丁香园

丁香公开课
CLASS.DXY.CN

临床预测模型构建&机器学习(R语言进阶)

第8章 多元回归分析中变量筛选方法

周支瑞

CONTENT

影响/危险因素筛选的思路

- 第一种思路：先做单因素分析，P值显著的变量（比如 $P < 0.1$）放入多元回归方程。
- 第二种思路：危险因素研究是根据对暴露因素（X）效应值的影响筛选协变量的。

影响/危险因素筛选的步骤

- 步骤 1：先看“Z”与Y有没有联系，用单因素分析，看“Z”的P值。
- 步骤 2：再看调整“Z”与不调整“Z”，X对Y的作用是否有变化。先运行基本模型，记录β1，再在该模型中加入“Z”，看β1变化多大？基本模型中可以加入一些必须要调整的变量。
- 步骤 3：再运行一个完整的模型，即调整所有可能的因素，然后从模型中剔除“Z”，看X的回归系数β1的变化。按照上述思路，比较不同的模型，观察X的回归系数的变化，通常认为变化不超过10%可以不调整。

- [举例]BMJ发表的一项研究，分析胎儿生长受限X与心血管风险Y的关系。 调整变量如果对X的作用影响超过10%才调整该变量“We selected these confounders on the basis of their associations with the outcomes of interest or a change in effect estimate of more than 10%.”。

BMJ. 2014 Jan 23;348:g14. doi: 10.1136/bmj.g14.

如何确定多因素回归分析的候选变量?

- [举例][A Prospective Natural-History Study of Coronary Atherosclerosis] Wrote: “Baseline variables that were considered **clinically relevant** or that showed a **univariate relationship with outcome** were entered into multivariate Cox proportional-hazards regression model. Variables for inclusion were carefully chosen, **given the number of events available**, to ensure parsimony of the final model.”

N Engl J Med. 2011 Jan 20;364(3):226-35

1. 临床专业知识

- 首先从临床专业知识的角度考虑，它的作用一定是被人们可接受的，可以从某个生理机制或途径去进行合理的解释。我们常见的候选变量包括以下几类：
  - 人口学资料：例如 性别、年龄、学历、职业、身高、体重 等
  - 生活习惯：例如 吸烟、饮酒、体育锻炼 等
  - 病史信息：例如 家族史、既往史（高血压、糖尿病、心梗等）等
  - 检查信息：例如 血液检查指标（LDL-C、CRP）、其他检查项目 等
  - 治疗信息：例如 既往用药、手术 等
  - 暴露/处理因素
- 针对以上很多候选变量无从下手时，我们可以参考既往发表的文献，总结出已公开发表报道过的对结局事件有独立作用的变量，将它们作为重点的候选变量以供备选。

2. 根据单因素分析结果筛选变量

- 第一，从统计分析角度讲，传统单因素分析方法的结果展示相对简单，它们仅能提示组间均值或率的分布差异有无统计学显著性；而采用单因素回归分析，除了定性的展示组间差异外，还可以提供更为丰富的信息，比如偏回归系数(β)的估计值、效应量估计值(OR、RR值)及可信区间等

2. 根据单因素分析结果筛选变量

- 第二，对于回归分析来说，先做单因素回归，再做多因素回归，这种分析思路展现了从单独一个因素到筛选多个混杂因素的变化过程。单因素回归分析的结果对于变量的筛选就显得很有意义，我们可以根据前后偏回归系数或者OR值的变化，来协助判断是否需要将其纳入到多因素回归中进行调整和控制。通过单因素分析结果，可以帮助我们来判断哪些因素是对结局事件有影响的可疑因素，从而将其作为多因素分析的候选变量。
- 第三、*p* value < 0.2 on univariate analysis were included。

2. 根据单因素分析结果筛选变量【举例】

- 【举例】2013年发表在JACC杂志(影响因子:19.9)上的文章《Predictors for Functionally Significant In-Stent Restenosis》作者在统计方法中这样写道：Candidate variables with a *p* value < 0.2 on univariate analysis were included in multivariable model.
- 单因素分析中，其结果之间的差异并不能真实得反映出该因素对结局事件的效应，我们可以将单因素分析结果有统计学显著性的变量（$p<0.05$），作为候选变量的第一梯队。当然，我们也可以适当地将纳入标准放宽到$p<0.1$，或者$p<0.2$，甚至有的研究放宽到$p<0.25$，这样可以有效避免遗漏一些重要变量。

3. 考虑样本量大小决定最终纳入变量个数

- 如果样本量足够大，统计效能足够，我们可以借助统计软件提供的变量筛选方法自动筛选变量，这是一种让研究者赏心悦目的神操作，软件提供的变量筛选方法有六种：
  ① 条件参数估计似然比检验（向前：条件）；
  ② 最大偏似然估计的似然比检验（向前：LR）；
  ③ Wald卡方检验（向前：Wald）；
  ④ 条件参数估计似然比检验（向后：条件）；
  ⑤ 最大偏似然估计的似然比检验（向后：LR）；
  ⑥ Wald卡方检验（向后：Wald）。

3. 考虑样本量大小决定最终纳入变量个数

- 在多因素回归分析中纳入的变量并非越多越好，我们还要从模型的稳健程度去考虑。文中已经提到控制混杂因素的个数主要取决于发生结局事件的多少，控制的混杂因素越多，所需要的结局事件的例数就越多。
- 对于多重线性回归模型，样本量应至少为10-15的自变量个数，而对于Logistic回归和Cox回归，结局事件则应至少为15-20倍的自变量个数。
- 需要注意的是，这里指的是结局事件的数量，而不是总的样本量，总样本量当然还要远远多于结局事件的数量。

N Engl J Med. 2011 Jan 20;364(3):226-35

European Journal of Cardio-Thoracic Surgery 54 (2018) 4–9

BMJ 2012;345:e5278 doi: 10.1136/bmj.e5278

4. 其他常见变量筛选方法

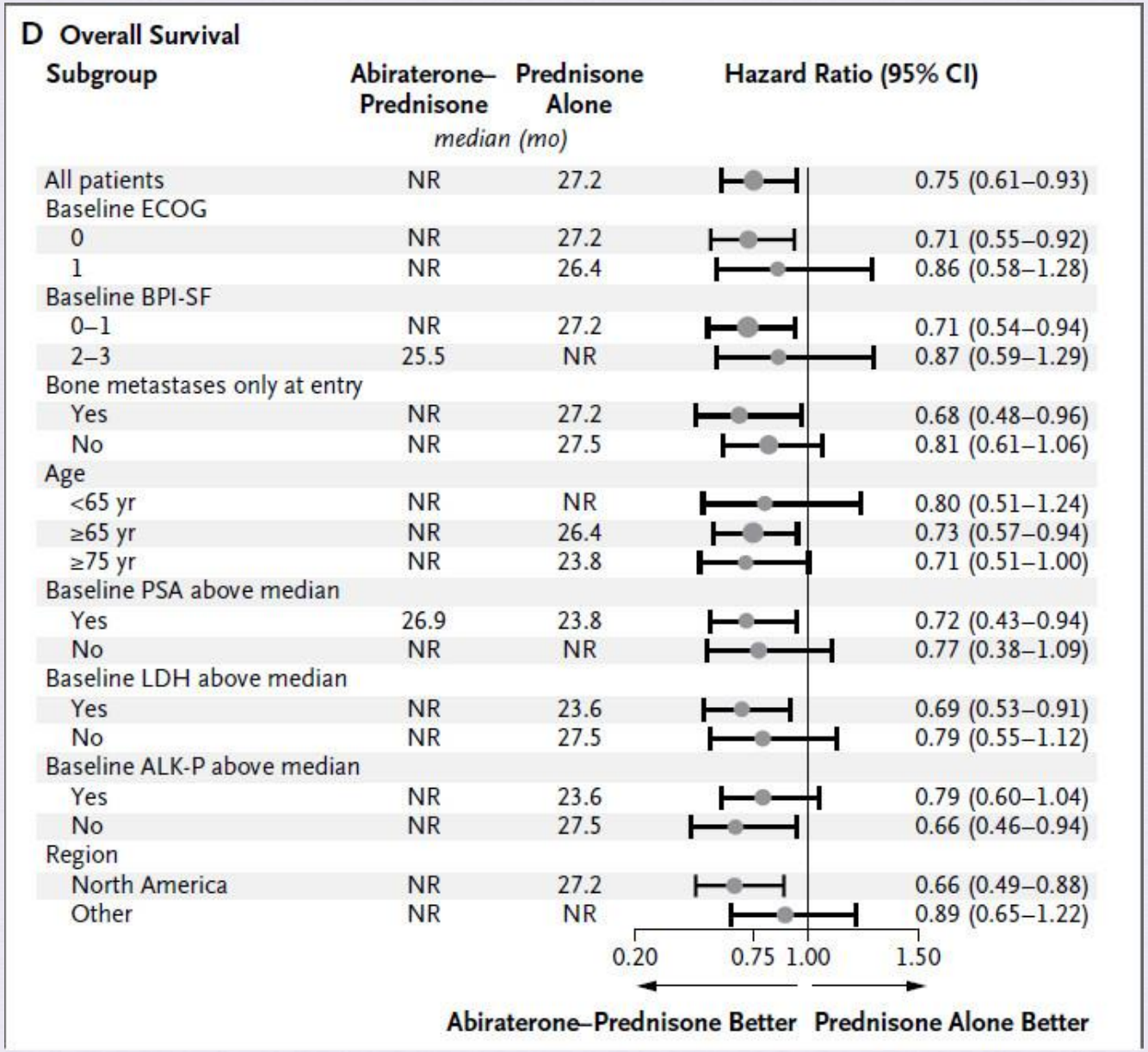
- 决定系数R^2，AIC，BIC，似然比对数、C-Statistics 等等。
- 纳入不同的变量，构建多个模型（mode1，model 2，model 3.....），客观报告每个模型的结果，这是一种敏感性分析。

CONTENT

亚组分析概念

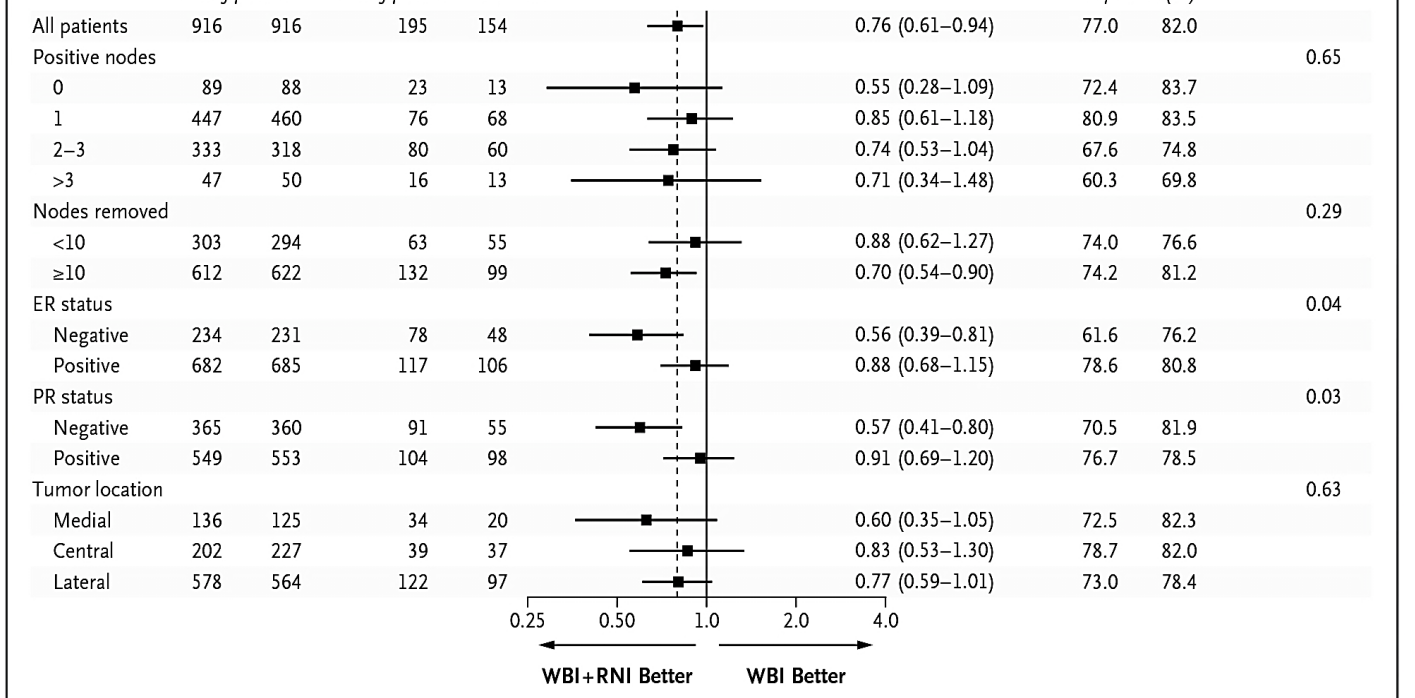
- 简言之，亚组分析即是在全数据集中按照某种分组因素把全数据集分为几个亚数据集，在每个亚数据集中分别比较实验组与对照组的试验效应差异。

亚组分析 案例1



N Engl J Med 2012.DOI:
10.1056/NEJMoa1209096

亚组分析 案例2



Subgroup	WBI	WBI+RNI	WBI	WBI+RNI	Hazard Ratio (95% CI)	WBI	WBI+RNI	P Value for Interaction
	no. of patients		*no. of patients with events*			*10-yr DFS (%)*		
All patients	916	916	195	154	0.76 (0.61–0.94)	77.0	82.0	
Positive nodes								0.65
0	89	88	23	13	0.55 (0.28–1.09)	72.4	83.7	
1	447	460	76	68	0.85 (0.61–1.18)	80.9	83.5	
2–3	333	318	80	60	0.74 (0.53–1.04)	67.6	74.8	
>3	47	50	16	13	0.71 (0.34–1.48)	60.3	69.8	
Nodes removed								0.29
<10	303	294	63	55	0.88 (0.62–1.27)	74.0	76.6	
≥10	612	622	132	99	0.70 (0.54–0.90)	74.2	81.2	
ER status								0.04
Negative	234	231	78	48	0.56 (0.39–0.81)	61.6	76.2	
Positive	682	685	117	106	0.88 (0.68–1.15)	78.6	80.8	
PR status								0.03
Negative	365	360	91	55	0.57 (0.41–0.80)	70.5	81.9	
Positive	549	553	104	98	0.91 (0.69–1.20)	76.7	78.5	
Tumor location								0.63
Medial	136	125	34	20	0.60 (0.35–1.05)	72.5	82.3	
Central	202	227	39	37	0.83 (0.53–1.30)	78.7	82.0	
Lateral	578	564	122	97	0.77 (0.59–1.01)	73.0	78.4	

Figure 2. Disease-free Survival at 10 Years, According to Subgroup.

Shown are hazard ratios and rates of disease-free survival among patients who underwent whole-breast irradiation plus regional nodal irradiation (WBI+RNI) and those who underwent whole-breast irradiation alone (WBI, control group). The dashed vertical line at 0.76 indicates the overall hazard ratio estimate. Hazard ratios are shown on a logarithmic scale. DFS denotes disease-free survival, ER estrogen receptor, and PR progesterone receptor.

N Engl J Med 2015;373:307-16.
DOI: 10.1056/NEJMoa1415340

亚组分析注意事项

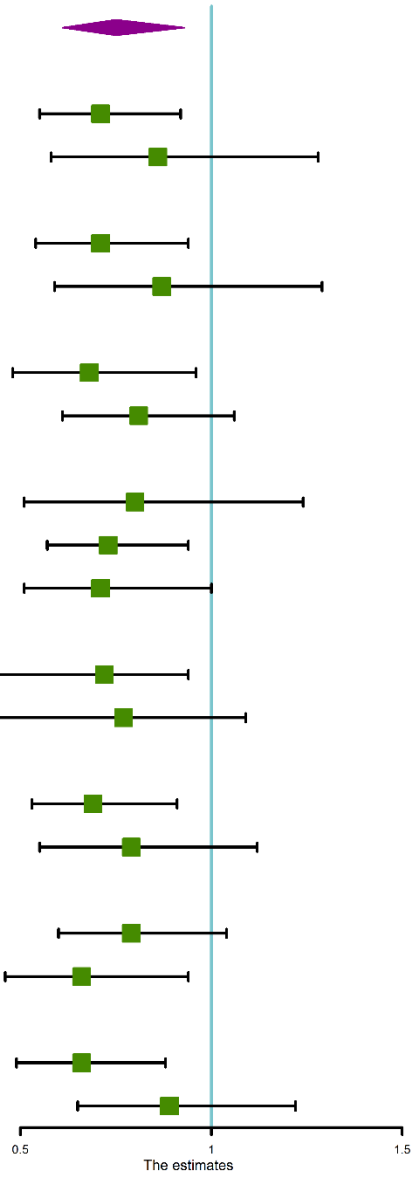
- 亚组分析是观察性分析
- 分组因素需要事先确定
- 由分层随机因素确定亚组数
- 亚组分析结果解读需要谨慎

观察性研究中是否可以有亚组分析?

亚组分析森林图绘制代码 案例1代码


```
install.packages("forestplot")
library(forestplot)
rs_forest <- read.csv('abiraterone.csv',header = FALSE)
# 读入数据的时候大家一定要把header设置成FALSE，保证第一行不被当作列名称。
tiff('Figure 1.tiff',height = 6000,width = 7000,res= 600)
forestplot(labeltext = as.matrix(rs_forest[,1:4]),
        #设置用于文本展示的列，此处我们用数据的前四列作为文本，在图中展示
        mean = rs_forest$V5, #设置均值
        lower = rs_forest$V6, #设置均值的lowlimits限
        upper = rs_forest$V7, #设置均值的uplimits限
        is.summary = c(T,T,T,F,F,T,F,F,T,F,F,T,F,F,F,T,F,F,T,F,F,T,F,F,T,F,F),
        #该参数接受一个逻辑向量，用于定义数据中的每一行是否是汇总值，若是，则在对应位置设置为TRUE，若否，则设置为FALSE；设置为TRUE的行则以粗体出现
        zero = 1, #设置参照值，此处我们展示的是HR值，故参照值是1，而不是0
        boxsize = 0.3, #设置点估计的方形大小
        lineheight = unit(8,'mm'),#设置图形中的行距
        colgap = unit(2,'mm'),#设置图形中的列间距
        lwd.zero = 2,#设置参考线的粗细
        lwd.ci = 2,#设置区间估计线的粗细
        col=fpColors(box='#458B00', summary= "#8B008B",lines = 'black',zero = '#7AC5CD'),
        #使用fpColors()函数定义图形元素的颜色，从左至右分别对应点估计方形，汇总值，区间估计线，参考线
        xlab="The estimates",#设置x轴标签
        lwd.xaxis=2,#设置X轴线的粗细
        lty.ci = "solid",
        graph.pos = 4)#设置森林图的位置，此处设置为4，则出现在第四列
```

Subgroup	Abirateron+Prednisone	Prednisone alone	Hazard Ratio(95%CI)
All patients	**NR**	**27.2**	**0.75(0.61-0.93)**
Baseline ECOG			
0	NR	27.2	0.71(0.55-0.92)
1	NR	26.4	0.86(0.58-1.28)
Baseline BPI-SF			
0~1	NR	27.2	0.71(0.54-0.94)
2~3	25.5	NR	0.87(0.59-1.29)
Bone metastases only at entry			
Yes	NR	27.2	0.68(0.48-0.96)
No	NR	27.5	0.81(0.61-1.06)
Age			
Less than 65 yr	NR	NR	0.8(0.51-1.24)
More than 65 yr	NR	26.4	0.73(0.57-0.94)
More than 75 yr	NR	23.8	0.71(0.51-1)
Baseline PSA above median			
Yes	26.9	23.8	0.72(0.43-0.94)
No	NR	NR	0.77(0.38-1.09)
Baseline LDH above median			
Yes	NR	23.6	0.69(0.53-0.91)
No	NR	27.5	0.79(0.55-1.12)
Baseline ALK-P above median			
Yes	NR	23.6	0.79(0.6-1.04)
No	NR	27.5	0.66(0.46-0.94)
Region			
North America	NR	27.2	0.66(0.49-0.88)
Other	NR	NR	0.89(0.65-1.22)



CONTENT

敏感性分析概念

- 改变统计分析条件，再次统计分析均称为敏感性分析
- 敏感性分析是一种验证研究结论是否稳健的分析策略

敏感性分析 案例1

Table 3. Sensitivity Analyses for Primary Outcome at 1 Year

Model	Effect	Aspirin			Clopidogrel-Aspirin			Hazard Ratio (95% CI)	*P* Value	P_{int} Value
		n	Patients With Event, n	Event Rate, %	n	Patients With Event, n	Event Rate, %			
Model 1*	Total	2586	450	17.4	2584	349	13.5	0.77 (0.67–0.89)	<0.001	0.25
	Subgroup 1†	2084	290	13.9	2129	222	10.4	0.73 (0.61–0.87)	<0.001	
	Subgroup 2‡	502	160	31.9	455	127	27.9	0.87 (0.68–1.12)	0.28	
Model 2§	Total	2586	362	14.0	2584	349	13.5	1.01 (0.86–1.19)	0.89	<0.001
	Subgroup 1†	2084	290	13.9	2129	222	10.4	0.73 (0.61–0.87)	<0.001	
	Subgroup 2‡	502	72	14.3	455	127	27.9	1.82 (1.34–2.47)	<0.001	
Model 3‖	Total	2586	450	17.4	2584	275	10.6	0.60 (0.51–0.70)	<0.001	<0.001
	Subgroup 1†	2084	290	13.9	2129	222	10.4	0.73 (0.61–0.87)	<0.001	
	Subgroup 2‡	502	160	31.9	455	53	11.7	0.38 (0.27–0.53)	<0.001	

CI indicates confidence interval.

*Model 1: all patients lost to follow-up were considered to have a stroke at last contact.

†Subgroup 1: patients with clopidogrel or aspirin treatment beyond month 3.

‡Subgroup 2: patients without clopidogrel or aspirin treatment beyond month 3.

§Model 2: only patients lost to follow-up in clopidogrel and aspirin group were considered to have a stroke at last contact.

‖Model 3: only patients lost to follow-up in aspirin group were considered to have a stroke at last contact.

Circulation. 2015;132:40-46. DOI: 10.1161/CIRCULATIONAHA.114.014791.

敏感性分析 案例2

丁香公开课 CLASS.DXY.CN

Table 2. Associations of Daily Coffee Consumption and All-Cause and Cause-Specific Mortality Among Men and Women

Variable	Coffee Consumption*					*P* Value for Trend	Per Cup Per Day
	Nonconsumers	Quartile 1 (Low)	Quartile 2 (Medium-Low)	Quartile 3 (Medium-High)	Quartile 4 (High)		
All-cause mortality							
Men							
Deaths, *n*	1039	4972	4440	4250	3601		–
HR (95% CI)							
Basic model†	1.00 (reference)	0.89 (0.83–0.95)	0.89 (0.83–0.95)	0.90 (0.84–0.96)	1.07 (0.99–1.15)	<0.001	–
Basic model plus smoking variables†	1.00 (reference)	0.88 (0.82–0.94)	0.83 (0.77–0.89)	0.78 (0.73–0.84)	0.83 (0.77–0.89)	<0.001	–
Multivariable model‡	1.00 (reference)	0.94 (0.87–1.00)	0.88 (0.82–0.95)	0.84 (0.78–0.90)	0.88 (0.82–0.95)	<0.001	0.97 (0.96–0.98)
Women							
Deaths, *n*	1817	6882	5236	5294	4162		–
HR (95% CI)							
Basic model†	1.00 (reference)	0.90 (0.85–0.95)	0.90 (0.85–0.95)	0.95 (0.90–1.01)	1.10 (1.04–1.16)	<0.001	–
Basic model plus smoking variables†	1.00 (reference)	0.91 (0.86–0.96)	0.87 (0.82–0.91)	0.87 (0.82–0.92)	0.90 (0.85–0.96)	0.004	–
Multivariable model‡	1.00 (reference)	0.94 (0.89–0.99)	0.90 (0.85–0.95)	0.90 (0.85–0.95)	0.93 (0.87–0.98)	0.009	0.99 (0.98–1.00)

Ann Intern Med. doi:10.7326/M16-2945

请在此处输入小标题

感谢观看

THANKS

丁香园特邀讲师 周支瑞