学习·发现·交流

时间相关受试者工作特征曲线及其在临床试验诊断分析中的应用

赵延延 赵维 王子悦 李卫 王杨
102300 北京，国家心血管病中心医学统计部
通信作者：王杨，Email：wangyang@mrbc-nccd.com
DOI：10.3760/cma.j.issn.0254-6450.2016.06.030

【摘要】 利用R语言通过实例介绍生存模型中诊断指标的两种时间相关受试者工作特征［ROC(t)］曲线估计方法，即以NNE（nearest-neighbor estimator of bivariate distribution）估计法获得累积/动态的$ROC^{C/D}(t)$曲线和以Cox估计法获得事件/动态的$ROC^{I/D}(t)$曲线。分析显示利用两种估计法获得的ROC曲线下面积（AUC）值均随时间变化而波动，其中以NNE估计法得到的值波动较大，而用Cox法得到的曲线波动较小，但两种方法所得AUC均值相近。由此表明利用ROC(t)可对临床试验中诊断指标的诊断能力进行评价，有助于对诊断指标选择最佳的诊断时间，但使用中应注意选择相应的估计方法以获得更准确的评价。

【关键词】 受试者工作特征曲线；时间相关受试者工作特征曲线；诊断能力

Time related receiver operation characteristic curves and its application in clinical trials and diagnostic analysis

Zhao Yanyan, Zhao Wei, Wang Ziyue, Li Wei, Wang Yang
Division of Medical Research, National Center for Cardiovascular Disease, Beijing 102300, China
Corresponding author: Wang Yang, Email: wangyang@mrbc-nccd.com

【**Abstract**】 By using R language to deal with practical problems, we introduce two methods of obtaining time related receiver operation characteristic [ROC(t)] curves from survival data: 1) nearest-neighbor estimator of bivariate distribution (NNE) estimation: to obtain cumulative/dynamic $ROC^{C/D}(t)$ curves; 2) Cox estimation: to obtain incident/dynamic $ROC^{I/D}(t)$ curves. The areas under the ROC(t) curves (AUC) obtained from the two methods fluctuate over time. The one obtained through NNE has bigger fluctuation than that obtained through Cox, while the mean of AUC of the two methods are similar. Time related ROC (t) can be effectively used to evaluate the diagnostic capacity of the marker in clinical trials, and help to select the best diagnostic time of the marker. According to the different scientific interests, researchers should select relevant methods for more accurate evaluation.

【**Key words**】 Receiver operation characteristic curves; Time related receiver operation characteristic curves; Diagnostic capacity

受试者工作特征（ROC）曲线作为经典的诊断试验分析方法，被广泛用于定量预后指标对二分类临床结局的诊断性能评价。ROC曲线下面积（AUC）则作为诊断综合准确性的量化评价指标。ROC曲线及AUC通常可以通过构建logistic回归模型获得，用于判断在特定时间点内可能发生的终点事件。

对于以临床终点事件为主要指标的临床研究，多采用生存分析作为统计方法，在评价是否发生事件的同时，还应考虑事件的发生时间。特定的预后指标，其诊断能力可能随时间而发生变化，采用传统方法无法处理。针对该问题，2005年Heagerty和Zheng[1]总结了3种时间依赖的诊断性能定义和评价方法，使研究者可对加入时间后的诊断能力及其动态变化进行评价，类似方法在国内研究中鲜有报道，为此笔者重点复习累积/动态（cumulative/dynamic）定义下的NNE（nearest-neighbor estimator of bivariate distribution）估计方法以及事件/动态（incident/dynamic）定义下的Cox估计方法，并通过实例分别采用两种方法估计与时间相关的ROC曲线及AUC，并比较其结果。

方法原理

对于生存数据，假设发生事件的个体所对应的事件时间为T_i（event time），而删失个体的删失时间为C_i（censoring time），令$Z_i=\min(T_i, C_i)$，定义δ_i，对应数据的删失情况，设$\delta_i=0$时，表示该个体数据为

删失，等于1则为非删失。那么，通常一组生存数据具有的信息将包括时间Z_i、状态δ_i，以及预后指标的取值M_i。c表示分界值(cut-off)。

(1) 累积/动态定义的NNE估计方法：当给定了cut-off值(c)，就能定义与时间相关的C/D型灵敏度及特异度

$$Sensitivity^C(c,t)=P(M>c\mid T\leqslant t) \quad (1)$$

$$Specificity^D(c,t)=P(M\leqslant c\mid T>t) \quad (2)$$

该定义方法是将整个人群在时间t划分为事件组与非事件组两类。当$T_i>t$时，个体被划分到非事件组；当$T_i\leqslant t$时，个体被划分到事件组，使个体i在不同时间t扮演的角色不同。由于采用传统KM估计方法估计每种可能的情况$X>c$(假设X为预测死亡事件的预测指标；c为区分预测的界值，可取不同的值)的灵敏度和特异度时，不能保证其单调性，而NNE估计方法进行估计时[即采用基于最近邻估计量的二维分布函数(X,T)估计，其中T为生存时间]，可保证其单调性。比较传统KM估计法，Heagerty等[2]提出了更为合理的NNE方法估计生存函数，具体步骤：首先估计联合生存函数，其形式为

$$S(c,t)=P(M>c,T>t)=\int_c^{\infty}S(t|M=s)dF_X(s) \quad (3)$$

$F_X(s)$为X的分布函数，$S(c,t)$的NNE方法估计值为

$$\hat{S}_{\lambda_n}(c,t)=\frac{1}{n}\sum_i\hat{S}_{\lambda_n}(t|M=M_i)1(M_i>c) \quad (4)$$

式中$1(M_i>c)$为示性函数，$\hat{S}_{\lambda_n}(t|M=M_i)$为利用近邻核函数

$$K_{\lambda_n}(M_i,M_j)=1\{-\lambda_n<\hat{F}_X(M_i)-\hat{F}_X(M_j)<\lambda_n\} \quad (5)$$

得到加权的KM生存函数估计

$$\hat{S}_{\lambda_n}(t|M=M_i)=\prod_{s\in\tau_n,s\leqslant t}\left(1-\frac{\sum_j K_{\lambda_n}(M_j,M_i)1(Z_i=s)\delta_j}{\sum_j K_{\lambda_n}(M_j,M_i)1(Z_j\geqslant s)}\right) \quad (6)$$

其中τ_n定义为所有不同Z_i值的集合，近邻核函数的定义中，λ_n代表选取“近邻”的范围，且$2\lambda_n\in(0,1)$，通常选取$\lambda_n=O(n^{-1/3})$即可[3]。另外还可以得到NNE生存函数的估计

$$\hat{S}_{\lambda_n}(t)=\hat{S}_{\lambda_n}(-\infty,t)$$

基于上述，再估计灵敏度和特异度。首先由Bayes条件概率公式

$$\hat{P}_{\lambda_n}(M>c\mid T\leqslant t)=\frac{\hat{P}_{\lambda_n}(M>c,T\leqslant t)}{\hat{P}_{\lambda_n}(T\leqslant t)}$$

$$=\frac{\hat{P}_{\lambda_n}(M>c)-\hat{P}_{\lambda_n}(M>c,T>t)}{\hat{P}_{\lambda_n}(T\leqslant t)}$$

$$\hat{P}_{\lambda_n}(M\leqslant c\mid T>t)=1-\hat{P}_{\lambda_n}(M>c\mid T>t)$$

$$=1-\frac{\hat{P}_{\lambda_n}(M>c,T>t)}{\hat{P}_{\lambda_n}(T>t)}$$

得到

$$\widehat{Sensitivity}^C(c,t)=P(M>c\mid T\leqslant t)=\frac{1-\hat{F}_X(c)-\hat{S}_{\lambda_n}(c,t)}{1-\hat{S}_{\lambda_n}(t)} \quad (7)$$

$$\widehat{Specificity}^D(c,t)=P(M\leqslant c\mid T>t)=1-\frac{\hat{S}_{\lambda_n}(c,t)}{S_{\lambda_n}(t)} \quad (8)$$

(2)事件/动态定义的Cox估计方法：2005年Heagerty和Zheng[1]给出了利用Cox模型对时间依赖的I/D型灵敏度与特异度的估计方法，从而可以得到时间依赖的ROC曲线以及AUC(t)。

I/D法定义为

$$Sensitivity^I(c,t)=P(M>c\mid T=t) \quad (9)$$

$$Specificity^D(c,t)=P(M\leqslant c\mid T>t) \quad (10)$$

即灵敏度表示在t时刻死亡的人群中，预后指标取值$>c$的被观察对象所占的比例；特异度表示在活过时间t的人群中，指标值$\leqslant c$的人所占的比例。该定义方法是将t时间上还处于研究观察的人群(风险人群)划分为事件组与非事件组两类，使得个体在其死亡时间点以前作为非事件组，而在其死亡时间t时作为事件组。该定义①与比例风险模型中对试验组数据与非事件组数据的划分紧密联系；②允许将其扩展运用于对时间相关的指标值上，并运用于非比例Cox模型中；③可得到进行时间平均的总结评估值，从而对该指标值的诊断能力进行与时间无关的整体评价。

Cox模型估计时间依赖的I/D型灵敏度与特异度的具体步骤：通过Cox回归，得到每个受试者的得分$M_i=Z_i^T\beta$，定义

$R_i(t)=1(Z_i\geqslant t)$

表示t时间的风险人群，以及

$\lambda(t|M_i)=\lambda_0(t)\exp(M_i\gamma)$

表示Cox模型下受试者i在时间t的风险函数，那么在Cox模型下，给定$R_i(t)=1$的人群中存在个体死亡条件下，得分为M_i的个体在时间t死亡概率为

$$\pi_k(\gamma,t)=P(\text{个体}k\text{发生事件}|\text{有一个个体发生事件})=\frac{R_k(t)\exp(M_k\gamma)}{\sum_j R_j(t)\exp(M_j\gamma)} \quad (11)$$

再定义

$$\pi_k(\gamma,t^+)=\frac{R_k(t^+)\exp(M_k\gamma)}{\sum_j R_j(t^+)\exp(M_j\gamma)} \quad (12)$$

Xu和O'Quigley[4]提出可用上述$\pi_k(\gamma,t)$来估计得分M_i在时间t下的分布，即

$$\hat{P}(M>c|T=t)=\sum_j \pi_j(\gamma,t)\cdot 1(M_j>c) \quad (13)$$

利用经验估计法可得到

$$\hat{P}(M\leqslant c|T>t)=\sum_j \pi_j(\gamma,t^+)\cdot 1(M_j\leqslant c) \quad (14)$$

即得到了灵敏度与特异度的估计值

$$\widehat{Sensitivity}(c,t)=\frac{\sum_k R_k(t)\exp(M_k\gamma)1(M_j>c)}{\sum_j R_j(t)\exp(M_j\gamma)} \quad (15)$$

$$\widehat{Specificity}(c,t)=\frac{\sum_k R_k(t^+)\exp(M_k\gamma)1(M_j\leqslant c)}{\sum_j R_j(t^+)\exp(M_j\gamma)} \quad (16)$$

实例分析

现有一组1 528例接受过PCI治疗的冠心病患者数据库，其中包括患者编号、死亡/随访时间、删失状态等生存分析相关变量，同时还包括对PCI患者术后死亡风险有预测作用的SYNTAX SCORE及SYNTAX SCORE Ⅱ组成。其中，SYNTAX SCORE是早期被验证过的，基于患者冠状动脉病变复杂程度获得的死亡风险预后指标；SYNTAX SCORE Ⅱ则是在SYNTAX SCORE的基础上，不仅仅通过冠脉复杂程度，而是进一步结合患者年龄、肌酐清除率、左心室射血分数（LVEF）、无保护的左主干（ULMCA）病变、周围血管病变、性别及慢性阻塞性肺病（COPD）等人口学和疾病史危险因素后，建立起的能够对PCI术后死亡风险进行更准确预测的指标[5]。

基于上述数据，分别应用累积/动态和事件/动态两种方法，估计不同时间下SYNTAX SCORE Ⅱ的诊断能力。运用R语言软件包survivalROC中的函数survivalROC，可以得到累积/动态法下每一时间点对应的ROC曲线（$ROC^{C/D}$）及$AUC^{C/D}(t)$的值；运用R语言软件包risksetROC中的函数risksetROC和risksetAUC，可以获得事件/动态法下每一时间点对应的ROC曲线（$ROC^{I/D}$）及其对应的$AUC^{I/D}(t)$估计值。

图1展示了不同时间的$ROC^{C/D}$和$ROC^{I/D}$曲线图，其中仅选取了有代表性的时间点，分别为0.5、1、2、4、6及7.2年的结果。事实上，对于任意时间点（数据中最后一个时间点之前）的ROC曲线，其结果与上述最接近的代表性时点结果基本一致。

从图1A可见，不同时间点对应的SYNTAX SCORE Ⅱ的诊断能力会发生一定的改变，2、4、6和7.2年的诊断能力相对优于0.5和1年的结果。而在图1B中，各时间点SYNTAX SCORE Ⅱ的诊断能力虽有改变，但无明显差异。

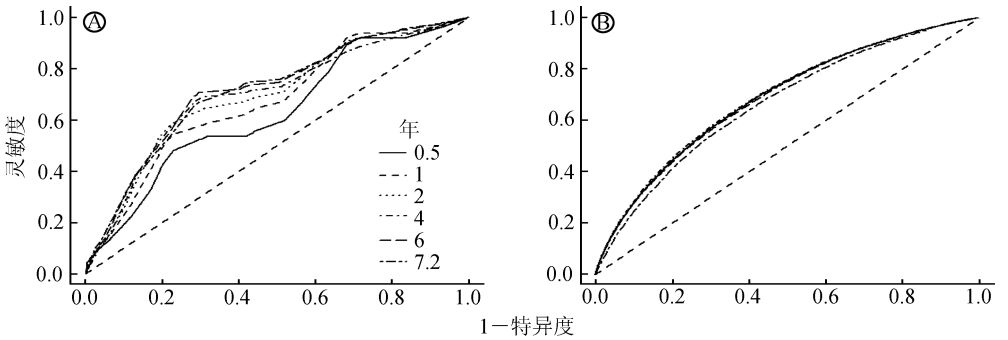
表1为不同时间点两方法分别对应的$AUC^{C/D}$值和$AUC^{I/D}$值。为便于比较，将其中$AUC^{C/D}(t)$和$AUC^{I/D}(t)$作为纵坐标，时间点作为横坐标，则获得图2。其结果与前类似，累积/动态法给出的SYNTAX SCORE Ⅱ诊断能力的估计结果，随时间会有较大的波动，在1年内的时点，出现了$AUC^{C/D}(t)$值＜0.65的情况，而在2年这一时点后，$AUC^{C/D}(t)$约为0.70，波动中有小幅上升趋势；而事件/动态法给出的$AUC^{I/D}(t)$结果，则整个观察期内相对稳定，在4～5年内基本在0.69左右，之后时点的诊断能力则有小幅下降趋势。

因此，临床研究者通过C/D型定义法，在利用SYNTAX SCORE Ⅱ对患者进行诊断分析时，其在2年以前的诊断能力相对较弱；而通过I/D型定义法，SYNTAX SCORE Ⅱ在4年后的诊断能力有所下降。

在获得SYNTAX SCORE Ⅱ于不同时间上的诊断能力后，就可以通过该指标对患者术后情况进行诊断预测。例如，通过患者的SYNTAX SCORE Ⅱ值，可对患者进行术后4年的死亡预测。

通过上述NNE估计法与Cox估计法的比较，显示两种方法得到的结论差异较大。以下分析两种估计法得到不同结果的原因。

图3是北京阜外医院冠心病患者术后发生事件



注：A.NNE方法的$ROC^{C/D}$曲线图；B.Cox方法的$ROC^{I/D}$曲线图

图1 诊断时间点分别为0.5、1、2、4、6和7.2年的ROC曲线

表1　不同时间点的$AUC^{C/D}$值与$AUC^{I/D}$值

t/年	$AUC^{C/D}$	$AUC^{I/D}$	t/年	$AUC^{C/D}$	$AUC^{I/D}$	t/年	$AUC^{C/D}$	$AUC^{I/D}$
0	0.686	0.691	0.758	0.658	0.690	3.398	0.682	0.693
0.003	0.743	0.692	0.813	0.666	0.690	3.806	0.686	0.694
0.005	0.744	0.690	1.016	0.673	0.690	3.874	0.690	0.693
0.008	0.743	0.691	1.021	0.680	0.690	3.926	0.695	0.694
0.016	0.758	0.691	1.087	0.683	0.690	3.995	0.700	0.694
0.019	0.763	0.691	1.492	0.691	0.690	4.027	0.703	0.694
0.022	0.727	0.690	1.717	0.696	0.691	4.197	0.707	0.694
0.049	0.725	0.691	1.766	0.701	0.691	4.318	0.711	0.694
0.123	0.699	0.691	1.771	0.707	0.690	5.027	0.707	0.687
0.142	0.675	0.691	1.791	0.710	0.691	5.051	0.705	0.688
0.151	0.628	0.691	1.854	0.714	0.690	5.062	0.686	0.688
0.175	0.638	0.691	1.892	0.720	0.690	5.388	0.696	0.684
0.277	0.653	0.691	1.927	0.716	0.690	5.530	0.699	0.685
0.334	0.636	0.690	2.029	0.693	0.690	5.574	0.703	0.686
0.411	0.620	0.690	2.261	0.698	0.689	5.848	0.710	0.690
0.441	0.634	0.690	2.533	0.703	0.691	5.982	0.717	0.690
0.507	0.629	0.690	2.549	0.702	0.690	6.012	0.716	0.688
0.528	0.638	0.690	2.724	0.704	0.690	6.059	0.720	0.689
0.548	0.643	0.690	2.784	0.708	0.691	6.204	0.716	0.686
0.600	0.651	0.690	2.932	0.713	0.691	6.672	0.722	0.685
0.726	0.650	0.690	2.995	0.697	0.691	7.233	0.714	0.670
0.728	0.646	0.690	3.102	0.700	0.692			
0.753	0.653	0.690	3.332	0.674	0.692			

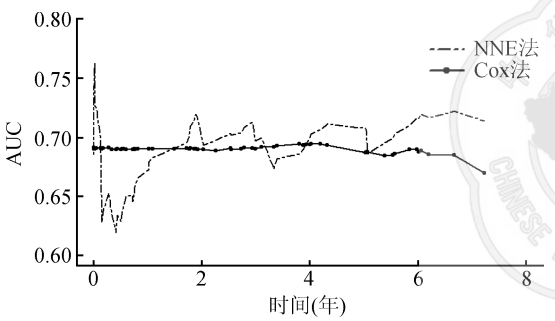


图2　NNE估计法和Cox估计法的AUC(t)

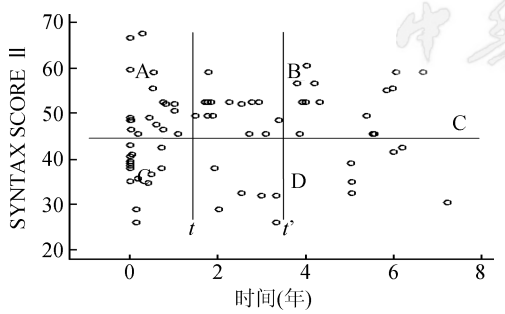


图3　术后76个发生事件个体的时间–得分散点分析

的76个个体SYNTAX SCORE Ⅱ值与发生时间的散点图。当给定预测时间t时，对于一系列c值：NNE方法估计时，通过灵敏度与特异度的定义，对所得到的ROC曲线上每一点的横纵坐标值（FP值和TP值）产生影响的分别是B区和A区个体的比例，且该方法中A、B、C、D四区的个体权重相等，这导致估计时间点t变化到t'时，所得的各估计值变化较大；Cox方法估计时，对每次估计产生不同影响的分别是B区个体和时间t对应的个体比例，且估计时是通过$\exp(M_k\gamma)$对不同个体赋予不同权重[1]，因此M_k较大的个体权重较大，导致有影响的B区个体对其结果影响更大，因此导致估计时间点t变化到t'时，所得的各估计值变化较小。由此导致当t变化的过程中，Cox方法得到的SYNTAX SCORE Ⅱ诊断能力随时间的变化明显减小。

事实上，Cox方法中危险因素（如本例中SYNTAX SCORE Ⅱ得分越高，个体死亡可能性越大）较高的个体其影响力更大，而NNE方法并未从这个角度进行考虑。因此，在实际临床实验中，若判断高得分者明显影响试验结果，则选取Cox方法应更为合理。

小结：由图2可见，两种方法总体变化的差异较大。主要原因是其灵敏度与特异度不同，且估计方法亦有差异。在临床试验中，如何准确选择更好的方法评价诊断指标，这与研究者关心的实际数据有关。一般有以下几点：①如研究者关心该诊断指标为受试者在特定时间点t之前发生的事件，并能够活过t，应选用NNE估计法；②若关心t时间以后的受试人群，在t时间发生事件并可活过t，应选Cox估计法；③若判断高得分者明显影响试验结果，则选取Cox方法更为合理；④Heagerty和Zheng[1]提出，由于Cox方法中灵敏度与特异度的定义在时间t对事件组与非事件组的划分上与Cox风险模型一致，可将其推广至随时间变化的指标值上，因此对于非比例风险模型同样适用；此外，还可通过均分时间得到平均时间的总体评价结果；⑤Cox估计法无法估计样本中最后一个事件时间点以后的诊断能力；⑥建议在临床试验中通过敏感性分析方式，分别计算两种模型结果，并综合考虑再得出相应结论，这样更助于提高结论的可靠性。

利益冲突　无

参　考　文　献

[1] Heagerty PJ, Zheng YY. Survival model predictive accuracy and ROC curves[J]. Biometrics, 2005, 61(1): 92–105. DOI: 10.1111/j.0006–341X.2005.030814.x.

[2] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker [J]. Biometrics, 2000, 56(2): 337–344. DOI: 10.1111/j.0006–341X.2000.00337.x.

[3] Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring[J]. Ann Statist, 1994, 22(3): 1299–1327. DOI: 10.1214/aos/1176325630.

[4] Xu RH, O' Quigley J. Proportional hazards estimate of the conditional survival function[J]. J Royal Stat Soc Ser B (Stat Methodol), 2000, 62(4): 667–680. DOI: 10.1111/1467–9868.00256.

[5] Farooq V, van Klaveren D, Steyerberg WW, et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score Ⅱ [J]. Lancet, 2013, 381(9867): 639–650. DOI: 10.1016/S0140–6736(13)60108–7.

（收稿时间：2015–09–25）
（本文编辑：张林东）