研究成果報告：死亡時間之個人化預測

國立中央大學統計研究所　江村剛志

中譯：施嘉翰（國立中央大學統計研究所）

前言

我的研究所達成的貢獻爲特別針對癌症病患的「個人化醫療」，癌症病患會希望知道自己的存活時間，而醫生可以告知病患他們的癌症分期（例如：第四期）、腫瘤大小（例如：直徑大於 1 公分）或是基因表現（*CXCL12* 之表現），但是病患們無法適當地解釋這些症狀，因此，我根據病患們的狀態發展了計算其存活機率的公式。我的研究因使用了「關聯結構（copula）」使其與現有之古典預測方法有巨大的不同。

古典方法

Cox 迴歸[1]是用來預測病患存活時間的古典統計方法，假設一位病患有 p 個共變量(covariate)，則其 t 年的存活機率可經由以下公式計算

$$\hat{S}(t) = \exp\left\{\begin{matrix} -\hat{\Lambda}_0(t)\exp(\hat{\beta}_1 \times \text{covariate } 1 \\ +\cdots+ \hat{\beta}_p \times \text{covariate } p) \end{matrix}\right\},$$

其中 $\hat{\beta}$ 以及 $\hat{\Lambda}_0(t)$ 是根據病患之存活資料計算得出。若研究者對於分類良好以及不良預後結果(prognosis)感興趣，則可以檢驗預後指標(prognostic index, PI)

$$\text{PI} = \hat{\beta}_1 \times \text{covariate } 1 + \cdots + \hat{\beta}_p \times \text{covariate } p \text{ 。}$$

低預後指標對應到良好預後結果（高存活機率），而高預後指標對應到不良預後結果（低存活機率），預後指標通常有助於辨識治療方法是否有效。

以下介紹我原創之統計方法，其延伸了古典預測方法以達到更準確且爲個人化之預測。

使用基因表現之存活預測：

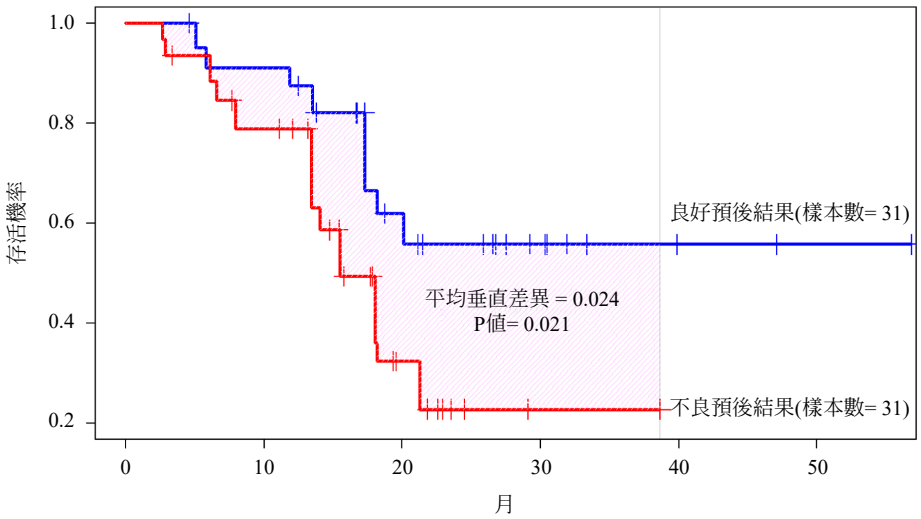
Emura and Chen [2]在使用病患之基因表現的基礎上發展了基於關聯結構(copula)之方法來預測其存活時間，藉由應用此方法至樣本數爲 63 的肺癌病患資料（訓練樣本），我們挑選出了 16 個基因並且建立了預後指標：

$$\begin{aligned}\text{P I} = {} & (0.51\times \textit{ZNF264}) + (0.50\times \textit{MMP16}) + (0.50 \\ & \times \textit{HGF}) + (-0.49\times \textit{HCK}) + (0.47\times \textit{NF1}) + \\ & (0.46\times \textit{ERBB3}) + (0.57\times \textit{NR2F6}) + (0.77\times \\ & \textit{AXL}) + (0.51\times \textit{CDC23}) + (0.92\times \textit{DLG2}) + \\ & (-0.34\times \textit{IGF2}) + (0.54\times \textit{RBBP6}) + (0.51\times \\ & \textit{COX11}) + (0.40\times \textit{DUSP6}) + (-0.37\times \textit{ENG}) \\ & + (-0.41\times \textit{IHPK1})\text{。}\end{aligned}$$

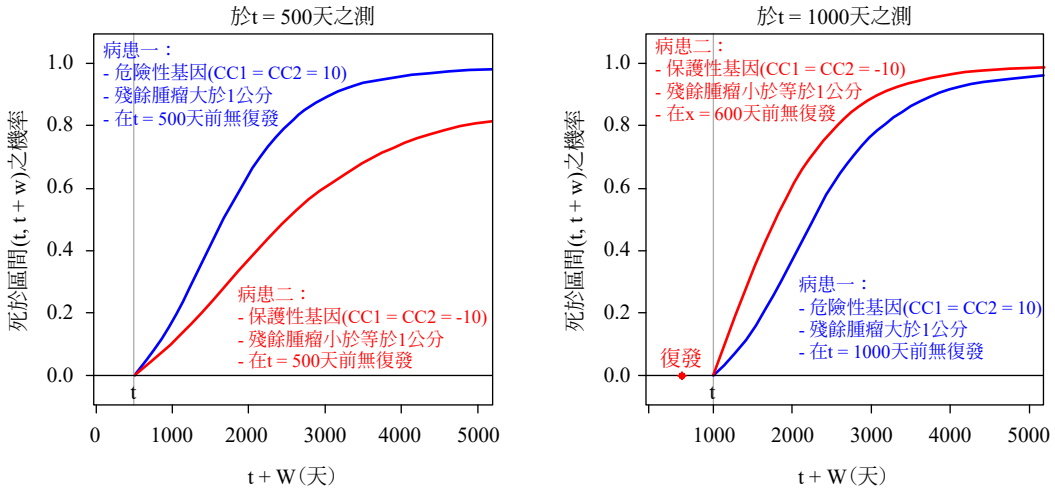
此預後指標應用來預測樣本數爲 62 的肺癌病患資料（測試樣本）之存活時間，圖一顯示預後指標成功地辨識出良好組（低預後指標）以及不良組（高預後指標），兩條曲線在良好組以及不良組之間顯著地分離（平均差異 $=0.224$；P 值 $= 0.021$），此結果證明了基於關聯結構之方法所推導出來的預後指標具有預測能力。

此預測演算法成功的關鍵是使用了關聯結構，首先，在 Clayton 關聯結構[3]下使用半母數最大概似估計量去估計 16 個基因的係數，第二，在圖一中的兩條存活曲線則是根據關聯結構圖像估計量(copula-graphic estimator) [4]進行計算。

R *compound.Cox* 套件[5]中提供了預測演算法，其中包含了基因的選取以及預後指標與關聯結構圖像估計量的計算，Emura and Chen [6]一書中則廣泛地討論了其背後之理論、方法以及程式。



圖一　良好以及不良預後結果之存活曲線，良好（不良）由低（高）預後指標所決定。



圖二　兩位病患（病患一與病患二）在以下設定之預測死亡機率：在預測時間 t = 500 天時兩位病患皆未復發，在預測時間 t = 1000 天時病患二復發而病患一則否。

腫瘤惡化後之存活預測：

若在手術過後癌細胞仍有殘留且生長，則癌症可能會復發，即使癌細胞已被成功地移除，新的腫瘤也有可能在其他不同的地方出現，此現象稱之爲「腫瘤惡化(tumour progression)」。

Emura *et al*. [7, 8]發展了基於關聯結構之方法來預測存活以及腫瘤惡化，藉由應用此方法至樣本數爲 912 的卵巢癌病患資料，我們建立了包含三個預後因素(prognostic factor)的預測公式：

- 術後殘餘腫瘤大小：(大於 1 公分 vs.小於等於 1 公分)
- 預測腫瘤惡化之預後指標（158 個基因）：

$$CC_1 = (0.249 \times CXCL12) + (0.235 \times TIMP2) + (0.222 \times PDPN) + \cdots + (-0.152 \times MMP12),$$

- 預測死亡之預後指標（128 個基因）：

$$CC_2 = (0.237 \times NCOA3) + (0.223 \times TEAD1) + (0.263 \times YWHAB) + \cdots + (-0.157 \times KCNH4),$$

其中兩個預後指標被稱爲複合共變量(compound covariate, CC) [9]，圖二演示了根據我們提出的預測公式所計算出的預測死亡機率。

圖二左顯示出了預測公式可以成功地分離出高危險病患（腫瘤大於 1 公分且高複合共變量）以及低危險病患（腫瘤小於等於 1 公分且低複合共變量），然而，若低危險病患腫瘤復發，則其死亡機會便會高於高危險病患（如圖二右），此結果表示了復發是一個非常重要的危險因子，因此，避免卵巢癌患者腫瘤復發對於增進其預後存活相當重要。

關聯結構在研發此預測演算法中扮演了相當重要的角色，死亡與復發之間的相關結構是由脆弱關聯結構聯合模型(joint fraity-copula model) [7]所描述，藉由將卵巢癌資料配適脆弱關聯結構聯合模型，我們發現死亡與復發之間具有正相關性(Kendall's tau = 0.49；95% CI: 0.32-0.65)，這解釋了爲什麼在腫瘤惡化後病患死亡機率會提升。R *joint.Cox* 套件[10]中提供了預測演算法給醫生以及病患使用，此外，我目前正撰寫一書其將廣泛地討論此預測演算法背後之理論、方法以及程式（此書將出版於[11]）。

參考文獻

[1] D. R. Cox, J R Stat Soc Series B Stat Methodol 34: 187-220 (1972).

[2] T. Emura and Y. H. Chen, Stat Methods Med Res 25(6): 2840-57 (2016).

[3] R. B. Nelsen, *An Introduction to Copulas*, Springer (2006).

[4] L. P. Rivest and M. T. Wells, J Multivar Anal 79: 138-55 (2001).

[5] T. Emura, H. Y. Chen and S. Matsui, *The compound.Cox R package*, CRAN (2018).

[6] T. Emura and Y. H. Chen, *Analysis of Survival Data with Dependent Censoring*, JSS Series in Statistics, Springer (2018).

[7] T.Emura, M. Nakatochi, K. Murotani and V. Rondeau, Stat Methods Med Res 26(6): 2649-2666 (2017).

[8] T. Emura, M. Nakatochi, S. Matsui, H. Michimae and V. Rondeau, Stat Methods Med Res, DOI:10.1177/0962280216688032 (2017).

[9] T. Emura, Y. H. Chen and H. Y. Chen, PLoS One 7(10): e47627. DOI:10.1371/journal.pone.0047627 (2012).

[10] T. Emura, *The joint.Cox* R package, CRAN (2018).

[11] T. Emura, S. Matsui, and V. Rondeau (2018-), Survival Analysis with Correlated Endpoints, JSS Research Series in Statistics, Springer.