中华流行病学杂志1994年2月第15卷第1期

分组Cox模型及其在癌症预后因素研究中的应用

项永兵　高玉堂

摘要　Cox比例危险回归模型是医学随访研究、临床试验研究中分析生存资料最常用的多因素分析方法，但它不适合于处理分组生存资料或重叠严重的大样本生存数据。笔者对分组比例危险回归模型及其在大样本寿命表生存资料分析中的应用进行了讨论。最后结合实例借助于GLIM软件探讨它在肺癌随访资料预后因素分析中的应用。

关键词　比例危险回归模型　分组Cox模型　广义线性模型　肺癌

Cox比例危险回归模型为多因素生存分析开辟了广阔的前景，是生存分析中最常用的多元分析方法之一[1, 2]。但Cox模型只能局限于样本含量并不很大且生存时间互不重叠或重叠很少的数据分析，不适合于分析生存时间重叠较多、样本含量很大或生存时间分组的数据，所以分组比例危险回归模型[2~8]应运而生。不同的作者所提出的模型形式不同，本文主要介绍有代表性的模型，并借助于GLIM软件[10]将其应用于肺癌预后因素的分析，取得了满意的效果。

分组比例危险回归模型及其建立

比例危险回归模型的基本形式是用下式定义研究个体在时刻t的危险率与回归向量间的关系。

$$\lambda(t, x)=\lambda_0(t)\exp(x\beta), \qquad (1)$$

$\lambda(t, x)$是具有协变量行向量x的个体在时刻t的危险率，$\lambda_0(t)$代表个体在时刻t的基准危险率，$x=(x_1,\cdots,x_p)$是解释变量行向量，β是与x相应的待估参数的列向量。而分组比例危险回归模型是首先把随访时间分成不同的时间区间，然后根据观察数据所建立的回归模型。假定有n个观察个体，随访区间分成k+1个区间$Ii=(\alpha_{i-1}, \alpha_i)$，$(i=1, \cdots, k+1)$，其中$\alpha_0=0$、$\alpha_k=T$、$\alpha_{k+1}=\infty$。用$R_i$、$D_i$分别表示在时刻$\alpha_{i-1}$的危险集、区间$(\alpha_{i-1}, \alpha_i)$内观察到的死亡个体集。先定义$P_i(x)$为具有协变量x的个体活过时刻$\alpha_i$的概率，$p_i(x)$为具有协变量x的个体活过时刻$\alpha_{i-1}$再活至$\alpha_i$的概率。在死亡与截尾于同一区间的发生是互为独立事件的前提下，可以构造如下似然函数

$$\prod_{i=1}^{k}\left(\prod_{l\in Di}[1-Pi(x_l)]\prod_{l\in Ri-Di}Pi(x_l)\right) \qquad (2)$$

在(1)式的研究基础上，很容易获得下式

$$Pi(x)=S(\alpha_i|x)=S_0(\alpha_i)^{\exp(x\beta)}=P_i^{\exp(x\beta)} \qquad (3)$$

式中$S(t|x)=S_0(t)^{\exp(x\beta)}$为生存函数，且$P_i=P_i(0)$，$P_i=\frac{P_i}{P_{i-1}}$，$i=1,\cdots,k$，则

$$pi(x)=\frac{P_i(x)}{P_{i-1}(x)}=p_i^{\exp(x\beta)} \qquad (4)$$

此式即所谓分组Cox模型。参数$\beta=(\beta_1, \cdots, \beta_p)$、$p=(p_1, \cdots, p_k)$的估计，Lawless[2]、Prentice和Gloeckler[3]有过详细的讨论。根据(2)式很容易建立样本的对数似然函数[2]，且Prentice和Gloeckler[3]对参数(β, p)进行了“重新参数化”，以改善参数极大似然估计中的Newton-Raphson算法的收敛性。

本文作者单位：上海市肿瘤研究所流行病学研究室　200032

参数β的假设检验，有记分统计量检验[3]。Lawless[2]、Prentice和Gloeckler[3]都对截尾数据存在时的似然函数的建立和参数的统计推断进行了讨论。

Aranda-Ordaz[7]、Tibshiriani和Ciampi[8]讨论了转化为广义线性模型理论[9]下的分组Cox模型拟合的问题。假定用$\theta_i(x)$表示具有协变量x的个体在区间I_i的区间危险度或区间条件死亡概率：

$$\theta_i(x)=Pr(\alpha_{i-1}\leqslant T<\alpha_i \mid T\geqslant\alpha_{i-1},x) \quad (5)$$

那么$p_i(x)=1-\theta_i(x)$是具有协变量x的个体活至α_{i-1}时刻后又活至α_i时刻的条件生存概率。根据（1）式可导出下列函数：

$$\ln[-\ln\{p_i(x)\}]=\ln[-\ln\{p(0)\}+x\beta \quad (6)$$

$$\ln[-\ln\{1-\theta_i(x)\}]=\ln[-\ln\{1-\theta i(0)\}]+x\beta \quad (7)$$

其中$\theta_i(0)$为x=0时的个体的基准死亡概率。令$c_i=\ln[-\ln\{1-\theta_i(0)\}]$，则（7）式右侧为一线性结构$c_i+x\beta$。再利用GLIM系统即不难拟合此类回归模型。Aranda-Ordaz[7]、Tibshiriani和Ciampi[8]根据偏似然函数的理论定义了下述广义线性模型。

$$\prod_{i=1}^{k}\prod_{j\in R(\alpha_{i-1})}\theta_i(x_i)^{\delta}\{1-\theta_i(x_i)\}^{1-\delta} \quad (8)$$

其中$R(\alpha_{i-1})$指时刻α_{i-1}的危险集，δ为指示变量，δ=1表示个体在区间I_i内死亡，δ=0表示截尾。在此基础上，假定r组病人中n_{ij}为j组i区间初的危险暴露个体数，d_{ij}为j组在i区间内的死亡数，则偏似然函数为：

$$L=\prod_{i=1}^{k}\prod_{j=1}^{r}\binom{n_{ij}}{d_{ij}}\theta_{ij}^{d_{ij}}(1-\theta_{ij})^{n_{ij}-d_{ij}} \quad (9)$$

对数偏似然函数为：

$$LnL=\sum_{i=1}^{k}\sum_{j=1}^{r}\{d_{ij}\ln(\theta_{ij})+(n_{ij}-d_{ij})\ln(1-\theta_{ij})\} \quad (10)$$

式中$\theta_{ij}=1-\exp\{-(1+c_i+x\beta)\}$为j组中个体在$I_i$区间的死亡概率[7,8]或$\theta_{ij}=\hat{\mu}_{ij}/n_{ij}$，$\hat{\mu}_{ij}$为$d_{ij}$的期望值[7]。略去常数项$\binom{n_{ij}}{d_{ij}}$后，利用上述似然函数即不难对参数进行统计推断，如果存在截尾数据，它们对似然函数的贡献应该考虑进去，Lawless[2]、Aranda-Ordaz[7]有过讨论。

实例分析

样本数据是1984～1986年诊断的上海市区35～69岁肺癌新病例，随访至1989年年底。为了说明上述方法的应用，我们取病例的性别、病期和治疗手段作为分类变量，其中性别二个水平、病期三个水平、治疗手段二个水平。在建立寿命表格式数据之前，首先确定寿命表的随访区间。Tibshiriani等[8]曾建议取5～10个区间为宜。本研究的样本资料随访了五年，随访区间以年为单位即有五个时间区间。然后借助于GLIM软件，以寿命表每格子中观察死亡数作为模型的因变量Y（用yvariate命令指定），误差结构取二项分布（用error命令指定），连接函数为互补ln(－ln)，连接（用link命令指定），因素水平数用factor命令指定。

表1给出了肺癌生存数据拟合分组Cox模型的步骤。模型3、4、5是为考虑单个因素的作用而拟合的模型。模型6、7、8是在模型2的基础上引入各个因素，与模型2比较，模型偏差度变化：性别$\triangle G^2=4.229$（$P<0.05$）、病期$\triangle G^2=331.805$（$P<0.001$）、治疗手段$\triangle G^2=216.514$（$P<0.001$），以病期对肺癌预后的统计学意义最强。模型9、10、11为三个因素中两两因素相互调整的模型。模型12则是引入全部三个因素，而模型13至16是在模型12的基础上，引进因素间交互作用项的模型，但从偏差度及自由度的变化上看，它们对模型均没有贡献。模型11从统计学上讲是拟合较佳的模型。考虑到性别因素的作用，模型12亦可看作拟合较佳的模型。

表1　肺癌预后因素分组Cox模型拟合步骤及模型拟合优度检验*

模　　型	偏差度(G^2)	自由度(df)	P值
1.总均数模型	803.08	59	$P<0.001$
2.随访时间	554.151	55	$P<0.001$
3.性别	798.975	58	$P<0.001$
4.病期	450.217	57	$P<0.001$
5.治疗手段	572.400	58	$P<0.001$
6.随访时间＋性别	549.922	54	$P<0.001$
7.随访时间＋病期	222.346	53	$P<0.001$
8.随访时间＋治疗手段	337.637	54	$P<0.001$
9.随访时间＋性别＋病期	220.890	52	$P<0.001$
10.随访时间＋性别＋治疗手段	337.611	53	$P<0.001$
11.随访时间＋病期＋治疗手段	69.699	52	$P>0.05$
12.随访时间＋性别＋病期＋治疗手段	68.885	51	$P>0.05$
13.模型12＋性别·病期	68.850	49	$P>0.05$
14.模型12＋性别·治疗手段	67.260	50	$P>0.05$
15.模型12＋病期·治疗手段	64.915	49	$P>0.05$
16.模型12＋性别·病期·治疗手段	59.446	44	$P>0.05$

*用*GLIM*软件拟合

表2　肺癌生存数据拟合分组Cox模型的参数估计值

因素及其水平	模型11				模型12			
	β	S.E.	exp(β)	95%CI	β	S.E.	exp(β)	95%CI
病期：								
早期			1.00				1.00	
中期	0.7361	0.1109	2.09	(1.68-2.59)**	0.7317	0.1110	2.08	(1.67-2.58)**
晚期	1.3339	0.0929	3.80	(3.16-4.55)**	1.3359	0.0929	3.80	(3.17-4.56)**
治疗手段：								
手术或联合			1.00				1.00	
化疗或放疗	0.7706	0.0631	2.16	(1.91-2.45)**	0.7804	0.6390	2.18	(1.93-2.40)**
性别：								
男							1.00	
女					−0.0556	0.0617	0.95	(0.84-1.07)*

注：①病期以早期为参照水平，治疗手段以手术或联合治疗为参照水平，性别以男性为参照水平。

②CI：可信区间。*$P<0.05$，**$P<0.01$

表2列出了基于模型11与12的参数估计值，给出了测度预后因素作用的常用指标。结果表明病期、治疗手段是影响肺癌预后的重要因素。中期与早期相比，危险比为2.1；晚期与早期相比，危险比为3.8。单一化疗或放疗等的治疗效果比手术或联合治疗的效果差，危险比为2.0以上。性别因素在模型中引进了病期和治疗手段以后没有显示出有统计学上的意义。

讨　论

Holford[4]、Thompson[5]、Prentice等[3]和Pierce等[6]提出了形式各异的处理生

存时间分组的统计模型。其中尤以Prentice和Gloeckler[3]提出的模型受到研究者们的重视，它保留了Cox模型“比例危险”的假设，故称为分组比例危险回归模型或称分组Cox模型。随后，Aranda-Ordaz[7]、Tibshiriani和Ciampi[8]讨论了整理成寿命表格式的生存数据分析。上述模型弥补了Cox回归模型难于胜任生存时间重叠严重或生存时间分组资料分析的缺陷，可用于分析大样本生存时间重叠严重的数据。转换为广大线性模型结构下的分组Cox模型，借助于GLIM软件更宜实现分组生存数据的分析，为大样本肿瘤预后因素的研究提供了很方便的手段。

通过拟合一系列模型，根据模型偏差度及自由度选择能够代表数据的最佳模型，并在此基础上估计模型中各参数的估计值。偏差度（deviance）是衡量由模型获得的拟合值与实际观察值之间差别大小的统计量。首先用于模型拟合好坏的拟合优度检验；其次用于衡量引进或剔除出模型的因素对模型是否有显著作用。

GLIM系统为模型拟合提供了非常理想的计算机环境，尤其适合于拟合因素间交互作用项的模型。但鉴于GLIM软件在整理数据资料功能上的欠缺，所以分析前要先对原始数据按寿命表格式进行分类整理，这是本研究的一个关键步骤，有条件的可利用一些计算机软件进行整理，以提高准确性或工作效率。

笔者利用GLIM软件拟合肺癌预后因素的分组Cox模型，并给出拟合步骤及各预后因素的β估计值及其标准误、危险比及其95%可信区间等。Aranda-Ordaz[7]、Tibshiriani等[8]所讨论的模型是一簇模型，包括相加和相乘模型。本文所拟合的分组Cox模型只是其中的一种，即连接函数为互补ln（－ln）的相乘模型情形。当然也可以直接采用Cox回归模型等多变量分析方法，将得到类似的结果。

应当看到，用分组Cox回归模型分析生存数据也存在一些不足。当引入模型的因素较多或因素水平较多或生存时间区间较多，则寿命表分层将越来越多，势必有些层次数据空缺，使模型拟合发生困难。对于研究因素较少的肿瘤临床试验和大样本随访生存资料的分析不失为一种十分有效的方法。总之，这一模型的实用性研究等有关问题值得深入探讨。

Grouped Cox Regression Model and its Application in Study of Prognostic Factors on Cancer *Xiang Yongbing, Gao Yutang, Department of Epidemiology, Shanghai Cancer Institute, Shanghai* 200032

Cox proportional hazards regression model is the most popular multivariate regression model for analysis of survival data in medical follow-up studies and clinical trials, but it is unable to handle grouped survival data or large data sets with many tied failure times adequately. This paper explores the grouped proportional hazards regression model (GPH model) and its use in analysis of large data sets presented in life tables. By use of the data in a lung cancer follow-up study conducted in urban area of Shanghai, the authors give an example in detail for analysing prognostic factors of lung cancer by using GLIM.

Key wodrs Proportional hazards regression model Grouped Cox model GLM Lung cancer

参考文献

1 Cox DR, Oakes D. Analysis of survival data. London: Chapman and Hall, 1984: 201.

2 Lawless JF. Statistical models and methods for lifetime data. New York: John Wiley and Sons, 1983：580.

3 Prentice RL, Gloeckler LA. Regression analysis of grouped survival da'a with application to breast cancer data. Biometrics, 1978, 34：57.

4 Holford TR. Life tables with concomitant information. Biometrics, 1976, 32：587.

5 Thompson WA. On the treatment of grouped

observations in life studies. Biometrics, 1977, 33：463.

6 Pierce DA. Distribution-free regression analysis of grouped survival data. Biometrics, 1979, 35：785.

7 Aranda-Ordaz FJ. An extension of the proportional hazards model for grouped data. Biometrics, 1983, 39：109.

8 Tibshiriani RJ, Ciampi A. A family of proportional-and additive-hazards models for survival data. Biometrics, 1983, 39：141.

9 Nelder JA, Wedderburn RWM. Generalized linear models. Journal of Royal Statistics Society, Series A 1972, 135：370.

10 GLIM Working Party. The GLIM system, Release 3.77, Payne, CD eds. Oxford：Numerical Algorithms Groups, 1985.

（收稿：1993-05-12　修回：1993-08-24）

传染病院医务人员乙型肝炎疫苗接种效果观察

李兴旺　付淑琴　赵翠琴　秦　靖　石景罡　丁静秋

为了解乙肝对医务人员的危害，我们对279名医务人员进行了HBV M检测，感染率为45%，又对HBV M阴性的154人进行了血源性乙肝疫苗接种，并随访一年，结果报告如下。

一、资料与方法：154人均为从事乙肝诊治、护理及检验工作的医务人员，其中男性14人，女性140人，年龄18～56岁，平均28.7岁，按0、1、6月程序接种，首剂30μg，二、三次为10μg，第7个月检测HBV M及免疫学指标。接种后一年对抗-HBs阳转者再次检测抗-HBs，以了解抗体维持水平。而对无效者给予再次接种。

二、结果：

1.概况：接种后106人为正常应答（抗-HBs/S/N≥10），7人为低应答（S/N：2.2～9.9），41人无应答（S/N：≤2.1），抗-HBs阳转率为73.4%（113/154），无应答的41人中34人接受了第2个程序，接种结果26人正常应答，1人低应答，7人无应答，抗-HBs阳转率79.4%，仍无应答者又有5人接受了第3个程序接种，结果抗-HBs全部阳转。

2.不同年龄与抗-HBs阳转率的关系：按人30岁、31～岁、41～50岁、>50岁分组，阳转率分别为91.3%、51.2%、64.7%和50%（$P<0.01$）。41～50岁组中有新调入人员，此5人阳转率为80%，其余阳转率为58.3%。

3.从事乙肝工作时间与抗-HBs阳转率的关系：按<1年、1～年、6～年、11～20年、>20年分组，阳转率分别为97%、89%、52.1%、43.6%和50%（$P<0.01$）。

4.阳转率与免疫学指标的关系：共对15例进行了观察，淋巴细胞亚群测定、干扰素水平测定与阳转率无明显关系，而淋转试验抗-HBs阳转者的转换率明显高于未阳转者。

5.随访结果：113例抗-HBs阳转者再次检测抗-HBs结果34人（30.1%）抗-HBs消失，比较不同年龄，不同工龄均有显著差异（$P<0.05$，$P<0.01$），对消失者给予20μg加强一次，结果抗-HBs全部阳转。

三、讨论：应用乙肝疫苗对普通人群进行预防接种，抗-HBs阳转率可达95%以上，而笔者对常年从事乙肝工作的医务人员接种效果显示，阳转率仅为73.4%，明显低于普通人群。据文献报道，45岁以上者抗-HBs阳转率逐渐下降，约为70%～85%，提示阳转率与年龄有关，而本文资料显示，接触乙肝病人的医务人员31岁以上者，阳转率明显下降（51.2%）年龄组明显提前，而此部分人员工作年限多在10年以上，提示，除与年龄有关外，还与接触乙肝时间的长短密切相关。

乙肝疫苗免疫后，抗-HBs可在体内维持数年，5年后抗-HBs消失率为8%～20%。而本组资料显示，1年后消失率即为30.1%，这一比率与年龄，尤其与工龄显著相关，似提示随着接触乙肝时间的延长，不仅抗-HBs阳转率降低，且产生抗体后消失的亦早。我们的结果还提示，对初免无应答者再重复给予第二，甚至第三程序接种，可明显提高抗-HBs阳转率。

综上所述，我们认为，对将要从事乙肝工作的医务人员应首先进行乙肝疫苗接种，而对已经从事这一工作的医务人员，接种后无效时，应增加接种次数或加大剂量，直至产生抗-HBs以达到防护的目的。

（收稿：1993—05—26　修回：1993—08—29）

本文作者单位：北京地坛医院　100011