國立交通大學

統計學研究所

碩 士 論 文

治癒模式之文獻回顧
Literature Review of the Cure Model

研 究 生：周欣茹

指導教授：王維菁 教授

中華民國 九十三 年 六 月

治癒模式之文獻回顧

Literature Review of the Cure Model

研 究 生：周欣茹　　Student：Hsin-Ru Chou

指導教授：王維菁　教授　　Advisor：Dr. Weijing Wang

國 立 交 通 大 學

統計學研究所

碩 士 論 文

A Thesis

Submitted to Statistics

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

摘　　要

傳統存活分析假設感興趣的事件一定會發生，然而當事件由“死亡”推廣到如“發病”、“復發”… 的情形時，這個假設便不盡合理。文獻修正的方向是承認存在所謂的“免疫者”永遠也不會發生此事件，這類方法統稱為“治癒模式”。本論文回顧治癒模式的部份文獻，將其分成三個方向討論。最常見的治癒模式對免疫者無明確的定義；有一類模式以存活是否超過某固定時間做為免疫的指標；另一方向則以競爭風險發生順序為指標，後兩者對免疫均有明確的定義。所考慮的推論方法，則有迴歸分析、非迴歸分析以及無母數分析三種，論文的重點放在迴歸分析方法的回顧。

ABSTRACT

In classical survival analysis, it is implicitly assumed that the event of interest will occur eventually. However, this assumption may not implausible when there exist a proportion of subjects who will never experience the event despite of long-term follow-up.

In the thesis, we will review the literature on cure models. Three types of cure models are considered according to the definition of “immune” or “cure”. In the first type, there is no explicit definition for the immune. In other words, cured individuals are always mixed with susceptible but censored ones. The second class defines cure as being able to survive beyond a pre-specified time period. For the third type, whether a subject is immune is determined by the order of competing events. Inference methods for cure models include parametric, semi-parametric and nonparametric analysis. The focus here is on semi-parametric regression analysis.

致 謝

這篇論文的完成，代表著我的學生生涯暫且告一個段落。將近二十年的求學過程中，有快樂、有滿足，也有低潮。在這兩年裡，經歷了許多，在此要感謝的人很多。

首先，要感謝我的指導老師王維菁老師，謝謝她這一年來在課業上和生活上對我的指導與幫助，尤其在我經歷最低潮的時候，老師所給予我的協助，使我能夠順利地度過，才有這篇論文的產生；感謝所上每位教授在這兩年來的教導以及學長姐、郭姐和靜怡這段時間以來對我們的照顧，還有班上同學們在日常生活上相互關心與支持；最後要感謝我的家人，我們一起度過了最艱難的時期。

即將邁入人生中的另一階段，未來在新的環境底下，希望能將所學有所發揮。

周欣茹僅誌

中國民國九十三年仲夏

于交通大學統計所

目錄

第一章 序論

傳統的存活分析假設感興趣的事件一定會發生，可以參考近年出版的教科書包含 Anderson et al.〔1993〕、Hougaard〔2000〕、Klein and Moeschberger〔1997〕...等。當感興趣的事件由“死亡”拓展到“發病”、“復發” … 時，傳統存活分析的假設便不盡合理。因此文獻上出現所謂“治癒模式”〔cure model〕承認只有部分的人會發生感興趣的事件，稱之為“可致病”〔susceptible〕；但有部份的個體永遠不會發生此事件，稱之為“免疫”〔immune〕。許多治癒模式的文獻採用混合模式〔mixture model〕做為分析的架構。混合模式的做法是將母體區分為兩類：免疫〔immune 或 cure〕和可致病〔susceptible〕。令T為到感興趣事件的發生時間，定義指標函數B，當$B=1$代表可致病，此時$T<\infty$；當$B=0$代表免疫，此時永遠觀察不到事件的發生，那麼可令$T=\infty$。可以將T的分佈拆解為

$$\Pr(T>t)=\Pr(T>t\mid B=1)\Pr(B=1)+\Pr(B=0)\text{。}$$

上式的右邊將T的分佈分成兩個部分：一部分是$\Pr(B=1)$，稱之為致病模式〔cumulative incidence model〕；另一部分為$\Pr(T>t\mid B=1)$，稱之為潛在發病時間模式〔latency model〕。可發現當免疫者存在時，T的分佈並不符合一般隨機變數的機率性質，因為

$$\lim_{t\to\infty}\Pr(T>t)=\Pr(B=0)=1-p>0\text{。}$$

我們可令$1-p$為 cure rate，稱之為治癒率或是免疫率，這個參數則是許多文獻感興趣的量。

統計推論的目的是根據所觀測的資料，對感興趣的參數做推論，內容包含估計和檢定…。若是在有限的時間點 t 已觀察到感興趣的事件發生，則可確知$B=1$且$T<t$。但是若是觀察不到事件發生，則有兩個可能情形 — 一是永遠免疫或是治癒，機率為$\Pr(B=0)$；另外一個只是暫時尚未發生，機率為$\Pr(T>t,B=1)$。文獻中最常見的治癒模式〔稱之為第一類模式〕並未直接定義如何區辨免疫的

觀測值和暫時設限的觀測值。第二類型的治癒模式則是將“治癒與否”定義為是否存活超過一個固定的年限(假設M年)，因此治癒率就是超過M年的存活機率$\Pr(T > M)$。第三類型的治癒模式把治癒與否定義為競爭風險的發生種類或是發生次序。我們將在第二章到第四章以這三個方向介紹不同類型的治癒模式，討論的重心在迴歸分析。混合模式所提供的架構可供研究者探討解釋變數對於“免疫與否”和“發病時間”是否存在不同的影響。在第五章我們提出對整篇論文的心得以及未來可能的研究方向。

第二章: 第一類治癒模式-無法直接辨識免疫者

第一類治癒模式只假設能觀察到部份有致病可能〔susceptible〕的個體，但無法由設限資料中分辨誰是免疫者。令發生感興趣事件發生的時間為T，但是只有部份的個體會發生此事件，以指標函數$B=1$代表這類人;免疫者以$B=0$表示，其發生事件的時間可令為$T=\infty$。在設限的情形下存在設限時間C，使得只能觀察到$X=T\wedge C$與指標函數$\delta=I(T\le C)$。假設樣本為隨機樣本，右設限的資料可表示為$\{(X_i,\delta_i)\ (i=1,...,n)\}$，可知若$\delta_i=1$則$B_i=1$;然而若$\delta_i=0$則不知$B_i$的值。然而若$X_i=C_i$的值很大，則當$\delta_i=0$時可以猜測$B_i=0$，這個概念和之後會定義的“充分追蹤時間”〔sufficient follow-up〕有關。

根據前述混合模式:

$$\Pr(T>t)=\Pr(T>t\mid B=1)\Pr(B=1)+\Pr(B=0)\text{。}$$

因為 $\lim_{t\to\infty}\Pr(T>t\mid B=1)\to 0$，所以$\lim_{t\to\infty}\Pr(T>t)\to\Pr(B=0)$。因為$T$的樣本符合傳統右設限資料，Maller and Zhou〔1992〕提議以無母數Kaplan-Meier估計量估計$\Pr(T>t)$，所建議的無母數估計量可以表示為

$$\hat{\Pr}(T>t)=\prod_{u\le t}\{1-\frac{\sum_{i=1}^{n}I(X_i=u,\delta_i=1)}{\sum_{i=1}^{n}I(X_i\ge u)}\}\text{。}$$

以$\hat{\Pr}(T>t)$做為$\Pr(T>t)$的估計量的優點早是存活分析領域被熟知的結果，然而以$\hat{\Pr}(T>t_{\max})$做為$\Pr(B=0)$的估計量的合理性〔$t_{\max}$為最大觀測到發生事件的值〕，卻需要特殊條件才會成立。Maller與Zhou〔1992〕的文章強調此估計量只有在充分追蹤時間〔sufficient follow-up〕成立時才會合理，否則會有高估$\Pr(B=0)$的情形。他們更在1994年的論文中提出檢驗充分追蹤時間是否成立的檢定方法，更在1996年的專書中將第一類治癒模式做有系統的討論。我們將在2.3節中摘錄Maller & Zhou〔1996〕書中的部份結果。Li et al.〔2001〕提及在第一類模式下，以無母數方法處理免疫問題因為缺乏額外資訊一般有不可辨識

的問題。Wang〔2004〕則具體指出當死亡這個無法避免的競爭風險存在時，無母數的估計方法有不可辨識性的問題。

當存在有解釋變數時，研究的重點在探討解釋變數對於“免疫與否”和“發病時間”是否存在不同的影響。令 Z 為解釋變數，迴歸模式架構於 incidence 的部份表示為：$\Pr(B=1|Z)$；於 latency 部份可表示為：$\Pr(T>t|B=1,Z)$。通常 $\Pr(B=1|Z)$ 以 binary regression 模式描述，如 logistic regression；針對 $\Pr(T>t|B=1,Z)$ 的模式，有的作者給定母數分配的假設，有的採半母數模式，有的則以無母數方法分析之。例如 Farewell〔1982〕利用 logistic/Weibull 做為兩部分的模式假設。有數篇文章則假設 logistic/Cox 模式，如 Kuk & Chen〔1992〕，Sy & Taylor〔2000〕，Peng & Dear〔2000〕。我們在以下數節中先介紹迴歸模式的推論方法，之後再討論無母數的推論問題。

2.1 Logistic/Weibull 模式

Farewell 在 1982 年的文章中針對致病模式的部分以 logistic regression 模式描述如下：

$$p_i(\beta)=\Pr(B_i=1|Z_i)=\frac{\exp(\beta' Z_i)}{1+\exp(\beta' Z_i)}\ ,$$

對於潛在發病模式的部分則是做 Weibull 的母數分配假設，

$$\Pr(T>t_i\,|\,B_i=1,Z_i)=\exp[-(\lambda t_i)^{\gamma}]\ 。$$

值得注意的是 Farewell〔1982〕的論文假設 $\Pr(T>t_i\,|\,B_i=1,Z_i)$ 與解釋變數無關，但是到了 1986 的論文則將解釋變數的影響納入 Weibull 分配的參數之中。根據傳統治癒模式的資料型態 $\{(\delta_i,X_i,Z_i),\mathrm{i}=1,...,\mathrm{n}\}$，可以得到概似函數為

$$L_F(\lambda,\gamma,\beta,x_i,z_i)=\prod_{i=1}^{n}[p_i(\beta)\gamma\lambda(\lambda x_i)^{\gamma-1}\exp(-(\lambda x_i)^{\gamma})]^{I(\delta_i=1)}$$

$$\times[p_i(\beta)\exp(-(\lambda x_i)^{\gamma})+(1-p_i(\beta))]^{I(\delta_i=0)}\}\ 。$$

因為取了對數之後再對參數微分計算太過複雜，所以利用 EM 演算法來簡化計

算。首先建構“完整資料”$\{(B_i,\delta_i,X_i,Z_i)\ (i=1,...,n)\}$之下的概似函數：

$$L_C(\lambda,\gamma,\beta)=\prod_{i=1}^{n}[p_i(\beta)\gamma\lambda(\lambda x_i)^{\gamma-1}\exp(-(\lambda x_i)^{\gamma})]^{I(\delta_i=1)}$$

$$\times[p_i(\beta)\exp(-(\lambda x_i)^{\gamma})]^{I(\delta_i=0,B_i=1)}[(1-p_i(\beta))]^{I(\delta_i=0,B_i=0)}\}\ 。$$

在 E-step 中對$\log L_F(\lambda,\gamma,\beta)$取條件期望值〔在給定觀測值下〕，其中牽涉到以下等式的計算：

$$E[I(\delta_i=0,B_i=1)\,|\,(\delta_i,x_i,z_i)]=\Pr(B_i=1\,|\,\delta_i=0,T_i>x_i,z_i)$$

$$=\frac{\Pr(B_i=1,\delta_i=0,T_i>x_i\,|\,z_i)}{\Pr(\delta_i=0,T_i>x_i\,|\,z_i)}$$

$$=\frac{\Pr(B_i=1\,|\,z_i)\Pr(T_i>x_i\,|\,B_i=1,z_i)}{\Pr(B_i=0\,|\,z_i)+\Pr(B_i=1\,|\,z_i)\Pr(T_i>x_i\,|\,B_i=1,z_i)}$$

$$=\frac{p_i(\beta)\exp(-(\lambda x_i)^{\gamma})}{[1-p_i(\beta)]+p_i(\beta)\exp(-(\lambda x_i)^{\gamma})}\ 。$$

此外可得

$$E[I(\delta_i=0,B_i=0\,|\,(\delta_i,x_i,z_i)]=\Pr(B_i=0\,|\,\delta_i=0,T_i>x_i,Z_i)$$

$$=\frac{1-p_i(\beta)}{[1-p_i(\beta)]+p_i(\beta)\exp(-(\lambda x_i)^{\gamma})}\ 。$$

第二個步驟 M-step 為對$E[\log L_C(\lambda,\gamma,\beta\,|\,data)]$取極大值，可將$E[\log L_C(\lambda,\gamma,\beta\,|\,data)]$對參數微分令其等於 0 求解以得估計值，這部分可以利用牛頓法以求得數值解。值得一提的是$E[I(\delta_i=0,B_i=1)]$與$E[I(\delta_i=0,B_i=0)]$會牽涉到未知參數，估計時為代入前階段之參數之估計值，在求極值的過程中視為定值。

利用母數分配的假設給定致病模式以及潛在發病時間模式，在模式正確的情形下，所得之估計量具有效性，但伴隨的問題是模式假設可能不夠一般化，使得應用性可能不夠廣；此外若是假設錯誤，所得之估計量可能有很大的偏誤因此不夠穩健。所以檢驗模式適合度有其必要性。

2.2 logistic/Cox 模式

Farewell〔1982〕在致病時間模式上，並未將解釋變數的影響考慮其中，這

是一大缺失。為了彌補這個問題，後續文章除了於致病模式做同樣的 logistic 模式假設外，

$$p_i(\beta) = \Pr(B_i = 1 \mid Z_i) = \frac{\exp(\beta' Z_i)}{1+\exp(\beta' Z_i)} \ ;$$

致病時間模式則是利用“等比例風險模式”(Cox proportional hazards model)做為假設：

$$\Pr(T_i > t \mid B_i = 1, Z_i) = \exp\{-\int_0^{t_i} h_0(u) e^{\alpha' z_i} du\} \ ,$$

其中 $h_0(t)$ 為 $B=1$ 群體發病時間的基底風險函數〔baseline hazard function〕。值得一提的是在加入了免疫群體的 Cox model 下，解釋變數對合併的群體就不再是成比例(proportional)的影響。此外，將迴歸分析用在混合模式的好處是可將解釋變數在“發病與否”和“發病時間”的影響分開討論，例如有的解釋變數只對發病與否有影響，有的則可能改變發病時間。

根據資料型態以及 logistic/Cox 模式的假設，概似函數可以寫成:

$$L_F(\alpha,\beta,h_0) = \prod_{i=1}^{n} [p_i(\beta) h_0(x_i) e^{\alpha' z_i} \exp(-\int_0^{x_i} h_0(u) e^{\alpha' z_i} du)]^{I(\delta_i=1)} \qquad (2.1)$$

$$\times [p_i(\beta) \exp(-\int_o^{x_i} h_0(u) e^{-\alpha' z_i} du) + (1 - p_i(\beta))]^{I(\delta_i=0)}\} \ 。 \qquad (2.2)$$

一般 MLE 的求解方法是將概似函數取對數以後再予以微分求解。由上式可發現，概似函數 $L_F(\alpha,\beta,h_0)$ 取對數後，(2.2)式變成為

$$\sum_{i=1}^{n} I(\delta_i = 0) \times \log[p_i(\beta) \exp(-\int_0^{x_i} h_0(u) e^{-\alpha' z_i} du) + (1 - p_i(\beta))] \ 。$$

其中 log 裡為連加的情況，一旦再對參數微分後，整個計算將會變得複雜許多。問題所在來自當 $\delta_i = 0$ 時，概似函數必須傳遞兩種可能性〔即 $B_i = 0$ 和 $B_i = 0$〕，使得函數有連加的部份。可發現若是 B_i 的值為已知，連加的部份會變為連乘，取對數後函數變為線性，求極值時計算亦得以簡化。以上的想法就是

EM 演算法的概念：先假設完整資料以建構簡單的概似函數，對於缺失值的部份以條件期望值估計，再求極值。先前討論的 Farewell 就是以此想法提出估計方法，當模式拓展到半母數的 Cox 模式時，在 E-step 時會牽涉到未知的基底函數 $h_0(t)$，推論的難度在於估計 (α,β) 時如何處理無窮維度的未知函數。

對於 Cox 模式的推論〔不包含免疫者〕，一般以部份概似函數〔partial likelihood function〕估計，文獻發現雖然忽略 $h_0(t)$ 的影響但所得之 α 估計量仍具有相當不錯的表現。然而當有免疫者存在時，若忽略基底風險函數 $h_0(t)$ 的資訊，則會對 α 的估計造成偏誤，代表部份概似函數估計法不可行。因此針對 logistic/Cox 模式有關致病時間模式部份的估計，便面臨到要估計基底風險函數的挑戰。文獻提出了數種方法針對基底風險函數的處理，我們在此簡述其概念。

2.2.1 Kuk and Chen 的邊際概似函數估計法

Kuk and Chen〔1992〕針對 logistic/Cox 模式估計的部分，考慮以邊際概似函數〔marginal likelihood〕的方法估計感興趣的迴歸參數 (α,β)，這部份可以不需處理基底風險函數。概念簡述如下：根據所得的右設限資料型態 $\{(\delta_i,X_i,Z_i)$，i=1,...,n$\}$，可將其分成兩部分，一部分為有設限的觀測值 (C)，另一部分為沒有設限的觀測值 (D)。將資料根據發生事件的時間做排序，再將排序資料做多重積分後建構邊際概似函數，這個做法可以巧妙的將基底函數去除，因此迴歸參數 (α,β) 的估計量為對邊際概似函數求極值獲得，無須在此階段處理基底函數估計。然而這個方法的難度在於多重積分即使經過簡化，仍牽涉大數目的排列組合個數，使得概似函數變成是龐大組合個數的連加，因此作者建議以 Monte Carlo 模擬法求概似函數的近似值後再求解。後續的論文在估計 (α,β) 時，選擇寧可直接面對基底函數存在的事實，而不願意為捨棄這個項目而引來更為複雜的計算問題。

2.2.2 EM 演算法

在 logistic/Cox 模式下的 EM 演算法的概念簡述如下：假設資料型態可表示為$\{(\delta_i, X_i, z_i, B_i), i=1,...,n\}$，此時概似函數可表示為

$$L_C(\alpha,\beta,h_0)=\prod_{i=1}^{n}\{p_i(\beta)^{I(B_i=1)}(1-p_i(\beta))^{I(B_i=0)}\} \quad (2.3)$$

$$\times\prod_{i=1}^{n}\{[h_0(x_i)e^{\alpha' z_i}]^{I(\delta_i=1,B_i=1)}[\exp(-\int_{o}^{x_i}h_0(u)e^{\alpha' z_i}du)]^{I(B_i=1)}\}\text{。} \quad (2.4)$$

則取對數以後可得，

$$\log L_C(\alpha,\beta,h_0)=\sum_{i=1}^{n}\{I(B_i=1)\log p_i(\beta)+I(B_i=0)\log(1-p_i(\beta))\}$$

$$+\sum_{i=1}^{n}\{I(\delta_i=1,B_i=1)[\log h_0(x_i)+\alpha' z_i]+I(B_i=1)[-\int_{0}^{x_i}h_0(u)e^{\alpha' z_i}du]\}\text{。}$$

可發現在$\log L_C(\alpha,\beta,h_0)$中的 log 函數裡不再有相加的問題，求極值的過程中對參數微分時也較好計算。

和前述 Farewell〔1982〕的做法相同：E-step 須對設限的個體求其得病可能性的期望值來作為權數〔weight〕，推導如下：

$$E[I(B_i=1)\,|\,(\delta_i,x_i,z_i)]=I(\delta_i=1)+E[I(B_i=1)\,|\,\delta_i=0,T_i>x_i,z_i]\text{，}$$

在 Cox 模式假設下可得

$$E[I(B_i=1)\,|\,\delta_i=0,T_i>x_i,z_i]=\Pr(B_i=1\,|\,\delta_i=0,T_i>x_i,z_i)$$

$$=\frac{\Pr(B_i=1,\delta_i=0,T_i>x_i\,|\,z_i)}{\Pr(\delta_i=0,T_i>x_i\,|\,z_i)}$$

$$=\frac{\Pr(B_i=1\,|\,z_i)\Pr(T_i>x_i\,|\,B_i=1,z_i)}{\Pr(B_i=0\,|\,z_i)+\Pr(B_i=1\,|\,z_i)\Pr(T_i>x_i\,|\,B_i=1,z_i)}$$

$$=\frac{p_i(\beta)S_0(x_i\,|\,B_i=1)^{\exp(\alpha' z_i)}}{[1-p_i(\beta)]+p_i(\beta)S_0(x_i\,|\,B_i=1)^{\exp(\alpha' z_i)}}\text{，}$$

$$E[I(B_i=0)\,|\,(\delta_i,x_i,z_i)]=1-\delta_i+E[I(B_i=0)\,|\,\delta_i=0,T_i>x_i,z_i]$$

$$=1-\delta_i+\frac{1-p_i(\beta)}{[1-p_i(\beta)]+p_i(\beta)S_0(x_i\,|\,B_i=1)^{\exp(\alpha' z_i)}}\text{。}$$

如前所述在 Cox 模式下這個權數包含未知的基底存活函數，需要額外處理。我們將在之後敘述兩篇文章的處理方法。EM 演算法的第二個部分為 M-step，將對數概似函數的期望值取極大值，其中權數部分經代入前階段的參數估計值可視為定

值，如此可使求極值的步驟不致太複雜。以下簡介兩篇文章以不同方式處理基底函數的估計問題。

(1) EM 演算法 -- Sy and Taylor

Sy and Taylor〔2000〕的文章中提出兩種方法估計基底風險函數。

i. Breslow-type 估計量：令 $H_0(t)$ 為基線累積風險函數，其定義為

$$H_0(t) = \int_0^t h_0(u)du \text{ 。}$$

以 Breslow 的方法估計 $H_0(t)$ 如下:

$$\tilde{H}_0(t \mid B=1) = \sum_{i:t_{(i)} \le t} \left(\frac{d_i}{\sum_{l \in R_i} w_l e^{\alpha' z_l}} \right) ,$$

其中 d_i 為在 t 時間事件發生的個數，分母的部分則是在 t 時間所有處於風險之個體的加權期望機率，此加權的權數為風險比例 $e^{\alpha' z_i}$ ，w_i 為個體被觀察到得病的可能性，若為非設限資料，$w_i = 1$；若為設限資料，$w_i = \pi_i$ ，即 $E[I(B_i = 1) \mid \delta_i = 0, T_i > x_i, z_i]$ 。基線存活函數 $S_0(t \mid B=1)$ 的估計量可由 $\tilde{H}_0(t \mid B=1)$ 求得如下:

$$\tilde{S}_0(t \mid B=1) = \exp(-\tilde{H}_0(t \mid B=1)) \text{ 。}$$

換句話說，Breslow-type 估計量是將未知的基底風險函數利用未知參數 α 和 w_i 來做表示，藉此降低未知參數的維度以簡化估計，代價是在 E-step 裡的權數 w_i 是一個非常複雜的 α 和 β 的函數。在 M-step 中依然視 w_i 為定值〔代入前階段 α 、β 估計值〕，利用 Cox 建議的部分概似函數將(2.4)式簡化後，取極大值，經過疊代的過程，所收斂的解即為參數估計量。

ii. Product-limit 估計量：前述 Breslow 的方法將未知的 $S_0(t \mid B=1)$ 表示成為 α 與 β 的“explicit”函數，以此降低未知參數的維度，卻同時增加函數的複雜度。而 product -limit 則是將基底存活函數利用 product-limit

表示成風險機率參數的連乘，對這些額外的參數再以 NPMLE 的概念求解，其解可表示為α、β的函數。其概念簡述如下：根據 Kalbfleisch and Prentice(1980)的想法，將資料分為設限(C_i)與非設限(D_i)兩部分，(2.4)式可以寫成

$$\prod_{i=0}^{k}\{\prod_{l\in D_i}[h(t_{(i)};z_i)S_0(t_{(i)}^{-}\mid B=1)^{\exp(\alpha' z_l)}]\times\prod_{l\in C_i}S_0(t_{(i)}\mid B=1)^{I(B_l=1)\exp(\alpha' z_l)}\}\ 。 \tag{2.5}$$

根據 PH 模式，基線存活函數可以 Product-limit 的形式表示為

$$S_0(t\mid B=1)=\prod_{j:t_{(j)}\le t}\gamma_j\ ，$$

(2.5)式即可利用 Product-limit 形式改寫為

$$\prod_{i=0}^{k}\{\prod_{l\in D_i}[1-\gamma_i^{\exp(\alpha' z_l)}]\times\prod_{l\in(R_i-D_i)}\gamma_i^{w_l\exp(\alpha' z_l)}\}\ ， \tag{2.6}$$

此處w_l的定義與 Breslow-type 估計量中所提到的相同。換句話說，經過以上的整理，概似函數表示成為α、$\gamma_i\,(i=1,...,k)$與β的函數，其中γ_i為“failure”的個數。要“同時”〔simultaneously〕針對這些函數求極值是似乎是難以達成的任務，因此 Sy and Taylor 建議以下的方式求解：固定α、γ_i的估計函數可表示為

$$\sum_{l\in D_i}\left\{\frac{e^{\alpha' z_l}}{1-\gamma_i^{\exp(\alpha' z_i)}}\right\}=\sum_{l\in R_i}w_l e^{\alpha' z_l},\quad i=1,...,k\ \ 。 \tag{2.7}$$

將以上具有 explicit-form 的 γ_i估計量後再代入概似函數(2.6)式中求α。中間權數w_l部分，牽涉到α、β與γ_i，經過層層疊代可以求得估計量。

(2) EM 演算法--Peng and Dear

Peng and Dear〔2000〕提出的做法與 Sy and Taylor〔2000〕的方法類似，兩者都是先根據完整資料建立對數概似函數，求其條件期望值後再求極值，兩篇論文的差異在 M-step 的處理三種參數α、β和$S_0(t\mid B=1)$的方法不同，此外疊代步驟亦有差異。在 M-step 的部分，Peng and Dear 的做法和前述 Sy and Taylor

所提 Breslow-type 估計量類似，$S_0(t \mid B=1)$化簡為同樣的α與β的函數。但因為(2.4)式較難處理，因此根據 Breslow 的做法，(2.4)式可寫為

$$\prod_{j=1}^{k} \frac{\exp(\alpha' s_j)}{\{\sum_{i \in R_j} w_i \exp(\alpha' z_i)\}^{d_j}}, \qquad (2.8)$$

s_j為所有非設限資料之解釋變數的總合，d_j為在t時間事件發生的個數。M-step 的部分被簡化成對(2.8)式取極大值，(2.8)式中只包含了α這個未知參數，並不含未知的基線風險函數，處理上來也較容易。

在 E-step 的部分，Peng and Dear 的做法則是和 Sy and Taylor 所提 Product-limit 估計量類似，根據 M-step 所估計出來的α、β以及 Kalbfleisch and Prentice(1980)的想法，(2.4)式可寫成(2.6)式的形式，可以得到γ_j $(j=1,\ldots,k)$的估計量

$$\hat{\gamma}_j \approx 1 - \frac{d_j}{\sum_{i \in R_j} w_i \exp(\alpha' z_i)},$$

進而可以估計基底存活函數$S_0(t)$，將$S_0(t)$帶入 E-step 中的w_i，如此疊代數次後即可求的估計量。

2.3 無母數分析

此類的治癒模式的資料和右設限模式毫無區別，然而選擇以傳統存活分析的方法或是以治癒模式的角度卻需要事先做決定。統計分析所牽涉到的問題在於分析方法需要做額外的假設，先前提到的方法是依靠模式的假設，在此節中所提到的無母數分析則需要靠資料品質的假設，才能夠確保估計值的唯一性以避免所謂“不可辨識”〔non-identifiability〕的問題。Farewell〔1986〕建議研究者以應用領域的專業判斷免疫者存在與否做為抉擇的準則。

無母數分析經常使用 Kaplan-Meier 估計量以估計$\Pr(T>t)$：

$$\hat{\Pr}(T>t)=\prod_{u\le t}\{1-\frac{\sum_{i=1}^{n} I(X_i=u,\delta_i=1)}{\sum_{i=1}^{n} I(X_i\ge u)}\}\text{。}$$

文獻中討論 Kaplan-Meier 估計量性質的文章相當多。我們在此節中整理 Maller and Zhou 〔1996〕 書中的結果，討論在治癒模式的假設下使用 $\hat{\Pr}(T>t_{\max})$ 以估計 $\Pr(B=0)=1-p$ 的可行性。經常使用存活分析方法的人都知道 K-M 估計量在估計時間尾端的分配時表現並不穩定。原因之一是當 t 值很大時 $\sum_{i=1}^{n} I(X_i\ge t)$ 的個數變小，這部份位居分母，使得估計的變異性增大。另一個更為相關的問題在於 $\hat{\Pr}(T>t)$ 的範圍只能估計限於能觀測到發生事件的最大時間點，如果觀察時間太短，會有一大部份本來是 susceptible 的人受到設限，此時若以 $\hat{\Pr}(T>t_{\max})$ 估計 $1-p$ 會造成嚴重的高估。在下一小節中我們以較為嚴謹的數學表示法來說明所謂“充分追蹤時間”〔sufficient follow-up〕的問題。

2.3.1 充份追蹤時間的條件

仿照 Maller and Zhou (1996)，定義以下幾個端點:

$$\tau_H=\inf_t\{t:\Pr(X\le t)=1\}\text{，}$$

$$\tau_G=\inf_t\{t:\Pr(C\le t)=1\}\text{，}$$

$$\tau_F=\inf_t\{t:\Pr(T\le t)=1\}\text{。}$$

假設觀察時間有限，$\tau_G<\infty$。在右設限下 $X=T\wedge C$，可知 $\tau_F\wedge\tau_G=\tau_H$。當不存在免疫者時，我們希望 $\tau_G>\tau_F$，這樣 $\tau_F=\tau_H$ 才能有機會觀察到最大可能的事件發生時間。若 $\tau_G<\tau_F$，代表觀察時間太短，會有 heavy censoring 出現，此時 K-M 曲線不會降到 0。然而當免疫者存在時，免疫者的發病時間定義為 $T=\infty$，此時 $p=\Pr(T\le\infty)<1$，T 的分配稱之為“improper”，無論觀測時間如何拉長都會有 $\tau_G\le\tau_F=\infty$。此時可以定義

$$\tau_0=\inf_t\{t:\Pr(T>t\mid B=1)=1\}\text{，}$$

代表可觀察到 susceptible 者最大可能發生事件的時間。此時我們反而希望 $\tau_G > \tau_0$，代表觀察時間充裕到足以使所有 susceptible 的人都有足夠時間發生事件。

Maller and Zhou 〔1996)〕在書中的第二章先將如何檢驗“免疫者是否存在”和“觀察時間是否充足”量化成數學的符號。之後再討論以 K-M 估計量做為檢驗方法的可行性。

<u>第一步:</u> 檢驗 $H_{01}:\Pr(T \le \tau_G)=1$。若是接受 H_{01}，代表觀察時間足以包含所有發生時間，可知 $\tau_F < \infty$ 而且免疫者不存在，觀察時間亦充份。然而若是 H_{01} 被拒絕，無法得知究竟是免疫存在使得 $\tau_F = \infty$ 或者是觀察時間不夠充份，而使得 τ_G 的值太小。此時我們要進行第二步驟的討論。

<u>第二步:</u> 在已知 $\Pr(T \le \tau_G)<1$ 的情形下，檢驗 $H_{02}:\tau_0 \le \tau_G$。若是 H_{02} 成立，代表觀測時間夠長到足以使得所有 susceptible 的個體都發生事件。此時可以確認免疫的族群存在即 $1-p>0$。若 H_{02} 被拒絕，代表研究時間不夠充份〔follow-up is not sufficient〕，無法由資料判別免疫者是否存在。此時如 Farewell 所言，研究者依自己的專業判斷免疫者是否存在，分析時避免以無母數方法而是靠模式假設做分辨“真正免疫者”和“暫時設限者”的工作。

2.3.2 Kaplan-Meier 估計量在尾端的性質

以下內容摘錄自 Maller and Zhou〔1996〕書中的第三章，我們討論如何檢驗前節所提出的兩個假說。討論估計量 $\hat{\Pr}(T > t_{\max})$ 的性質可分為以下幾個方向:

a. <u>*Theorem 3.2:*</u> $t_{\max} \to \tau_H$ *almost surely*。

在 K-M 的估計式中 $t_{\max}$ 被用來估計 τ_H，很顯然 $t_{\max} \le \tau_H$。這個定理說明當樣本數很大時，所觀察到的值會很接近真正被估計的端點。

b. <u>*Theorem 3.4:*</u> *當* $\Pr(T \le t)$ *在* τ_H *是連續的情形下，* $\hat{\Pr}(T \le t_{\max}) \to \Pr(T \le \tau_H)$ *in probability*。

這個結果保證了 $\hat{\Pr}(T \le t_{\max})$ 做為 $\Pr(T \le \tau_H)$ 的點估量具有一致性。值得注意的是最終目標是以 $\hat{\Pr}(T \le t_{\max})$ 估計 p 的可行性，所以我們要討論 p 與 $\Pr(T \le t)$ 在 t 發生於不同界限時的關係。

c. *Theorem 3.5:* *當* $\tau_G \le \tau_F$ *，則* $\tau_G = \tau_H$ *，此時* $\hat{\Pr}(T \le t_{\max}) \to \Pr(T \le \tau_G)$ *in probability*。

當 $\tau_G \le \tau_F$ 觀察時間未能包含最大可能觀測到的發病時間，值得一提的是這個結果未討論免疫者存在的情形。當免疫者不存在時，$\tau_G \le \tau_F$ 是不理想的狀態;然而當免疫者存在時，$\tau_G < \tau_F$ 卻是必然，而我們希望以 $\hat{\Pr}(T \le t_{\max})$ 估計 $p<1$，所以有以下的結果。

d. *Theorem 3.6:* *只有當* $\tau_0 \le \tau_G$ *，才有* $\hat{\Pr}(T \le t_{\max}) \to \hat{\Pr}(T \le \tau_0) = p$ 。

這個結果說明了充分追蹤時間（$\tau_0 \le \tau_G$）的重要性，此時以 Kaplan-Meier 尾端的估計量用來估計免疫的比例才會合理。值得注意的是若 $\tau_G = \tau_H < \tau_0$，則 $\hat{\Pr}(T > t_{\max})$ 會高估 $1-p$，因為若將觀察時間延長會繼續記錄到事件發生，使得 $\hat{\Pr}(T > t)$ 還有下降空間。

2.3.3 Kaplan-Meier 估計量尾端會降到 0 的機率

一旦 $\hat{\Pr}(T \le t_{\max}) = 1$，我們就會得到 $\hat{\Pr}(B = 0) = 0$。當免疫者實際存在時〔尤其當比例很小〕，資料仍有可能得到 $\hat{\Pr}(B = 0) = 0$ 的情形，雖然實際上 $\Pr(B = 0) > 0$。因此 Maller and Zhou 在 3.2 節的主題就以計算了在重複抽樣下會得到 $\hat{\Pr}(T \le t_{\max}) = 1$ 的 frequency。熟悉存活分析的人都知道當最大的觀測值是觀察到的 failure 則 Kaplan-Meier 曲線會降到 0 即 $\hat{\Pr}(T \le t_{\max}) = 1$。令 $\delta_{\max}$ 代表 $t_{\max}$ 對應的指標函數，則

$$\Pr(\hat{\Pr}(B=0)=0)=\Pr(\delta_{\max}=1)。$$

書中的 Theorem 3.9，推導了在 $p\le 1$ 的情形下 $\Pr(\delta_{\max}=1)$ 的理論值; Theorem 3.10，推導了在 $p\le 1$ 的情形下 $\Pr(\delta_{\max}=0)$ 的理論值。這些值都以積分或是連加的方式表示，與變數 p 值和 $T|B=1$ 與 C 的母數分配有關。實際的計算在一般情形下需要用到數值方法。但是若分配是 exponential distribution，則可以得到 explicit 的結果。此外兩個定理的結果包含所有的樣本大小，所以是 finite-sample 的結果。

Theorem 3.11 和 3.12 則是針對 τ_0，τ_F，τ_G 大小的不同排列情形推導 $\Pr(\delta_{\max}=1)$ 在大樣本的極限值。基本上我們要看當 $p=1$ 時，是否 $\lim_{n\to\infty}\Pr(\delta_{\max}=1)=1$，因為當免疫者不存在且樣本很大時，K-M 曲線降到 0 的機率應該是 1 才合理。同理當 $p<1$ 代表免疫者存在，我們應該得到 $\lim_{n\to\infty}\Pr(\delta_{\max}=0)=1$，代表 K-M 曲線不會降到 0 的機率也應該是 1 才合理。我們重述其理論:

Theorem 3.11: *當 $p=1$ 且觀測時間太短以致 $\tau_G<\tau_F$，則 $\lim_{n\to\infty}\Pr(\delta_{\max}=0)=1$。當 $\Pr(B=1)<1$，則亦會得到 $\lim_{n\to\infty}\Pr(\delta_{\max}=0)=1$。*

Theorem 3.12: *當 $p=1$ 且觀察時間夠長到使得 $\tau_G>\tau_F$，則 $\lim_{n\to\infty}\Pr(\delta_{\max}=1)=1$。*

書中約略提到 $\tau_G=\tau_F$ 的情形，需要加入特殊的條件才會使得 $\lim_{n\to\infty}\Pr(\delta_{\max}=1)$ 的極限存在。

2.3.4 Kaplan-Meier 估計量的分配理論與收斂速率

令 $\hat{p}=\hat{\Pr}(T\le t_{\max})$。Maller and Zhou 的 Theorem 4.1 證明只有當觀察時間充份的情形下 $(\tau_0\le\tau_G)$，$\hat{p}\to^{p}\Pr(B=1)=p$。這個定理說明 $\hat{p}$ 具有一致性的充分必要條件就是 $\tau_0\le\tau_G$。此時 $\Pr(T\le t|B=1)$ 的無母數估計量為 $\hat{\Pr}(T\le t)/\hat{p}$，在同樣的條件下 $\hat{\Pr}(T\le t)/\hat{p}$ 具有 uniform consistency 的性質。

以上討論的是 $\hat{p}$ 做為點估計量的性質，欲做後續的推論〔如區間估計和假設檢定〕則需要有關於 $\hat{p}$ 分配理論的推導結果。在 Maller and Zhou〔1996〕書中第四章的 Theorem 4.1 證明了當 $0<p<1$ 的情形下〔即免疫者存在〕，$\sqrt{n}(\hat{p}-p)$ 會收斂到常態分配，此外書中亦提供了 $Var\{\sqrt{n}(\hat{p}-p)\}$ 的公式。然而當 $p=1$，只知 $\sqrt{n}(\hat{p}-1)\to 0$。換言之當免疫者不存在時〔 $p=1$ 〕，$n^{-q}(\hat{p}-p)$ 會收斂到一個 non-degenerate distribution，只知 $\hat{p}$ 的收斂速率比一般 $n^{-1/2}$ 的速率來得快〔即 $q>1/2$ 〕，但是確切的 q 值未知，而且 $n^{-q}(\hat{p}-p)$ 的分配型態亦未知。文獻上稱 $p=1$ 時的性質討論為“boundary problem”，因為 $p=1$ 是比例的上界。在一般統計分析中研究這個問題都是困難的，因為許多有用的數學工具〔如泰勒展開式 …〕在邊界無法使用。因為理論推導的問題牽涉到困難的數學分析能力，我們不做進一步討論。在此我們討論 $p=1$ 的性質在實際問題的應用。

點估計量 $\hat{p}$ 的性質在所有 $p\le 1$ 的情形都適用。但是做進一步的推論卻需要判斷是否免疫者存在。先前提到的第一步驟檢驗 $H_{01}:\Pr(T\le\tau_G)=1$ 就包含檢定免疫者存在的情形。因為若是 H_{01} 被接受了，可以推得 $p=1$；若被拒絕則要繼續做第二步驟 H_{02} 的檢定。Maller and Zhou〔1994〕的論文提出討論檢定充份追蹤時間的方法。欲檢定 H_{01}，可以利用 $\hat{p}-1$ 的距離做為判別標準，然而 $n^{-q}(\hat{p}-1)$ 的分配型態和 q 值在 $p=1$ 卻是未知的。根據 Neyman-Pearson 法則，棄卻範圍由控制型一錯誤 α 決定：$\Pr(|\hat{p}-1|>c_n(\alpha)\,|\,p=1)=\alpha$。換言之若是 $\hat{p}>1-c_n(\alpha)$ 則拒絕 H_{01}，承認免疫不存在。然而 $c_n(\alpha)$ 的值卻因為 $\hat{p}$ 的性質在 $p=1$ 缺乏理論依據所以無法求得，使得以無母數的方法檢驗 H_{01} 的目標遇到很大的挑戰。Maller and Zhou 以模擬分析討論 H_{01} 的檢定。

第三章:以固定時間為切點之治癒模式

有一些文獻把存活超過某時間(M 年)定義為治癒，此時治癒比例為

$$\Pr(B=1)=\Pr(T\le M)\ 。$$

在解釋變數存在的情形下，模式建構於 incidence 部份:

$$\Pr(B=0\mid Z)=\pi(\beta^{'}Z)=\frac{\exp(\beta^{'}Z)}{1+\exp(\beta^{'}Z)}\ 。$$

在沒有設限的情形下，資料可以表示為 $(T_1,...,T_n)$，可自動轉換成為$(B_1,...,B_n)$，其中 $B_i=I(T_i\le M)$，則概似函數方法可表示為：

$$L(\beta)=\prod_{i=1}^{n}\left(\frac{\exp(\beta^{'}Z_i)}{1+\exp(\beta^{'}Z_i)}\right)^{I(B_i=1)}\left(\frac{1}{1+\exp(\beta^{'}Z_i)}\right)^{I(B_i=0)}\ ,$$

對參數微分後可得 score equation：

$$\sum_{i=1}^{n}\{I(T_i\ge M)-\pi(\beta^{'}Z_i)\}\frac{\pi_{\phi}(\beta^{'}Z_i)}{\pi(\beta^{'}Z_i)\bar{\pi}(\beta^{'}Z_i)}Z_i=0\ ,\qquad (3.1)$$

其中 $\pi_{\phi}(\phi)=\frac{\partial\pi(\phi)}{\partial\phi}$，$\bar{\pi}=1-\pi$。然而當設限存在時，我們只能觀察到$\{(X_i,.\delta_i),i=1,...,n\}$，其中 $X_i=T_i\wedge C_i$, $\delta_i=I(T_i\le C_i)$，C_i 代表設限變數，也因此B_i 的值有可能發生未知的情形。當 $\delta_i=1$，很顯然 $B_i=I(X_i\le M)$；當 $\delta_i=0$，此時B_i 的值是未知的。以此為方向的治癒模式並非我們主要的興趣，因此我們只選兩篇文章簡述其概念。

3.1 Jung〔1996〕

Jung 發現$E(I(X_i\ge t)\mid Z_i)=\Pr(T_i\ge t\mid Z_i)G(t)$，其中$G(t)=\Pr(C_i>t)$。因此建議以$I(X_i\ge M)/\hat{G}(M)$做為$I(B_i=0)$的“代理者”(proxy)，修正(3.1)的估計函數可得：

$$\sum_{i=1}^{n}\{\frac{I(X_i\ge M)}{\hat{G}(M)}-\pi(\beta^{'}Z_i)\}\frac{\pi_{\phi}(\beta^{'}Z_i)}{\pi(\beta^{'}Z_i)\bar{\pi}(\beta^{'}Z_i)}Z_i=0\ 。\qquad (3.2)$$

這個方法使用的原則是處理 missing data 常用的“inverse probability weighting”的方法。

3.2 Subramanian (2001)

Subramanian 文章的概念是以 $E[I(T_i \geq M) | \delta_i, X_i, Z_i]$ 的無母數估計量(以 $\hat{E}_i$ 表示之)取代(3.1)式中的 $I(T_i \geq M)$。此文章所做的額外假設為解釋變數 Z_i 為離散變數，如此才可能針對個別可能的 Z 值估計 $\Pr(X_i \geq M | Z_i)$。所提出的估計函數可表示為

$$\sum_{i=1}^{n} \{\hat{E}_i - \pi(\beta' Z_i)\} \frac{\pi_\phi(\beta' Z_i)}{\pi(\beta' Z_i)\overline{\pi}(\beta' Z_i)} Z_i = 0 \ , \qquad (3.3)$$

其中

$$\hat{E}_i = \hat{E}[I(T_i \geq M) | \delta_i, X_i = x_i, Z_i]$$

$$= I(X_i \geq M) + I(X_i = x_i < M, \delta_i = 0) \frac{\hat{\Pr}(T > M | Z_i)}{\hat{\Pr}(T > x_i | Z_i)} \ 。 \qquad (3.4)$$

這個方法所應用的原則被稱為“imputation by conditional mean”。這個方法的缺點在於 $\Pr(T > t | Z)$ 的估計與 Z 有關，所需要的資訊非 $\Pr(B = 0 | Z) = \pi(\beta' Z)$ 的假設所能夠涵蓋。若是解釋變數為離散型，則可以根據 Z 值將資料做切割，再以 Kaplan-Meier 的無母數估計量估計 $\Pr(T > t | Z)$。

第四章:考慮競爭風險之治癒模式

傳統的治癒模式並不直接定義哪個狀態為免疫,因此真正的免疫者是與非免疫的設限資料混合在一起。當模式假設不夠強時〔指在無母數的情形下〕,則強烈的需要依賴資料的“良好品質”(指追蹤時間充份),才能做正確的推論。第三章討論的治癒模式,分析的結果和人為決定的切點有關,所以應用有限。在本章中我們討論第三類型的治癒模式,免疫與否的定義取決於競爭風險發生的種類或是次序。

我們先討論一個簡化的例子,是將 Betensky 與 Schoenfeld。2001〕討論的新生兒因急性肺炎住院改編成因 SARS 住院的例子。令

T_1 = time to hospital discharge (因 SARS 入院到活著出院的時間)

T_2 = time to death (因 SARS 入院到在醫院死亡的時間)。

在此例中,“活著出院”可視為痊癒。也就是說“免疫與否”取決於競爭風險(“死亡”與“出院”)發生的次序,因此若是在沒有外來設限的情形下,“痊癒”是可被觀察到的事件。因為免疫與否有清楚的定義,這樣的架構不致發生前述不可辨識性的問題。

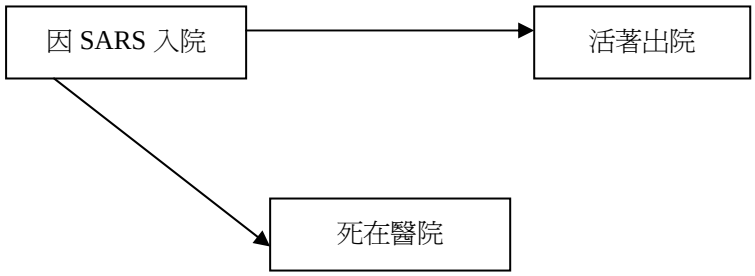

圖 4-1:SARS 入院的例子

以競爭風險的角度來看上述的問題，在這個簡化的例子中，只有兩種失敗型態(failure types)。令 $\tilde{B}=1$，代表“活著出院”的情形；$\tilde{B}=2$，代表“在醫院中死亡”的情形。若令“活著出院”代表痊癒，則無法痊癒的比例為 $\Pr(\tilde{B}=2)$。然而對 $\tilde{B}=2$ 的觀測值，T_1 的定義不明，有的做法是對死在醫院者時，令 $T_1=\infty$；或是只定義 $T_1 \mid \tilde{B}=1$。可得

$$\Pr(T_1>t)=\Pr(T_1>t\mid\tilde{B}=1)\Pr(\tilde{B}=1)+\Pr(\tilde{B}=2)\ 。$$

可以清楚看到

$$\lim_{t\to\infty}\Pr(T_1>t)=\Pr(\tilde{B}=2)\ ，$$

代表不會發生活著出院的比例。值得一提的是，$1-\Pr(T_1>t)=\Pr(T_1\le t)$ 是競爭風險文獻裏經常討論的“累積發生函數”(cumulative incidence function)，代表到 t 時間所累積觀察到發生活著出院的比率。同理可得

$$\Pr(T_2>t)=\Pr(T_2>t\mid\tilde{B}=2)\Pr(\tilde{B}=2)+\Pr(\tilde{B}=1)\ ，$$

其尾端機率 $\Pr(\tilde{B}=1)$ 代表不會死於醫院的比例。

文獻中考慮競爭風險的治癒模式有 Greenhouse and Wolfe (1984)，Larsen and Dinse (1985)，Taylor (1995)，Ng and McLanchlan (1998)，Betensk and Schoenfeld (2001)，Maller and Zhou (2002)，Wang (2004)…等。由發表年代看來，這個研究方向似乎有變得熱門的趨勢。我們將統整這些文章的符號，以便做有系統的整理。這一系列文獻中，經常引用的文獻始自 Greenhouse and Wolfe (1984)，多數的文章討論不只兩種競爭風險。在此，我們由 Larsen & Dinse (1985)開始介紹，因為此文提供的思路為後續文章主要的脈絡。

4.1 Larson-Dinse 的迴歸分析

令 $\tilde{B}$ 代表失敗的型態，假設共有 J 種可能，每一種失敗型態發生的機率定為

$\Pr(\tilde{B}=j)=p_j$ ， $p_1+\cdots+p_J=1$。再定義 $T_j \mid \tilde{B}=j$ 為給定第 j 個事件會發生的時間長度，則條件存活函數為

$$Q_j(t)=\Pr(T_j>t \mid \tilde{B}=j)。$$

Larson-Dinse〔1985〕以迴歸模式來描述 p_j 與 $Q_j(t)$。令 $Z: p\times 1$ 代表解釋變數，在失敗型態的部分用 logistic regression 做為模式的假設，則每一種失敗型態發生的機率可以表示成

$$\Pr(\tilde{B}=j \mid Z)=\frac{\exp(\beta_j^T Z)}{\sum_{j=1}^{J}\exp(\beta_j^T Z)}，$$

其中 $\beta_j=(\beta_{j1},...,\beta_{jp})$；發生第 j 個事件的時間長度，$T_j \mid Z,\tilde{B}=j$，則假設服從 Cox 模式，則條件存活函數可表示為

$$Q_j(t \mid Z)=\Pr(T_j>t \mid \tilde{B}=j)=\exp\{1-\int_0^t h_j(u)\exp(\gamma^T Z)du，$$

其中風險函數 $h_j(t)$ 的型態給定為 piece-wise exponential 模式，可將時間資料分成 M 個區段 $(I_1,\cdots,I_M)$，第 j 個風險型態的風險函數為

$$h_j(t)=\exp(\alpha_{jm}) \quad (t\in I_m, m=1,\cdots,M)，$$

因此在每個區段的 $h_j(t)$ 為常數。在風險函數 $h_j(t)$ 給定的情況下，可用傳統的最大概似法以估計參數。然而若是基底風險函數不給定，估計將因未知參數的維度太高而變得複雜。然而根據第一類治癒模式所發展的方法概念，應可以類推到處理競爭風險的情形。

我們在此回顧基本的推論架構。在沒有設限的情形下，觀察到的資料型態為 $(\tilde{B}_i,x_i,z_i)$， $\tilde{B}_i=1,...,J$ ， $i=1,...,n$ ，則概似函數可以表示為：

$$L(\beta,\gamma,h_j, j=1,...,J)=\prod_{i=1}^{n}\prod_{j=1}^{J} f_j(x_i \mid z_i)\Pr(\tilde{B}=j \mid z_i)\}^{I(\tilde{B}_i=j)}，$$

其中 $f_j(t\mid z)=-\dfrac{\partial Q_j(t\mid z)}{\partial t}$。而在有設限的情況下，有的觀測值無法得知真正的失敗型態，此時所觀測到的時間 $x_i=C_i$，且 $\min(T_{1i},...,T_{Ji})>x_i$。為了傳遞重要的概念，我們假設較簡單的情形，令 $J=2$，所觀察到的資料型態為: (b_i,x_i,z_i)，$b_i=0,1,2\quad(i=1,...,n)$，其中 $b_i=1\Rightarrow\tilde{B}_i=1$，$x_i=T_{1i}$，代表確知痊癒(or immune)；$b_i=2\Rightarrow\tilde{B}_i=2$，$x_i=T_{2i}$，代表確知未痊癒(or susceptible)；然而當 $b_i=0$ 時，$\tilde{B}_i$ 的值未知，即不知道免疫與否，此時 $x_i=C_i<\min(T_{1i},T_{2i})$。設限下的概似函數可以表示為

$$L_F=\prod_{i=1}^{n}[f_1(x_i\mid z_i)p_1(z_i)]^{I(b_i=1)}[f_2(x_i\mid z_i)p_2(z_i)]^{I(b_i=2)}$$
$$\times\{p_1(z_i)Q_1(x_i\mid z_i)+p_2(z_i)Q_2(x_i\mid z_i)\}^{I(b_i=0)}\ 。$$

和第二章的問題類似，最後一項在取對數後再對參數微分的計算會變得太過複雜，因此可用 EM 演算法來估計。假設 $\tilde{B}_i$ 可觀察到完整的資料之概似函數為

$$L_C=\prod_{i=1}^{n}[f_1(x_i\mid z_i)p_1(z_i)]^{I(b_i=1)}[f_2(x_i\mid z_i)p_2(z_i)]^{I(b_i=2)}$$
$$\times\{p_1(z_i)Q_1(x_i\mid z_i)\}^{I(b_i=0,B_i=1)}\{p_2(z_i)Q_2(x_i\mid z_i)\}^{I(b_i=0,B_i=2)}\ 。$$

取對數之後為

$$\log L_C=\sum_{i=1}^{n}\{I(b_i=1)[\log p_1(z_i)+\log f_1(x_i\mid z_i)]$$
$$+I(b_i=2)[\log p_2(z_i)+\log f_2(t_i\mid z_i)]$$
$$+I(b_i=0,\tilde{B}_i=1)[\log p_1(z_i)+\log Q_1(x_i\mid z_i)]$$
$$+I(b_i=0,\tilde{B}_i=2)[\log p_2(z_i)+\log Q_2(x_i\mid z_i)]\}\ 。$$

EM 演算法分為兩個部分，第一部分為先對 $\log L_C$ 取條件期望值，稱為 E-step，也就是求 $E[\log L_C\mid b_i,x_i,z_i]$，其中牽涉到計算

$$E[I(\tilde{B}_i=j)\mid b_i,x_i,z_i]=I(b_i=j)+I(b_i=0)\Pr(\tilde{B}_i=j\mid X=x_i,z_i)$$

$$= I(b_i = j) + I(b_i = 0) w_j(x_i \mid z_i) \text{ ，}$$

在這裡 $w_j(x_i \mid z_i)$ 可表示為

$$w_j(x_i \mid z_i) = \frac{p_j(z_i) Q_j(x_i \mid z_i)}{p_1(z_i) Q_1(x_i \mid z_i) + p_2(z_i) Q_2(x_i \mid z_i)} \text{ 。}$$

第二部分為對對數概似函數的期望值取極大值，稱為 M-step，也就是求 $\max E[\log L_C]$。經過疊代，就可以得到所求的估計值。值得一提的是 $w_j(x_i \mid z_i)$ 牽涉到未知參數的部分為代入前階段之估計值，在求極大值的過程中視為定值。以上是以 2 種失敗型態來作為 Larsen-Dinse 在考慮競爭風險的治癒模式下的觀念闡述，至於 J 種失敗型態也可以利用此觀念加以推廣。

文章分析了著名的心臟移植資料(Stanford Heart Transplant)，希望對於接受心臟移植後的病人(共 65 人)能探討解釋變數對於不同失敗型態的影響。共有兩個失敗型態，其中 $\tilde{B} = 1$ 代表因為發生排斥而死亡，$\tilde{B} = 2$ 代表因為其他原因而死亡。在實際資料中，$\hat{\Pr}(b = 1) = 0.45$（29 人），$\hat{\Pr}(b = 2) = 0.18$（12 人），值得注意的是 $\hat{\Pr}(b = 0) = 0.37$（24 人），代表缺失值頗嚴重。解釋變數的維度 $p = 4$，除了截距項以外還包含“mismatch score”（越大代表捐贈者和接受者組織之相容度差），“age”代表接受移植時的年紀，“waiting time”代表等到心臟移植的時間。其中“mismatch score”和“age”變數都經過標準化，“waiting time”則是轉成是否超過 31 天的 0/1 變數。對於 $Q_j(t \mid Z)$ 的模式採 piecewise-exponential 分配，試了三種配置法:當 $M = 3$ 時，將時間分成 3 個區段 $[0,45),[45,90),[90,\infty)$；當 $M = 2$ 時，將時間分成 2 個區段 $[0,60),[60,\infty)$；另一個選擇是 $M = 1$。

當發現“waiting time”沒有太大影響時這個解釋變數便被捨棄，此時作者只考慮將“mismatch score”和“age”並以 $M = 3$ 的 piecewise-exponential 分配做為 $Q_j(t \mid Z)$ 的模式。分析結果發現這兩個解釋變數對於 $\Pr(\tilde{B} = 1)$ 並無顯著

影響，倒是對於反映在$Q_1(t)$上的排斥時間有顯著影響，年紀越輕且組織 match 較好的人到排斥發生的時間越長。換言之年紀大小和是否因排斥而死亡無關，但若是終究因排斥而死，年輕者之排斥現象會較年長者更晚發生〔年輕者手術後可以撐比較久〕。

4.2 Logistic/Weibull 模式 -- Taylor

Larson and Dinse〔1985〕的分析，對於每一種失敗型態發生的機率(p_j)和給定第 j 個事件會發生的條件存活函數($Q_j(t)$)皆做了母數模式的假設。而在 Taylor〔1995〕一文中，對於每一種失敗型態發生的機率(p_j)仍做 logistic regression 的假設，以 $J=2$ 為例令

$$p_1 = \Pr(\tilde{B}=1|Z) = \frac{\exp(\beta' Z)}{1+\exp(\beta' Z)} \quad ,$$

然而對於 $Q_j(t|Z)$ $(j=1,2)$ 這個部分，對於可能發病者($\tilde{B}=2$)假設 $Q_2(t|Z)=Q_2(t)$ 並以無母數 product-limit 的方式分析之;但是對於免疫的人($\tilde{B}=1$)視$T=\infty$，所以避免了對$Q_1(t|Z)$做分配的假設。在前述假設下，可以得到概似函數

$$L_F = [p_1(z_i)]^{I(b_i=1)}[f_2(x_i|z_i)p_2(z_i)]^{I(b_i=2)}$$

$$\times[p_1(z_i)+Q_2(x_i|z_i)p_2(z_i)]^{I(b_i=0)} \quad ,$$

在這裡 $f_2(t|z)=h(t)Q_2(t|z)$。和 Larson and Dinse (1985) 所提出的估計一樣，最後一項在取對數後再對參數微分的計算會變得太過複雜，所以 L_F 需要先做處理簡化，所得到的概似函數為

$$L_C = [p_1(z_i)]^{I(b_i=1)}[h(x_i)Q_2(x_i|z_i)p_2(z_i)]^{I(b_i=2)}$$

$$\times[p_1(z_i)]^{I(b_i=0,\tilde{B}_i=1)}[Q_2(x_i|z_i)p_2(z_i)]^{I(b_i=0,\tilde{B}_i=2)} \quad ,$$

取對數之後為

$$\log L_C = \sum_{i=1}^{n} \{ I(b_i = 1)[\log p_1(z_i)]$$

$$+ I(b_i = 2)[\log p_2(z_i) + \log Q_2(x_i \mid z_i) + \log h(x_i)]$$

$$+ I(b_i = 0, \tilde{B}_i = 1)[\log p_1(z_i)]$$

$$+ I(b_i = 0, \tilde{B}_i = 2)[\log p_2(x_i) + \log Q_2(x_i \mid z_i)] \} \text{。}$$

Taylor 也是建議以 EM 演算法來做估計，所需要的工作亦是針對簡化後概似函數中的 $I(b_i = 0, \tilde{B}_i = 1)$ 和 $I(b_i = 0, \tilde{B}_i = 2)$ 求條件期望值。對於 $I(b_i = 0, \tilde{B}_i = 2)$ 的條件期望值

$$E[I(\tilde{B}_i = 2) \mid b_i, x_i, z_i] = I(b_i = 2) + I(b_i = 0) \Pr(\tilde{B}_i = 2 \mid X_i = x_i, z_i)$$

$$= I(b_i = 2) + I(b_i = 0) w_2(x_i \mid z_i) \text{，}$$

Taylor 導出

$$w_2(x_i \mid z_i) = \Pr(\tilde{B}_i = 2 \mid X_i = x_i, z_i) = \frac{p_2(z_i) Q_2(x_i \mid z_i)}{p_1(z_i) + p_2(z_i) Q_2(x_i \mid z_i)} \text{，}$$

$$w_1(x_i \mid z_i) = 1 - w_2(x_i \mid z_i) \text{，}$$

此處 $Q_2(x_i \mid z_i)$ 利用 Kaplan-Meier 所表示成的 product-limit 形式代入，透過 EM 演算法即得所求之估計值。

利用以上 logistic / Kaplan-Meier 做為模式假設會發生的問題是過多的未知參數造成估計的繁雜，此外理論上 $\lim_{t \to \infty} Q_1(t \mid Z) = 0$，但是在模擬分析上卻依然可能產生致病時間模式不會遞降到 0 的情形，尤其是當樣本數太小或是設限資料的比例太大的時候。針對這個問題 Taylor 亦提出修補的方法可強迫 $Q_1(t \mid Z)$ 的無母數估計值遞降至 0。

4.3 母數分析 - 非迴歸模式:

Greenhouse and Wolfe(1984)提出以最大概似法做為推論方法，並針對常見模式做細節討論。Ng and McLanchlan (1998)亦是採取母數模式的架構，因為

提出的推論方法較為創新，因此回顧這篇文章。該文主要目的是估計 $T_1 \mid \tilde{B}=1$ 的分配與 p ，至於 $T_2 \mid \tilde{B}=2$ 的分佈較不重要。值得一提的是當 $\tilde{B}=j$ ，令 $T_k=\infty\,(k\neq j)$。在 mixture 的表示法下：

$$\begin{aligned} H(t) &= \Pr(T_1>t, T_2>t) \\ &= \Pr(T_1>t, \tilde{B}=1)+\Pr(T_2>t, \tilde{B}=2) \\ &= Q_1(t)p+Q_2(t)(1-p)\text{。} \end{aligned}$$

論文中討論 likelihood-based 的推論方法。如果 $Q_1(t)$ 和 $Q_2(t)$ 的母數分配均為已知，則 log-likelihood 可以表示為

$$\log L(\psi)=\sum_{j=1}^{n} I(b_j=1)\log\{pf_1(x_j;\theta_1)\} \quad +\sum_{j=1}^{n} I(b_j=2)\log\{(1-p)f_2(x_j;\theta_2)\}$$
$$+\sum_{j=1}^{n} I(b_j=0)\log\{H(x_j;\psi)\}\text{，}$$

其中 $f_j(t)=-\dfrac{\partial Q_j(t)}{\partial t}$ ，$\psi=(\theta_1,\theta_2,p)$。假設我們只關心 (θ_1,p) ，視 θ_2 為不感興趣的參數。以此為前提下作者認為 full-likelihood 需要對兩個分佈均做母數分配的假設並沒有必要，並進一步提出了 partial Maximum likelihood 的估計方法。他們的想法是只對 $Q_1(t)$ 的函數型態做母數的假設，對於 $b_i=1$ 的資料給予密度函數，即 $pf_1(t_j;\theta_1)$ ，但是對於 $b_i=0$ 或是 $b_i=2$ 的觀察值，都視為設限並假設觀察到研究的結束(即 C_i)，此時在概似函數中給予 $(1-p)+pQ_1(C_i;\theta_1)$。所以修正後的部份概似函數可以表示為

$$\begin{aligned} \log\tilde{L}(\psi)=&\sum_{i=1}^{n} I(b_i=1)\log\{pf_1(x_i;\theta_1)\} \\ &+\sum_{i=1}^{n} I(b_i=2)\log\{(1-p)+pQ_1(x_i^*;\theta_1)\} \\ &+\sum_{i=1}^{n} I(b_i=0)\log\{(1-p)+pQ_1(x_i;\theta_1)\}\text{，} \end{aligned}$$

其中最值得注意的是對 $b_i=2$ 的觀測值，不能使用 $T_2=x_i$ 而是更大的 $C=x_i^*$。因為若是將 x_i^* 以 x_i 取代(即競爭風險視為外生的設限情形)，所得之估計量會有偏

誤。

Maller and Zhou（2002）考慮更一般化的競爭風險架構，可以容許更多的失敗型態，所以 $\sum_{j=1}^{J}\Pr(\tilde{B}=j)=\sum_{j=1}^{J}p_j=1$。再定義以下存活函數

$$Q_j(t)=\Pr(T_j>t\mid\tilde{B}=j)\ ,$$

$$H(t)=\Pr(T_1>t,...,T_J>t)=1-\sum_{j=1}^{J}\Pr(T_j\le t,\tilde{B}=j)=1-\sum_{j=1}^{J}\{1-Q_j(t)\}p_j\ 。$$

令 $(T_{ij},C_i)\ (i=1,...,n)$ 為 (T_j,C) 的隨機樣本 $(j=1,...,J)$，代表實際發生事件類型的指標函數定義為 $\delta_{ji}=I(\tilde{B}_i=j,\min(T_{1i},...,T_{Ji})<C_i)$，$\delta_i=\sum_{j=1}^{J}\delta_{ji}$，其中所以若是第 i 個人設限，$\delta_i=0$，否則 $\delta_i=1$，此外令 $X_i=\min(T_{1i},...,T_{Ji},C_i)$，資料可表示為 $\{(\delta_{1i},...,\delta_{Ji},X_i,\delta_i),(i=1,...,n)\}$。概似函數就可表示為

$$L(\psi)=\prod_{i=1}^{n}\{\prod_{j=1}^{J}[p_j f_j(x_i)]^{\delta_{ji}}\}\{1-\sum_{j=1}^{J}p_j[1-Q_j(x_i)]\}^{1-\delta_i}\ ,$$

其中 $-\partial Q_j(t)/\partial t=f_j(t)$。

以上討論的兩篇以母數分析為出發點的文章，如何在“最大概似法”架構下，提出值得發表的創意。Ng and McLanchlan（1998）提出較具穩健性的 partial ML 方法；Maller and Zhou（2002）則是理論推導，做為檢定 $\sum_{j=1}^{J}p_j=1$ 的假說的基礎。在 $J=J^*$ 給定下，而且外生的設限存在，我們往往不知道 $\{i:\delta_i=0\}$ 的設限觀測值中究竟是否包含未出現過的失敗型態(指 $J>J^*$，此時 $\sum_{j=1}^{J}p_j<1$）。這個問題的方向在於希望得知參數的值是否發生在邊界(boundary)，是非典型的推論問題，許多數學技巧都不能使用。

4.4 無母數分析 - Wang

Wang（2004）只考慮 2 種競爭風險，並以 multi-state model 描述問題，文中令 $\tilde{B}$ 做為路徑的指標，並提出以無母數角度估計 $p,Q_1(t),Q_2(t)$ 的方法。為求

符號的一致性，資料型態可以表示為$\{(b_i, x_i)(i = 1,...,n)\}$。可知當$b = 0$，$\tilde{B}$的值未知，在給定資料後$\tilde{B} = 1$的條件機率為

$$p(c) = \Pr(\tilde{B} = 1 \mid b = 0, X = c)$$

$$= \frac{\Pr(T_1 \wedge T_2 > c, \tilde{B} = 1)\Pr(C = c)}{\Pr(T_1 \wedge T_2 > c)\Pr(C = c)}$$

$$= \frac{1}{H(c)} \int_{v>c} \Pr(T_1 \in [v, v + dv), \tilde{B} = 1) \text{ ，}$$

這裡$H(t) = \Pr(T_1 > t, T_2 > t)$。在假設$T_1$、$T_2$和$C$為連續型隨機變數，則可得到

$$P(c) = \frac{1}{H(c)} \int_{v>c} \frac{\Pr(X \in [v, v + dv), b = 1)}{\Pr(C > v)}) + \Pr(\tilde{T} < T_1, \tilde{B} = 1) \text{ ，}$$

其中$\tilde{T} = \sup\{t : \Pr(T_1 > t, T_2 > t)\Pr(C > t) > 0\}$，代表可觀察時間的上界。值得一提的是$\Pr(\tilde{T} < T_1, \tilde{B} = 1)$永遠無法被準確估計，除非是$\Pr(\tilde{T} < T_1, \tilde{B} = 1) = 0$，這個條件就是追蹤時間為充份。可利用 Kaplan-Meier 估計量估計$H(t)$和$G(t) = \Pr(C > t)$如下：

$$\hat{H}(t) = \prod_{u \le t} \{1 - \frac{\sum_{i=1}^{n} I(X_i = u, b_i \neq 0)}{\sum_{i=1}^{n} I(X_i \ge u)}\} \text{ ，}$$

$$\hat{G}(t) = \prod_{u \le t} \{1 - \frac{\sum_{i=1}^{n} I(X_i = u, b_i = 0)}{\sum_{i=1}^{n} I(X_i \ge u)}\} \text{ 。}$$

對於$\Pr(X = v, b = 1)$，因為是可觀察變數之機率函數則可用 empirical function 來做估計。而$\Pr(\tilde{T} < T_1, \tilde{B} = 1)$可寫成$p(\tilde{T})H(\tilde{T})$，對於$p(\tilde{T})$和$H(\tilde{T})$可以利用觀察到最大$b \neq 0$的值$x_{(n)}$來做估計。因此可以得到$p(c)$的估計量為

$$\hat{p}(c) = \frac{1}{\hat{H}(c)} \int_{v>c} \sum_{i=1}^{n} \frac{I(X_i = v, b_i = 1)}{n\hat{G}(v)} + \tilde{p}(\tilde{X}_{(n)})\hat{H}(\tilde{X}_{(n)}) \text{ ，}$$

其中$\tilde{p}(\tilde{X}_{(n)})$的值需要額外假設才能得知。

$p(c)$的資訊對於估計$Q_1(t)$、$Q_2(t)$與p很有幫助。$Q_1(t)$和$Q_2(t)$與p可利用下列三個式子來做估計，

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n}\{I(b_i = 1) + I(X_i = x_i, b_i = 0)\hat{p}(x_i)\}\ ,$$

$$\hat{Q}_1(t) = \prod_{u \le t}\left\{1 - \frac{\sum_{i=1}^{n} I(X_i = u, b_i = 1)}{\sum_{i=1}^{n} I(X_i \ge u, b_i = 1) + \sum_{i=1}^{n} I(X_i \ge u, b_i = 0)\hat{p}(x_i)}\right\}\ ,$$

$$\hat{Q}_2(t) = \prod_{u \le t}\left\{1 - \frac{\sum_{i=1}^{n} I(X_i = u, b_i = 2)}{\sum_{i=1}^{n} I(X_i \ge u, b_i = 2) + \sum_{i=1}^{n} I(X_i \ge u, b_i = 0)\{1 - \hat{p}(x_i)\}}\right\}\ 。$$

利用無母數方法來做分析，因為不需要做模式分佈假設，所以估計量應該具有穩健性，但是先決條件仍是資料的品質要夠好，在此是以“充份追蹤時間”的條件是否成立來衡量。雖然如此，第三類型的資料因為多了可以辨別部份免疫者的資訊，即使是假設不符仍有可能做合理的猜測和修正。

第五章 結論

在有關治癒模式的文獻中，我們將文章依“治癒”或是“免疫”是否明確定義區分成“傳統”、“以固定時間為切點”和“考慮競爭風險”三個方向，並以混合模式的架構，分別討論“致病與否”和“發病時間”的推論問題。一般統計的分析方法可區分為討論解釋變數影響的迴歸分析與非迴歸分析。此外依假設強弱可分為母數分析、半母數分析、以及無母數分析三種方法。本論文所選取的文獻以迴歸模式的半母數分析為主，在混合模式的架構下可分別探討解釋變數對於“致病與否”和“發病時間”的影響。

在傳統的治癒模式下，因為免疫者的定義不明確，所以推論方法往往必須依賴模式假設以避免在參數估計時有不可辨識的問題。母數分析方法在模式假設正確的情形下，估計量具有效性;然而若是模式假設錯誤，所得之估計量可能不夠穩健。我們發現文獻上對於治癒模式討論模式適合度〔goodness-of-fit〕的方法似乎不多。當以混合模式架構將模式分成兩個部份討論時，如何針對兩部份的模式假設檢驗其適合度，以及探討兩者如何互相影響，應該是值得繼續討論的課題。無母數分析所得之估計量，雖然假設最弱但是需要“充份追蹤時間”成立下才能得到合理的性質。我們發現在無母數分析和母數分析做取捨時，除了穩健性和有效性的考量外，尚要評估資料的品質是否滿足“充分追蹤時間”的標準。因此檢驗這個條件是否成立是一個重要的研究課題，Maller and Zhou 〔1994〕提出的檢驗方法，依據“最大事件發生值”和“最大觀測值”的距離做為判斷觀測時間是否充份的依據，然而這方法的實用性並不大。我們認為可以嚐試的一個思考方向是在“充份追蹤時間”不成立時，推導估計量造成的偏誤的範圍〔bound〕，或是發展一種準則使能夠達到最短的充份的追蹤時間。

在討論迴歸分析的文獻時，我們發現免疫者的存在會使得概似函數因為多了連加項而在分析上變得複雜。當模式假設弱時，如何處理維度龐大不感興趣的參數是推論過程挑戰的來源。本論文的討論集中在以 EM 演算法簡化概似函數的估

計，所研讀的數篇論文差異在於估計權數時處理 nuisance parameters 的方法有所不同。EM 演算法的優點是提供具體的方式以求得複雜函數的解，但是起始值的選取、收斂的判斷、收斂值的理論性質、所得到估計量的變異數是否能求得 … 都是伴隨來的新問題，換言之以概似函數為出發點的方向並非萬靈丹。在論文即將完成時，我們也發現有數篇類似的文獻，模式由 Cox PH 模式拓展到更具一般性的轉換模式〔transformation model〕，推論方法以估計函數的概念為主，可以參考 Chen et al.〔2002〕和 Lu and Ying〔2004〕。

對於考慮競爭風險下的治癒模式來說，免疫者的定義取決在於競爭風險型態發生的順序，相對於傳統的治癒模式，免疫者有明確的定義。所以在考慮競爭風險下，無論在迴歸分析或是在無母數分析上，所需要的假設都較傳統治癒模式要來的弱些。Wang (2004)的無母數方法亦需要充分追蹤時間假設，但所需條件較 Maller and Zhou (1992) 的架構為弱。我們將討論過治癒模式的數篇論文依其類別整理於附錄二中。

附錄一、參考文獻

1. Andersen, P. K., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.
2. Betensky, R. A. and Schoenfeld, D. A. (2001). “Nonparametric estimation in a cure model with random cure times.” *Biometrics*, **57**, 282-286.
3. Chen, K., Jin, Z. and Ying, Z. (2002). “Semiparametric regression analysis of transformation models with censored data.” *Biometrika*, **89**, 659-668.
4. Farewell, V. T. (1982). “The use of mixture models for the analysis of survival data with long-term survivors.” *Biometrics*, **38**, 1041--1046.
5. Farewell, V. T. (1986). “Mixture models in survival analysis.” *Can. J. Statist.*, **4**, 257-262.
6. Greenhouse, J. B. and Wolfe, R. A. (1984). “A competing risks derivation of a mixture model for the analysis of survival data.” *Commun. Statist. – Thero. Meth.*, **13**, 3133-3154.
7. Hougaard, P. (2000). *Analysis of Multivariate Survival Analysis.* New York: Springer-Verlag.
8. Jung, S.-H. (1996). “Regression analysis for long-term survival rate.” *Biometrika*, **83**, 227-232.
9. Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer-Verlag.
10. Kuk, A. Y. C. and Chen, C. (1992). “A mixture model combining logistic regression with proportional hazards regressions.” *Biometrika*, **79**, 531--541.
11. Larson, M. G. and Dinse, G. E. (1985). “A mixture model for the regression analysis of competing risks data.” *Appl. Statist.*,**34**,No.3,201-211.
12. Laska, E. M. and Meisner, M. J. (1992). “Nonparametric estimation and testing in a cure model.” *Biometrics*, **48**, 1223--1234.

13. Li, C.-S., Taylor, J. M. G. and Sy, J. P. (2001). “Identifiability of cure models.” *Statistics & Probability Letters*, **54**, 389--395.

14. Lu, W. and Ying, Z. (2004). “On semiparametric transformation cure model.” To appear in *Biometrika*.

15. Maller, R. A. and Zhou, S. (1992). “Estimating the proportion of immunes in a censored sample.” *Biometrika*,**79**,731--739.

16. Maller, R. A. and Zhou, S. (1994). “Testing for sufficient follow-up and outliers in survival data.” *J. Am. Statistic. Assoc.*, **89**, 1499--1506.

17. Maller, R. A. and Zhou, S. (1996). *Survival Analysis with Long-Term Survivors.* Wiley: New York.

18. Ng, S.K. and McLachlan, G.J. (1998). On modifications to the long-term survival mixture model in the presence of competing risks. *Journal of Statistical Computation and Simulation*, **61**, 77-96.

19. Peng, Y. and Dear, K. B. G. (2000). “A nonparametric mixture model for cure rate estimation.” *Biometrics*, **56**, 237-243.

20. Pepe, M. S. (1991). “Multiple failure endpoint studies.” *J. Am. Statistic. Assoc.*, **86**, 770-778.

21. Prentice, R.L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). “The analysis of failure times in the presence of competing risks.” *Biometrics*, **34**, 541--554.

22. Simon R. and Makuch, R. W. (1984). “A nonparametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias.” *Statistics in Medicine*,**3**, 35-44.

23. Slud, E. V. and Rubinstein, L. V. (1983). “Dependent competing risks and summary survival curves.” *Biometrika*, **70**, 643-649.

24. Subramanian, S. (2001). “Parameter estimation in regression for long-term

survival rate from censored data." *Journal of Statistical Planning and Inference*, **99**, 211-222.

25. Sy, J. P. and Taylor, J. M. G. (2000). "Estimation in a Cox proportional hazards cure model." *Biometrics*, **56**, 227-236.
26. Taylor, J. M. G. (1995). "Semi-parametric estimation in failure time mixture models." *Biometrics*,51, 899--907.
27. Tsai, W. Y. and Crowley, J. (1998). "A note on nonparametric estimators of the bivariate survival function under univariate censoring." *Biometrika*, **85**, 573--580.
28. Turnbull, B. W., Brown, B. W. and Hu, M. (1974) "Survivorship Analysis of Heart Transplant Data." *J. Amer. Statist. Association*, **69**,74-80.
29. Voekel, J. G. and Crowley, J. J. (1984). "Nonparametric inference for a class of semi-Markov processes with censored observations." *Ann. Statist.*,**12**, 142-160.
30. van der Laan, M. J. and Hubbard, A. E. (1998). "Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed." *Biometrika*, **85**, 771-783.
31. Wang, W. and Wells, M. T. (1998). "Nonparametric estimation of successive duration times under dependent censoring." *Biometrika*, **85**, 561-572.
32. Wang, W. (2004). "Nonparametric estimation of the sojourn time distributions for a multi-path Model". *Journal of the Royal Statistical Society, Series B.* **65**, 921-936.
33. Zhao, L. P. and LeMarchand, L. (1992). "An analytical method for assessing patterns of familiar aggregation in case-control studies." *Genetic Epidemiology*,9, 141-154.

附錄二、治癒模式文獻一覽

	作者與發表年代	可能發病與否	可能發病時之發病時間
傳統治癒模式	Farewell (1982,1986)	Logistic regression	Weibull
	Kuk and Chen (1992)	Logistic regression	假設 Cox model marginal likelihood
	Sy and Taylor Peng and Dear(2000)	Logistic regression	假設 Cox model EM (baseline 處理不同)
	Lu and Ying (2004)	Logistic regression	Transformation model
	Maller and Zhou (1992)	未做假設	未做假設但需要充份觀察時間
固定切點	Jung (1996), Subramanian (2001)	Logistic regression	定值
競爭風險模式	Ng & McLanchlan (1998) Maller & Zhou (2002)	未做假設	(部份)母數分配
	Larson and Dinse (1985)	Logistic regression	Cox model (基底函數在區段為常數)
	Taylor (1995)	Logistic regression	Nonparametric
	Wang (2004)	未做假設	未做假設但需要充份觀察時間