
基于Cox回归模型构建疾病风险评分工具

作者：龚志忠 来源：医咖会

本文原地址：<https://www.iikx.com/news/statistics/22660.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

基于Cox回归模型构建疾病风险评分工具。2015年BMJ杂志发表了题为《Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement》的论文，即TRIPOD声明，对于疾病诊断和预后的预测模型研究统一了报告规范。随着疾病预防关口的前移，越来越多的研究者关注疾病预测模型的研究，评分工具作为一种简便快速的评价方法，得到了广泛的应用。本期我们就来介绍基于Cox回归模型构建疾病风险评分工具的方法。

研究实例.

我们以D'Agostino等人2001年发表在JAMA期刊的一篇研究为例(《Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation》), 该研究利用多种族人群的数据，对Framingham冠心病预测评分工具进行了验证。

我们选取其中FHS队列白人男性的数据结果进行说明，队列共包含2439名健康男性，年龄30-74岁，随访长达10年，观察研究对象冠心病的发病情况。研究人员采用多因素Cox回归模型进行分析，最终筛选纳入模型的危险因素包括：年龄、血压、总胆固醇(TC)、高密度脂蛋白胆固醇(HDL-C)、吸烟、糖尿病。

步骤.

1. 构建多因素Cox回归模型.

通过构建多因素Cox回归模型，将我们主要考虑的危险因素纳入到回归模型中，从而估计各个危险因素的回归系数 β 。

同时，计算各个危险因素的均值，或者危险因素每个分类所占的比例，我们在后面的第7步中会用到这部分数值，这里先留下一个伏笔。

例如在本例研究中，年龄为连续型变量，计算研究人群的平均年龄为48.3岁。HDL虽然也为连续型变量，但研究人员对HDL进行了分组，将其转化为分类变量。共分为5组，每组所占的比例分别为19%、36%、15%、19%、11%，并且以中间一组(HDL 45-49)作为参照组(Base)。

Risk factor	Regression coefficient	Mean or proportion
Age, years	0.0533	48.3
Optimal BP (DBP [†] < 80 and SBP [‡] < 120)	0.0948	0.20
Normal BP (DBP 80–84 or SBP 120–129)	Base	0.24
Hi Normal BP (DBP 85–89 or SBP 130–139)	0.4225	0.20
Stage I Hypertension (DBP 90–99 or SBP 140–159)	0.6596	0.23
Stage II+ Hypertension (DBP ≥ 100 or SBP ≥ 160)	0.8964	0.13
Total cholesterol < 160	−0.3781	0.07
Total cholesterol 160–199	Base	0.31
Total cholesterol 200–239	0.5696	0.39
Total cholesterol 240–279	0.7438	0.17
Total cholesterol ≥ 280	0.8284	0.06
HDL < 35	0.6074	0.19
HDL 35–44	0.3684	0.36
HDL 45–49	Base	0.15
HDL 50–59	0.0000	0.19
HDL ≥ 60	−0.4608	0.11
Current smoker	0.7277	0.40
Diabetes	0.5252	0.05

Average 10-year survival = 0.943*

2. 将各个危险因素进行分类，并指定每组的参考值Wij.

我们按照临床意义或使用习惯将各个危险因素进行分组，并在每个分组中选择合适的数值作为参考值Wij，通常选择组内的中间值作为参考值。

例如在本例中，研究人群的年龄范围为30-74岁，我们按照5岁一个年龄段来进行划分，共分为9组，每组选择中间值为参考值Wij，例如30-34岁这一组的参考值Wij为 $(30+34)/2=32$ 。

在本例中，吸烟和糖尿病为二分类变量，而血压、TC和HDL在构建Cox回归模型时，研究人员已经事先进行了分组，转换为了哑变量，因此统一设定状态为No时的参考值Wij为0，Yes时为1。

Risk factor	Categories	Reference value (W_{ij})
Age*	30–34	32
	35–39	37
	40–44	42
	45–49	47
	50–54	52
	55–59	57
	60–64	62
	65–69	67
	70–74	72
Optimal BP (DBP < 80 and SBP < 120)	No	0
	Yes	1
High normal BP (DBP 85–89 or SBP 130–139)	No	0
	Yes	1
Stage I hypertension (DBP 90–99 or SBP 140–159)	No	0
	Yes	1
Stage II+ hypertension (DBP ≥ 100 or SBP ≥ 160)	No	0
	Yes	1
Total cholesterol < 160	No	0
	Yes	1
Total cholesterol 200–239	No	0
	Yes	1
Total cholesterol 240–279	No	0
	Yes	1
Total cholesterol ≥ 280	No	0
	Yes	1
HDL < 35	No	0
	Yes	1
HDL 35–44	No	0
	Yes	1
HDL 50–59	No	0
	Yes	1
HDL ≥ 60	No	0
	Yes	1
Current smoker	No	0
	Yes	1
Diabetes	No	0
	Yes	1

3. 确定各个危险因素的基础风险参考值WiREF.

对于每一个危险因素，我们需要选择一个合适的分组来作为基础风险参考值WiREF，在后续构建评分工具时，该组分值将记为0分，危险因素的值高于WiREF时记正分，得分越高则风险越高，相反低于WiREF时记负分。

例如在本例研究中，对于年龄变量，我们可以选择年龄为42岁作为基础风险参考值WiREF。

Risk factor	Categories	Reference value (W_{ij})
Age	30–34	32
	35–39	37
	40–44	42 = W_{1REF}
	45–49	47
	50–54	52
	55–59	57
	60–64	62
	65–69	67
	70–74	72
Optimal BP (DBP < 80 and SBP < 120)	No	0 = W_{2REF}
	Yes	1
High normal BP (DBP 85–89 or SBP 130–139)	No	0 = W_{2REF}
	Yes	1
Stage I hypertension (DBP 90–99 or SBP 140–159)	No	0 = W_{2REF}
	Yes	1
Stage II+ hypertension (DBP ≥ 100 or SBP ≥ 160)	No	0 = W_{2REF}
	Yes	1
Total cholesterol < 160	No	0 = W_{3REF}
	Yes	1
Total cholesterol 200–239	No	0 = W_{3REF}
	Yes	1
Total cholesterol 240–279	No	0 = W_{3REF}
	Yes	1
Total cholesterol ≥ 280	No	0 = W_{3REF}
	Yes	1
HDL < 35	No	0 = W_{4REF}
	Yes	1
HDL 35–44	No	0 = W_{4REF}
	Yes	1
HDL 50–59	No	0 = W_{4REF}
	Yes	1
HDL ≥ 60	No	0 = W_{4REF}
	Yes	1
Current smoker	No	0 = W_{5REF}
	Yes	1
Diabetes	No	0 = W_{6REF}
	Yes	1

4. 计算每一个危险因素的分组与基础风险参考值之间的距离D.

结合多因素Cox回归模型估计的回归系数 β_i ，以及危险因素各组的参考值 W_{ij} ，来计算危险因素的每一分组与基础风险参考值 W_{iREF} 之间的距离D，计算公式为 $D = (W_{ij} - W_{iREF}) * \beta_i$

例如在本研究中，年龄的基础风险参考值 W_{iREF} 为42，在Cox回归中年龄对应的回归系数 β_i 为0.0533，那么对于50-54岁组，其参考值 W_{ij} 为52，该组与基础风险参考值的距离D为 $(52 - 42) * 0.0533 = 0.5330$ 。

Risk factor	Categories	Reference value (W_{ij})	β_i	$\beta_i(W_{ij} - W_{iREF})$
Age			0.0533	
	30-34	32		-0.5330
	35-39	37		-0.2665
	40-44	42 = W_{1REF}		0
	45-49	47		0.2665
	50-54	52		0.5330
	55-59	57		0.7995
	60-64	62		1.0660
	65-69	67		1.3325
	70-74	72		1.5990
Optimal BP			0.0948	
	No	0 = W_{2REF}		0
	Yes	1		0.0948
High normal BP			0.4225	
	No	0 = W_{2REF}		0
	Yes	1		0.4225
Stage I hypertension			0.6596	
	No	0 = W_{2REF}		0
	Yes	1		0.6596
Stage II+ hypertension			0.8964	
	No	0 = W_{2REF}		0
	Yes	1		0.8964
Total cholesterol <160			-0.3781	
	No	0 = W_{3REF}		0
	Yes	1		-0.3781
Total cholesterol 200-239			0.5696	
	No	0 = W_{3REF}		0
	Yes	1		0.5696
Total cholesterol 240-279			0.7438	
	No	0 = W_{3REF}		0
	Yes	1		0.7438
Total cholesterol ≥ 280			0.8284	
	No	0 = W_{3REF}		0
	Yes	1		0.8284
HDL <35			0.6074	
	No	0 = W_{4REF}		0
	Yes	1		0.6074
HDL 35-44			0.3684	
	No	0 = W_{4REF}		0
	Yes	1		0.3684
HDL 50-54			0	
	No	0 = W_{4REF}		0
	Yes	1		0
HDL ≥ 60			-0.4608	
	No	0 = W_{4REF}		0
	Yes	1		-0.4608
Current smoker			0.7277	
	No	0 = W_{5REF}		0
	Yes	1		0.7277
Diabetes			0.5252	
	No	0 = W_{6REF}		0
	Yes	1		0.5252

5. 设定评分工具中1分对应的常数B.

我们需要设定评分工具中每记1分时，对应的各个危险因素变化的距离常数。

例如在本例中，如果我们设定年龄每增加5岁时记为1分，那么此时常数 $B=5*0.0533=0.2665$

6. 计算危险因素每个分类对应的分值Points_{ij}.

在第5步确定常数B的基础上，来计算危险因素每一个分类所对应的分值，计算公式为 $Points_{ij}=D/B=(W_{ij}-W_{iREF})*\beta_i/B$ ，最后将计算出来的数值四舍五入取整，即为该组对应的分值。

例如在本研究中的年龄50-54组，其计算的分数为 $0.533/0.2665=2$ 分。

Risk factor	Categories	Reference value (W_{ij})	β_i	$\beta_i(W_{ij} - W_{iREF})$	Points _{ij} = $\beta_i(W_{ij} - W_{iREF})/B$
Age			0.0533		
	30–34	32		−0.5330	−2
	35–39	37		−0.2665	−1
	40–44	42 = W_{1REF}		0	0
	45–49	47		0.2665	1
	50–54	52		0.5330	2
	55–59	57		0.7995	3
	60–64	62		1.0660	4
	65–69	67		1.3325	5
	70–74	72		1.5990	6
Optimal BP			0.0948		
	No	0 = W_{2REF}		0	0
	Yes	1		0.0948	0
High normal BP			0.4225		
	No	0 = W_{2REF}		0	0
	Yes	1		0.4225	2
Stage I hypertension			0.6596		
	No	0 = W_{2REF}		0	0
	Yes	1		0.6596	2
Stage II+ hypertension			0.8964		
	No	0 = W_{2REF}		0	0
	Yes	1		0.8964	3
Total cholesterol <160			−0.3781		
	No	0 = W_{3REF}		0	0
	Yes	1		−0.3781	−1
Total cholesterol 200–239			0.5696		
	No	0 = W_{3REF}		0	0
	Yes	1		0.5696	2
Total cholesterol 240–279			0.7438		
	No	0 = W_{3REF}		0	0
	Yes	1		0.7438	3
Total cholesterol ≥280			0.8284		
	No	0 = W_{3REF}		0	0
	Yes	1		0.8284	3
HDL <35			0.6074		
	No	0 = W_{4REF}		0	0
	Yes	1		0.6074	2
HDL 35–44			0.3684		
	No	0 = W_{4REF}		0	0
	Yes	1		0.3684	1
HDL 50–54			0		
	No	0 = W_{4REF}		0	0
	Yes	1		0	0
HDL ≥60			−0.4608		
	No	0 = W_{4REF}		0	0
	Yes	1		−0.4608	−2
Current smoker			0.7277		
	No	0 = W_{5REF}		0	0
	Yes	1		0.7277	3
Diabetes			0.5252		
	No	0 = W_{6REF}		0	0
	Yes	1		0.5252	2

7. 计算总分与风险预测概率的对应表.

根据第6步的结果，可以将每个危险因素的分值相加来计算总分，理论上每个危险因素取最低值时，可以得到总分最低值为 $(-2)+0+(-1)+(-2)+0+0=-5$ ，同理可得到总分最高值为 $6+3+3+2+3+2=19$ ，因此总分的范围为： $-5\sim 19$ 分。

然后再根据多因素Cox回归模型的方程，来计算每一分值对应的风险预测概率值，计算公式如下：

$$\hat{p} = 1 - S_0(t)^{\exp\left(\sum_{i=1}^P \beta_i X_i - \sum_{i=1}^P \beta_i \bar{X}_i\right)}$$

在第5步中我们设定了年龄每增加5岁时记为1分，年龄的回归系数为0.0533，基础风险参考值WiREF为42，因此我们可以估算公式中对应的数值，其中：

$$\sum_{i=1}^P \beta_i X_i \approx 0.0533 * 42 + B * \text{总分}; \quad \sum_{i=1}^P \beta_i \bar{X}_i \approx \text{Mean or Proportion} * \beta_i$$

$$\begin{aligned} \text{在本例中 } \sum_{i=1}^P \beta_i \bar{X}_i &= 0.0533 * 48.3 + 0.0948 * 0.2 + 0 * 0.24 + 0.4225 * 0.2 + \\ &0.6596 * 0.23 + 0.8964 * 0.13 + (-0.3781) * 0.07 + 0 * 0.31 + 0.5696 * 0.39 + \\ &0.7438 * 0.17 + 0.8284 * 0.06 + 0.6074 * 0.19 + 0.3684 * 0.36 + 0 * 0.15 + 0 * \\ &0.19 + (-0.4608) * 0.11 + 0.7277 * 0.4 + 0.5252 * 0.05 = 3.8326 \end{aligned}$$

此外，公式中 $S_0(t)$ 表示10年平均生存率，本例中根据原始数据采用Kaplan-Meier方法可算出 $S_0(t)=0.943$ (见步骤1中的表格)。

最终即可计算出总分与风险预测概率的对应表，如下表所示。

Point total	Estimate of risk	Point total	Estimate of risk
–5	0.0032	8	0.0950
–4	0.0041	9	0.1221
–3	0.0053	10	0.1564
–2	0.0069	11	0.1991
–1	0.0090	12	0.2516
0	0.0118	13	0.3150
1	0.0153	14	0.3897
2	0.0200	15	0.4751
3	0.0260	16	0.5689
4	0.0338	17	0.6666
5	0.0439	18	0.7616
6	0.0569	19	0.8462
7	0.0736		

Risk factor	Categories	Points
Age	30–34	–2
	35–39	–1
	40–44	0
	45–49	1
	50–54	2
	55–59	3
	60–64	4
	65–69	5
	70–74	6
Blood pressure	Optimal	0
	Normal	0
	High normal	2
	Stage I hypertension	2
	Stage II+ hypertension	3
Total cholesterol	< 160	–1
	160–199	0
	200–239	2
	240–279	3
	≥ 280	3
HDL	< 35	2
	35–44	1
	45–49	0
	50–59	0
	> 60	–2
Current smoker	No	0
	Yes	3
Diabetes	No	0
	Yes	2

评分工具.

与Cox回归模型结果比较.

风险评分工具已经做好，为了进一步验证它的准确性，我们举一个实例，来比较一下评分预测结果与原始的Cox回归模型预测的结果之间的差距。

例如有一名患者，63岁，收缩压130mmHg，舒张压85mmHg，总胆固醇TC为215，HDL为48，糖

尿病史，否认吸烟，根据评分系统里各个危险因素的分值，分别记为4、2、2、2和0分，总分为10分，根据刚刚做好的评分工具，可以得出其对应的风险概率为15.64%。

我们再根据多因素Cox回归模型进行一次计算：

$$\begin{aligned}\sum_{i=1}^p \beta_i X_i &\approx 0.0533 * 62 + 0.0948 * 0 + 0.4225 * 1 + 0.6596 * 0 + 0.8964 * 0 + \\ &(-0.3781) * 0 + 0.5696 * 1 + 0.7438 * 0 + 0.8284 * 0 + 0.6074 * 0 + 0.3684 * \\ &0 + (-0.4608) * 0 + 0.7277 * 0 + 0.5252 * 1 = 4.8219 \\ \sum_{i=1}^p \beta_i \bar{X}_i &= 3.8326\end{aligned}$$

对应的风险概率值为：

$$\hat{p} = 1 - 0.943^{\exp(4.8219 - 3.8326)} = 0.1461 = 14.60\%$$

可以看出，评分工具与Cox回归模型预测结果之间仅仅只相差1%，满足疾病风险预测评估的要求，而且应用起来也很直观和便捷。

至此，我们已经介绍完了分别利用Logistic和Cox回归模型构建疾病风险评分工具的方法，两者步骤基本相同，但在细节上也有一些区别，希望能够给大家的临床研究带来一些帮助。

更多统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发