1 - \pi)$ 為患者的治癒率， $S(t)$ 為未痊癒患者的存活函數。1982 年，Farewell 提出用邏輯斯迴歸表示治癒率，並假設存活分析呈韋伯分配(Weibull distribution)<sup>[8]</sup>，然後把式(1)改寫成

$$S_{pop}(t) = \pi(z)S(t) + (1 - \pi(z))$$

$S_{pop}(t)$ 為存活函數， $1 - \pi(z)$ 為患者的治癒率，

$\pi(z) = \exp(z'\alpha) / (1 + \exp(z'\alpha))$ ， $z$ 為影響治癒率的共變數向量 (Covariate vector)， $\alpha$ 為邏輯斯迴歸的係數向量，用來估計 $z$ 的效應， $S(t)$ 為未痊癒患者的存活函數。此模型同時允許將共變數加入痊癒與未痊癒患者的存活函數中。令  $O = \{(t_i, \delta_i, z_i, x_i), i = 1, \dots, n\}$  為所觀測的存活資料，而  $\Theta = (\alpha, \beta, S_0(t))$  為未知的參數， $t_i$  為患者的存活時間， $\delta_i$  為設限指標 (Censoring indicator)， $(z_i, x_i)$  分別代表第  $i$  個觀測值影響治癒率與存活函數的共變異數。其概似函數可以寫成

$$L(\alpha, \beta) = \prod_{i=1}^n \{ [1 - \pi(z_i)]^{1-y_i} \pi(z_i)^{y_i} h(t_i | Y = 1, x_i)^{\delta_i y_i} S(t_i | Y = 1, x_i)^{y_i} \}$$
  
 $Y$  為特定事件發生的指示變量， $Y=1$  代表非痊癒， $h(t_i | Y = 1, x_i)$  為非痊癒患者的風險函數。由於在實際資料裡是無法觀察到是否治癒，也就是無法觀察到  $Y$ ，故使用 EM 演算法去估計未知的參數。

### 3. 決策樹(Decision Tree)

決策樹(Decision Tree)是一個多功能但邏輯單純的監督式機器學習模型，可以進行分類，也可以進行無母數迴歸，優點是不需先假設資料的線性關係和分布。決策數比較常見的演算法有 ID3、C4.5、CHAID 和 CART，下面會簡單回顧每個方法的演算法和優缺點<sup>[9]</sup>。首先，先定義決策樹用來衡量資料純度的名詞—熵(Entropy)，用  $I_H(t)$  可以表示為

$$I_H(t) = - \sum_{i=1}^c p(i|t) \log_2 p(i|t)$$

$H$  代表同質性(Homogeneity)。當熵為 0 時，代表資料是純的(Pure)，反之，若熵為 1 表示資料是不純的(Impurity)。

ID3(Iterative Dichotomiser 3)是 Quinlan Ross 於 1986 年提出，適用在離散的資料上，是一種由上而下，透過資訊增益(Information gain)判斷切割前後的熵，熵越低就優先做為切割點，再測試每個節點上的屬性。此缺點是若有某個特徵把每筆資料都切開，會得到很低的熵，但對預測卻沒有幫助，故提出了 C4.5。C4.5 適用離散資料，改善了 ID3，使用資訊獲益比率(Gain ratio)，考量了分支的資訊和整體效益，故比資訊增益更好地決定切割點。卡方自動交叉檢驗法(Chi-squared Automatic Interaction Detector, CHAID)適用於離散資料，使用卡方檢定(Chi-square tests)去計算切割前後 p-value 的差異，差異最大的當作切割點。最後是 CART(Classification And Regression Tree)，由 Breiman 在 1984 年提出，它的特點是同時建立在分類和回歸上，所以可以適用於連續資料，概念是利用基尼係數(Gini index)計算節點中任一一筆資料被分錯的機率。

### 4. Time-dependent ROC & Time-dependent AUC

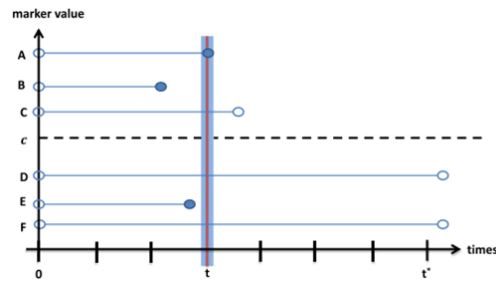
目前 ROC (Receiver Operator Characteristic Curve) 已被廣泛使用在疾病診斷評估上<sup>[12]</sup>，而 AUC (Area under the ROC curve) 為 ROC 曲線下的面積，也是常用的指標，主要概念都為透過真陽性(True positive)、偽陽性(False positive)、真陰性(True negative)和偽陽性(False negative)彼此之間的比例，求出敏感度(Sensitivity)和特異度(Specificity)來評估。然而

傳統的 ROC 曲線是將疾病狀態和觀測值當作固定的，但越多研究有隨時間變化的疾病狀態的資料，故使用時間相關 ROC 曲線分析是更合適的。時間相關 ROC 曲線有三種定義，分別為 cumulative/dynamic(C/D)、incident/dynamic(I/D)和 incident/static(I/S)，其中 cumulative/dynamic 為比較適合用於生存分析的定義，也是目前大多數使用的方法。參考(圖二)<sup>[10]</sup>，當 A、B、E 都為病例組(Cases)，而 C、D、F 為對照組(Controls)時的情況。在閾值為 $c$ 的時候，敏感度和特異度分別用 $S_e^C(c, t)$ 和 $S_p^D(c, t)$ 表示為

$$S_e^C(c, t) = P(X_i > c | T_i \leq t)$$

$$S_p^D(c, t) = P(X_i \leq c | T_i > t)$$

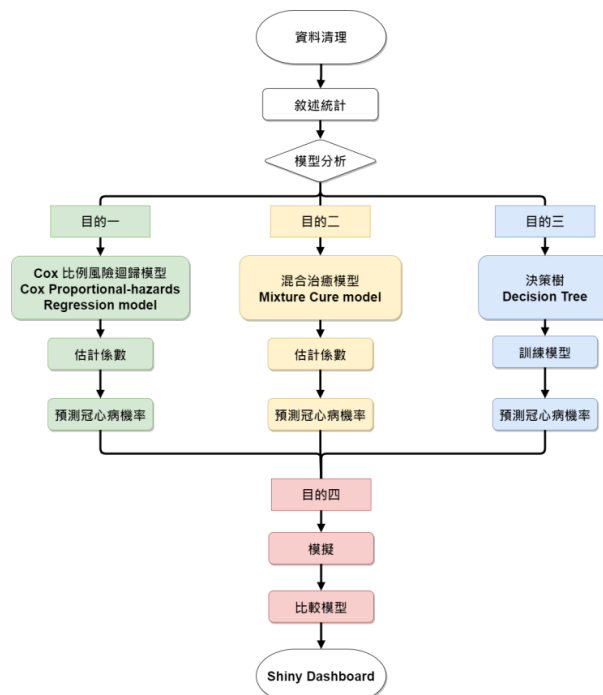
經由上述公式，可以透過不同的閾值，計算出不同的敏感度和特異度，就可畫出時間 $t$ 下的 ROC 曲線。



(圖二)

#### (四) 研究方法與步驟

##### 1. (圖三)為研究流程圖。



(圖三)研究流程圖



## 2. 資料清理

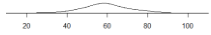



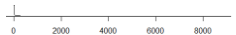


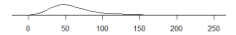
此筆資料是來自成大醫院胸心血管放射線科蔡依珊主任所提供的心臟血管電腦斷層資料，從 2010 年至 2019 年共有 4873 筆患者資訊(重複測量患者當成獨立個體)與 147 個變數。除了傳統危險因子，我們特別想針對四大動脈的狹窄程度和鈣化分數、存活時間和死亡的關係。

首先，我們的資料將四大動脈分為 0 到 5 的狹窄程度，0 為最沒有狹窄，5 為最狹窄。針對遺失值的部分，我們採取眾數的方式填補。又左迴旋支(LCX)有分成前、後段的狹窄程度，而左前降支(LDA)和右冠狀動脈(RCA)則分成前、中、後段的狹窄程度，我們將各段的狹窄程度相加當成各個主動脈的總狹窄程度。再來，由於此筆資料只有檢測日沒有研究結束時間，故取最新的日期 2020 年 2 月當結束日期，計算存活時間的變數"survival month"。








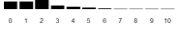

## 3. 敘述統計

(表一)和(表二)為資料清理完後重要的變數

(表一)連續變數說明

變數	描述	遺失值	平均	標準差	中位數	最小值	最大值	分布圖
age	年齡	0	59.34	10.85	58.9	18.4	101.5	
LM cac	左主冠狀動脈的鈣化分數	0	11.64	47.87	0	0	1059.6	
LDA cac	左前降支的鈣化分數	0	74.22	206.19	0.1	0	4467.2	
LCX cac	左迴旋支的鈣化分數	0	31.18	135.38	0	0	3124	
RCA cac	右冠狀動脈的鈣化分數	0	70.81	298.96	0	0	8827.1	
sbp	收縮壓	98.5%	115.2	38.73	124.5	0	197	
total cac	總鈣化分數	0	187.42	600.77	5.22	0	17477.7	
survival month	自檢測日至 2020 年 2 月的存活月數	0	49.73	27.98	44	0	115	
EAT	心包油量化指標	0	57.4	26.2	53.29	0.02	243.3	

(表二)離散變數說明

變數	描述	個數	比例	分布圖	變數	描述	個數	比例	分布圖
Gender	性別	4873			RCA	右冠狀動脈狹窄程度	4873		
0	女性	1611	33.1%		0	無狹窄	637	13.1%	
1	男性	3262	66.9%		1		541	11.1%	
LM	左主冠狀動脈狹窄程度	4873			2		1010	20.7%	
0	無狹窄	1411	28.9%		3		1172	24.1%	
1		2977	61.1%		4		597	12.2%	
2*		483	10%		5		347	7.1%	
LDA	左前降支狹窄程度	4873			6**		569	11.7%	
0	無狹窄	413	8.5%		Death		4873		
1		690	14.1%		0	未死亡	4783	98.2%	
2		1209	24.8%		1	死亡	90	1.8%	
3		798	16.4%		hl		4873		
4		589	12.1%		0	無高血壓	1431	29.4%	
5		442	9.1%		1	有高血壓	3442	70.6%	
6**		732	15%		htn		4873		
LCX	左迴旋支狹窄程度	4873			0	無高血壓	2739	56.2%	
0	無狹窄	1127	23.1%		1	有高血壓	2134	43.8%	
1		1136	23.4%		dm		4873		
2		1404	28.8%		0	無糖尿病	3841	78.8%	
3		513	10.5%		1	有糖尿病	1032	21.2%	
4***		693	14.2%						

\*LM 狹窄程度 $\geq 2$ \*\*LDA 和 RCA 狹窄程度 $\geq 6$ \*\*\*LCX 狹窄程度 $\geq 4$ 

由於糖尿病(sbp)的遺失值過多，所以在建模時不適合當解釋變數。

#### 4. Cox 迴歸模型建構

此研究為追蹤 4873 位患者自檢測日到死亡之間的存活時間，定義發生的特定事件為死亡與否。接下來，建立 Cox 比例迴歸模型，模型如下：

$$h(t|X) = h_0(t)e^{\beta^T x}$$

將資料放入模型並使用最大概似估計法估計出 $\beta$ 。然而使用 Cox 比例迴歸模型忽略掉的 $h_0(t)$ ，會使我們無法預測冠心病的風險機率，故必須預測生存曲線(Survival curve)，而在此之前必須先估計 $h_0(t)$ 。假設基線

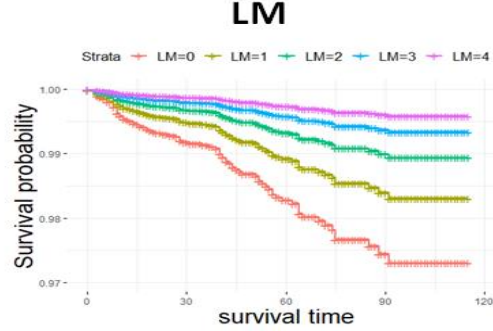
組(Baseline group)患者的解釋變數為 $X^* = (x_1^*, x_2^*, \dots, x_p^*)^T$ ，而在時間段 $t_k \leq t < t_{k+1}$ 上的基礎累積風險(Baseline cumulative hazard)為

$$\widehat{H_0}(t) = \sum_{j=1}^k \frac{d_j}{\sum_{l \in R_j} \exp\{\hat{\beta}^T(x_l - x^*)\}}$$

$d_j$ 是在時間 $t_j$ 時發生死亡的事件數量，接者可以進一步估計生存函數

$$\hat{S}(t; x) = \exp\{-\widehat{H_0}(t) \exp(\hat{\beta}^T x)\}$$

最後就可以畫出 Cox PH 模型估計出的生存曲線，這裡先畫左主冠狀動脈(LM)的生存曲線，如(圖四)所示



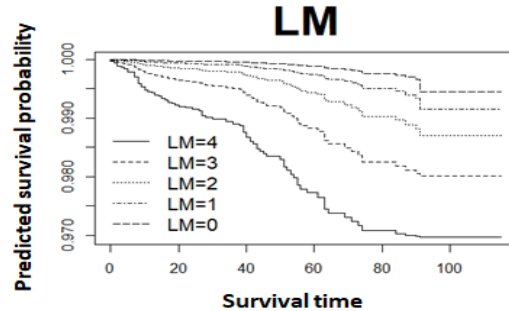
(圖四)左主冠狀動脈(LM)在 Cox PH 模型下的存活函數

## 5. 混合治癒模型建構

令 $O = \{(t_i, \delta_i, z_i, x_i), i = 1, \dots, 4873\}$ 為所此研究的存活資料，而令 $\theta = (\alpha, \beta, S_0(t))$ 為未知的參數。 $t_i$ 為存活的月數(Survival month)，若在2020年2月前追蹤為死亡， $\delta_i$ 定義為1，若追蹤為存活， $\delta_i$ 則定義為0。 $(z_i, x_i)$ 分別代表第 $i$ 個觀測值影響治癒率與存活函數的共變異數， $\alpha$ 為邏輯斯迴歸的係數向量，用來估計 $z$ 的效應， $\beta$ 為非痊癒患者存活函數的係數向量， $S_0(t)$ 為未痊癒患者在 $x$ 都為0時的基礎危險函數。接著建立混合治癒模型，模型如下：

$$s_{pop}(t|x, z) = \pi(z)S(t|x) + (1 - \pi(z))$$

利用其概似函數去估計出 $\alpha$ 和 $\beta$ 。最後，一樣先畫出左主冠狀動脈(LM)在混合治癒模型下的生存曲線，如(圖五)所示



(圖五)左主冠狀動脈(LM)在混合治癒模型下的存活函數

## 6. 決策樹建構

此研究目的是要預測冠心病的風險機率，屬於連續資料，故使用決



策樹裡 CART 演算法。又此資料的死亡率只有 1.8%，所以可能會有過度配適的問題，故之後還需要研究該如何改善決策樹。

## 7. 模擬資料

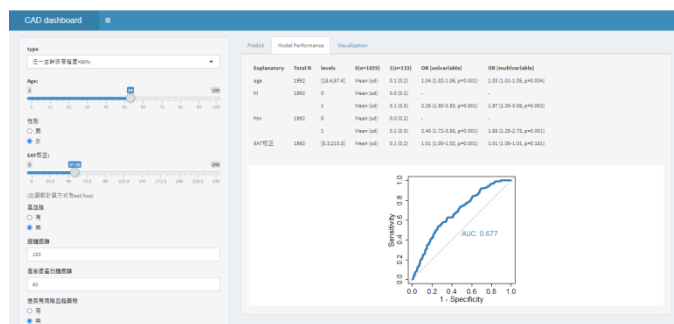
我們將模擬不同的資料結構，分別去配適上述三種模型。首先先假設時間資料呈韋伯分配，並將存活時間模擬出三種結構，分別為正偏態、負偏態和常態，在樣本資料為 25、50、75 和 100 的設定下，重複試驗 1000 次。

## 8. 比較模型

本研究將採取模擬資料的 75% 的當訓練資料(Training data)，25% 當測試資料(Testing data)，分別用 Cox PH 模型、混合治癒模型和決策樹去預測罹患冠心病的風險機率。然後使用 time-dependent ROC 和 time-dependent AUC 來評估模型並比較三種模型，選出在不同資料結構下最適用的模型。

## 9. Shiny Dashboard

(圖六)為此計畫目前做的 Shiny dashboard



(圖六) Shiny dashboard

## (五) 預期結果

直觀來說，可能會覺得越狹窄、鈣化分數越高，罹患冠心病的機率也會越高，死亡率可能也較高，但由(圖四)和(圖五)可看出左主冠狀動脈越狹窄(LM=4)反而生存機率越高，而且確實使用混合治癒模型後不同狹窄程度有更明顯的差異，所以這裡更值得深入研究到底是什麼因素較會影響冠心病和死亡。

## (六) 參考文獻

- [1] McClelland, R. L., Jorgensen, N. W., Budoff, M., Blaha, M. J., Post, W. S., Kronmal, R. A., ... & Burke, G. L. (2015). 10-year coronary heart disease risk prediction using coronary artery calcium and traditional risk factors: derivation in the MESA (Multi-Ethnic Study of Atherosclerosis) with validation in the HNR (Heinz Nixdorf Recall) study and the DHS (Dallas Heart Study). *Journal of the American College of Cardiology*, 66(15), 1643-1653.
- [2] Neves, P. O., Andrade, J., & Monção, H. (2017). Coronary artery calcium score: current status. *Radiologia brasileira*, 50(3), 182-189.

- [3] Ming-Ting Wu, Xing-Zhong Lin, Kuan-Rau Chiou(2011), Quantitative CT Study on the Role of Epicardial Adipose Tissue in Coronary Artery Disease and Metabolic Syndrome: Adding assessment of regional-specific epicardial fat thickness on coronary artery calcification computed tomography improves the prediction of obstructive coronary artery disease.
- [4] Kumar, D., & Klefsjö, B. (1994). Proportional hazards model: a review. Reliability Engineering & System Safety, 44(2), 177-188.
- [5] Kleinbaum, D. G., & Klein, M. (2012). The Cox proportional hazards model and its characteristics. In Survival analysis (pp. 97-159). Springer, New York, NY.
- [6] Abt, M., & Welch, W. J. (1998). Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes. Canadian Journal of Statistics, 26(1), 127-137.
- [7] Berkson, J., & Gage, R. P. (1952). Survival curve for cancer patients following treatment. Journal of the American Statistical Association, 47(259), 501-515.
- [8] Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 1041-1046.
- [9] Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334-337.
- [10] Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. BMC medical research methodology, 17(1), 1-19.

## **(七) 需要指導教授指導內容**

### **(1) 建立模型**

存活模型的建立和係數的解釋，還有決策數遇到的問題，都需要再和指導教授討論。

### **(2) 臨床知識**

臨床方面的實驗研究和系上所學的實驗設計稍有不同，背後的假設和設計都需要再向指導教授請教。

### **(3) 模擬資料**

如何模擬出好的資料，背後需要的假設和方法需要再請教指導教授。

### **(4) 比較模型**

本計畫使用 time-dependent ROC 去評估模型，但因為該方法在臨床上的應用還很少，故需要再和指導教授討論。