·循证医学中的医学统计学问题·

COX回归模型在临床医学科研中的价值

李 河，郭 兰，孙家珍

（广东省人民医院、广东省心血管病研究所、广东省医学科学院，广州 510080）

[摘 要] 在临床医学科研中常会遇到生存数据的统计学处理分析问题，而比例危险率回归模型（简称COX回归模型）常用来进行影响生存时间的多因素分析；故在论文撰写中需根据科研设计类型和研究变量特征，进行合理的临床流行病学和医学统计学思维，以便得到正确结果。

[关键词] 生存数据；完全数据；截尾数据；比例危险率回归模型（COX回归模型）

[中图分类号] R195.1 [文献标识码] A [文章编号] 1671-5144（2011）01-0051-03

The Values of COX Regression Model in the Clinical Medicine Research

LI He, GUO Lan, SUN Jia-zhen

(*Guangdong General Hospital, Guangdong Cardiovascular Institute, Guangdong Academy of Medical Sciences, Guangzhou 510080, China*)

Abstract: It's a common situation for a survival data statistics in the clinical researches. Proportional hazards regression model is often applied for the multivariate analysis on survival time. So, it's necessary to have a correct thinking in clinical epidemiology and medical statistics for the good paper-writing, which is based on the research-type and variable-characteristic, as to achieve a correct result.

Key words: survival data; completed data; censored data; proportional hazards regression model (COX regression model)

在临床医学科研中常常会遇到如下生存研究数据，如心脏瓣膜置换术后瓣膜存活时间研究、肿瘤病人接受治疗后其生存时间研究等。生存数据（寿命数据）可分为完全数据和截尾数据。完全数据如当病人确诊为某病（或开始治疗）到病人死亡（或到治愈）为止所经历的全部时间。截尾数据（或删失数据）中，左截尾数据为记录之前的某个时刻发生了终点事件（如死亡）但确切发生（死亡）时间未知；右截尾数据为记录之后仍存活但确切发生终点事件（死亡）的时间未知；区间截尾数据为某个时间区间发生了终点事件（如死亡）但确切发生（死亡）时间未知。生存分析是需要同时考虑生存结果和生存时间的一种统计学方法，可以充分利用截尾数据的信息，对生存时间分布特征及生存时间的主要影响因素进行分析。医学统计学中处理分析生存数据的统计学方法可简单划分为：参数分析法（如指数分布、威布尔分布）、非参数分析法（寿命表法、KAPLAN-MEIER乘积极限法）、半参数分析法（COX回归模型）。本文以下列模拟数据为例来说明COX回归模型在临床医学科研中的应用价值。

以接受心脏瓣膜置换术后病人随访数据库（aa1.sas7bdat）为基础，截取其中128例研究对象的相关变量数据为例，并且对有关数据进行变量变换，来举例分析说明瓣膜置换术后瓣膜的生存

[作者简介] 李河（1963-），男，内蒙古商都人，主任医师，医学博士，主要研究方向为流行病学与医学统计学在临床医学科研中应用、心血管病的二级预防研究。

[通讯作者] 孙家珍，Tel：020-83827812；E-mail：sjz36@163.com

状况。用于本例的数据中，研究对象男 60 例、女 68 例，年龄 7~51 岁，平均年龄及标准差 29.33±9.32 岁，其中二尖瓣置换术 121 例，非二尖瓣置换术 7 例，对研究对象的平均随访时间为 92.3 个月（5~219 个月）。本例数据相关研究变量定义如下：编号 ID 为研究对象统一编号，性别 SEX（男性=1，女性=2），年龄 AGE3（从小到大每 10 岁为一年龄等级），手术类型 SSLX2（二尖瓣置换术=1，非二尖瓣置换术=0），瓣膜生存时间 SCSJ（实际生存时间），终点事件 SCSJ1（终点事件 SCSJ1=1 为发生瓣膜生存时间小于 5 年事件、否则 SCSJ1=0），OK1（完全数据 OK1=1，截尾数据 OK1=0），见表 1。

表 1 心脏瓣膜置换术后瓣膜生存状况数据

ID	SEX	AGE3	SSLX2	SCSJ	SCSJ1	OK1
2	1	3	1	125	0	1
3	2	5	1	140	0	1
4	1	2	1	47	1	1
6	2	4	1	132	0	1
10	1	2	1	138	0	1
11	1	4	1	162	0	1
…	…	…	…	…	…	…
…	…	…	…	…	…	…
194	2	3	1	117	0	0
195	2	4	1	145	0	1
199	2	3	0	87	0	0

为了分析性别、年龄及手术类型对瓣膜生存状况的影响，甲医生对上述生存数据进行了如下统计学处理分析：首先依据上数据库定义相关协变量为性别 SEX（男性=1，女性=2），年龄 AGE3（从小到大每 10 岁为一年龄等级），手术类型 SSLX2（二尖瓣置换术=1，非二尖瓣置换术=0）；定义因变量为“瓣膜生存时间小于 5 年即为发生终点事件，并且赋值 SCSJ1=1，反之赋值 SCSJ1=0”，然后运行 SAS 程序执行二分类因变量的多变量 Logistic 回归分析、二分类因变量的多变量逐步 Logistic 回归分析（取协变量引入方程的检验水准 α= 0.05，剔出方程的检验水准 α=0.10），表 2 结果显示年龄每增加一个等级（增加 10 岁），发生瓣膜生存时间小于 5 年（发生终点事件）的比数比（odds ratios，OR）=0.456（95%可信区间 0.287~ 0.725），差异有统计学意义（χ^2=10.989 3，P=0.000 9）；而性别及手术类型对瓣膜生存时间的影响无统计学意义（χ^2=0.896 6，P=0.343 7；χ^2=0.669 2，P=0.413 3）。表 3 结果显示年龄每增加一个等级（增加 10 岁），发生瓣膜生存时间小于 5 年（发生终点事件）的比数比 OR=0.482（95%可信区间 0.306~0.759），差异有统计学意义（χ^2=9.924 2，P=0.001 6）。SAS 参考程序及结果如下[1-3]：

```
data aa.aa2; set aa.aa1;
if id='.' then delete;
if scsj='.' then delete;
if scsj<=60 then scsj1=1;
else if scsj>60 then scsj1=0;
if id >200 then delete; run;
ods rtf;
proc logistic data=aa.aa2 descending;
model scsj1=sex age3 sslx2;
run; ods rtf close;
ods rtf;
proc logistic data=aa.aa2 descending;
model scsj1=sex age3 sslx2/selection=stepwise sle=
0.05 sls=0.10; run;
ods rtf close;
```

而乙医生对上述数据进行了如下处理：为了分析性别、年龄及手术类型对瓣膜生存状况的影响，定义相关协变量为性别 SEX（男性=1，女性=2），年龄 AGE3（从小到大每 10 岁为一年龄等级），手术类型 SSLX2（二尖瓣置换术=1，非二尖瓣置换术=0）；因变量取瓣膜生存时间 SCSJ（生存时间），并且定义其数据特性变量 OK1（完全数据赋值 OK1=1，截尾数据赋值 OK1=0）。然后进行多变量 COX 回归分析、多变量逐步 COX 回归分析（协变量引入方程的检验水准 α=0.05，剔出方程的检验水准 α=0.10），表 4 结果显示性别、年龄及手术类型对瓣膜生存时间的影响皆未见差异有统计学意

表 2　瓣膜生存时间 5 年影响生物瓣膜生存的危险因素 Logistic 回归分析

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Odds Ratio Estimates		
						Point Estimate	95% Wald	Confidence Limits
Intercept	1	2.153 4	1.265 7	2.894 3	0.088 9	…	…	…
性　　别	1	−0.417 7	0.441 1	0.896 6	0.343 7	0.659	0.277	1.563
年　　龄	1	−0.785 1	0.236 8	10.989 3	0.000 9	0.456	0.287	0.725
手术类型	1	−0.600 1	0.733 6	0.669 2	0.413 3	0.549	0.130	2.311

表 3　瓣膜生存时间 5 年影响生物瓣膜生存的危险因素逐步 Logistic 回归分析

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Odds Ratio Estimates		
						Point Estimate	95% Wald	Confidence Limits
Intercept	1	0.806 3	0.732 1	1.213 1	0.270 7	…	…	…
年　　龄	1	−0.729 4	0.231 5	9.924 2	0.001 6	0.482	0.306	0.759

义（χ^2=0.416 6，P=0.518 6；χ^2=1.738 3，P=0.187 4；χ^2=0.692 7，P=0.405 2）。表 5 结果显示在检验水准取α=0.05 时，进行多变量逐步 COX 回归分析未见协变量引入方程（NOTE：No（additional）variables met the 0.05 level for entry into the model.）。SAS 参考程序及结果如下[3-4]：

```
data aa.aa2;
set aa.aa1;
if id='.' then delete;
if scsj='.' then delete;
if id >220 then delete;
run;
ods rtf;
proc phreg data=aa.aa2;
model scsj * OK1(0)=sex age3 sslx2/RL; run;
ods rtf close;
ods rtf;
proc phreg data=aa.aa2;
model scsj * OK1 (0)=sex age3 sslx2/RL selection=
stepwise sle=0.05 sls=0.10; run;
ods rtf close;
```

由上可见，甲医生采用二分类因变量多变量

表 4　影响生物瓣膜生存时间危险因素 COX 回归分析

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio	Confidence Limits
SEX	1	0.130 79	0.202 63	0.416 6	0.518 6	1.140	0.766	1.695
AGE3	1	0.122 95	0.093 25	1.738 3	0.187 4	1.131	0.942	1.358
SSLX2	1	0.313 15	0.376 24	0.692 7	0.405 2	1.368	0.654	2.859

表 5　影响生物瓣膜生存时间危险因素逐步 COX 回归分析

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio	Confidence Limits
SEX	…	…	…	…	…	…	…	…
AGE3	…	…	…	…	…	…	…	…
SSLX2	…	…	…	…	…	…	…	…

NOTE：No（additional）variables met the 0.05 level for entry into the model.

Logistic 回归分析、二分类因变量多变量逐步 Logistic 回归分析（取协变量引入方程的检验水准α=0.05，剔出方程的检验水准 α=0.10）的统计学结果，与乙医生采用多变量 COX 回归分析、及多变量逐步 COX 回归分析（协变量引入方程的检验水准 α=0.05，剔出方程的检验水准 α=0.10）的结果不同，甲乙两医生何者在医学统计学思维上科学合理？

从科学合理的统计学思维来分析，乙医生对本例资料的统计学处理分析是恰当的，而甲医生的统计学处理分析存在错误。因为生存分析的因变量是终点事件结局及出现该终点事件结局所经历的时间大小。生存数据常常不满足正态分布要求、不满足一般多变量线性回归模型（下转第 59 页）

European Caucasian population [J]. Ann Rheum Dis, 2010, 69(11):1958-1964.
[10] Kato M, Sanada M, Kato I, et al. Frequent inactivation of A20 in B-cell lymphomas [J]. Nature, 2009,459(7247):712-716.
[11] Coornaert B, Baens M, Heyninck K, et al. T cell antigen receptor stimulation induces MALT1 paracaspase-mediated cleavage of the NF-kappaB inhibitor A20 [J]. Nat Immunol, 2008,9(3):263-271.
[12] Düwel M, Welteke V, Oeckinghaus A, et al. A20 negatively regulates T cell receptor signaling to NF-kappaB by cleaving Malt1 ubiquitin chains [J]. J Immunol, 2009,182(12):7718-7728.
[13] Chu Y, Vahl JC, Kumar D, et al. B cells lacking the tumor suppressor TNFAIP3/A20 display impaired differentiation, hyperactivation, cause inflammation and autoimmunity in aged mice [J]. Blood, 2010 Nov 18. [Epub ahead of print]
[14] Song XT, Evel-Kabler K, Shen L, et al. A20 is an antigen presentation attenuator, and its inhibition overcomes regulatory T cell-mediated suppression [J]. Nat Med, 2008,14(3):258-265.
[15] Verstrepen L, Verhelst K, van Loo G, et al. Expression, biological activities and mechanisms of action of A20 (TNFAIP3)[J]. Biochem Pharmacol, 2010,80 (12):2009-2020.
[16] Compagno M, Lim WK, Grunn A, et al. Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma [J]. Nature, 2009,459(7247):717-21.
[17] Won M, Park KA, Byun HS, et al. Novel anti-apoptotic mechanism of A20 through targeting ASK1 to suppress TNF-induced JNK activation [J]. Cell Death Differ, 2010,17(12):1830-1841.
[18] Hjelmeland AB, Wu Q, Wickman S, et al. Targeting A20 decreases glioma stem cell survival and tumor growth [J]. PLoS Biol, 2010,8(2):e1000319.
[19] Shembade N, Ma A, Harhaj EW. Inhibition of NF-kappaB signaling by A20 through disruption of ubiquitin enzyme complexes [J]. Science, 2010, 327(5969):1135-1139.
[20] Tavares RM, Turer EE, Liu CL, et al. The ubiquitin modifying enzyme A20 restricts B cell survival and prevents autoimmunity [J]. Immunity, 2010,33(2):181-191.
[21] Chanudet E, Huang Y, Ichimura K, et al. A20 is targeted by promoter methylation, deletion and inactivating mutation in MALT lymphoma [J]. Leukemia, 2010,24(2):483-487.
[22] Montesinos-Rongen M, Schmitz R, Brunn A, et al. Mutations of CARD11 but not TNFAIP3 may activate the NF-kappaB pathway in primary CNS lymphoma [J]. Acta Neuropathol, 2010,120 (4):529-535.
[23] Novak U, Rinaldi A, Kwee I, et al. The NF-κB negative regulator TNFAIP3 (A20) is inactivated by somatic mutations and genomic deletions in marginal zone lymphomas [J]. Blood, 2009,113(20):4918-4921.
[24] Philipp C, Edelmann J, Bühler A, et al. Mutation analysis of the TNFAIP3 (A20) tumor suppressor gene in CLL [J]. Int J Cancer, 2010 Jun 7. [Epub ahead of print]
[25] Frenzel LP, Claus R, Plume N, et al. Sustained NF-kappaB activity in chronic lymphocytic leukemia is independent of genetic and epigenetic alterations in the TNFAIP3 (A20) locus [J]. Int J Cancer, 2010 Jul 28. [Epub ahead of print]
[26] Guo Q, Dong H, Liu X, et al. A20 is overexpressed in glioma cells and may serve as a potential therapeutic target [J]. Expert Opin Ther Targets, 2009,13(7):733-741.

[收稿日期] 2010-12-27

(上接第 53 页)的要求,故以生存时间为因变量建立一般多变量线性回归模型不妥。而以某一时点的终点事件结局(如 1=发生终点事件,0=未发生终点事件)为因变量,进行二分类因变量多变量 Logistic 回归分析,则没有充分利用生存时间大小的信息,也没有考虑生存时间数据还存在的数据删失问题,故二分类因变量多变量 Logistic 回归模型也不适合分析此类数据。而且一般多变量线性回归模型模型和二分类因变量多变量 Logistic 回归模型都无法充分利用这类删失数据提供的信息。

故在进行生存数据统计学处理分析、进行影响生存时间大小的多因素分析时,应该采用比例危险率回归模型(COX 回归模型)。这也要求我们在临床医学科研实际工作或论文撰写过程中,需要根据科学研究设计类型 (包含专业研究设计和统计研究设计)及所收集数据研究变量的特征,进行科学合理的临床流行病学思维、医学统计学思维,运行国际上承认统计分析软件(如 SAS),得出正确的统计学结果[5]。

[参 考 文 献]

[1] 方积乾. 医学统计学与电脑实验 [M]. 第 3 版. 上海:上海科学技术出版社, 2006:62-80.
[2] 高惠璇,李贵斌,耿直,等编译. SAS 系统·SAS/STAT 软件使用手册 [M]. 北京:中国统计出版社, 1997:309-338.
[3] 刘勤,金丕焕. 分类数据的统计分析及 SAS 编程 [M]. 上海:复旦大学出版社, 2002.
[4] 方积乾,孙振球. 卫生统计学 [M]. 第 6 版. 北京:人民卫生出版社, 2008:123-155.
[5] 王家良. 临床流行病学—临床科研设计、衡量与评价 [M]. 第 2 版. 上海:上海科学技术出版社, 2001:61-73.

[收稿日期] 2011-01-18