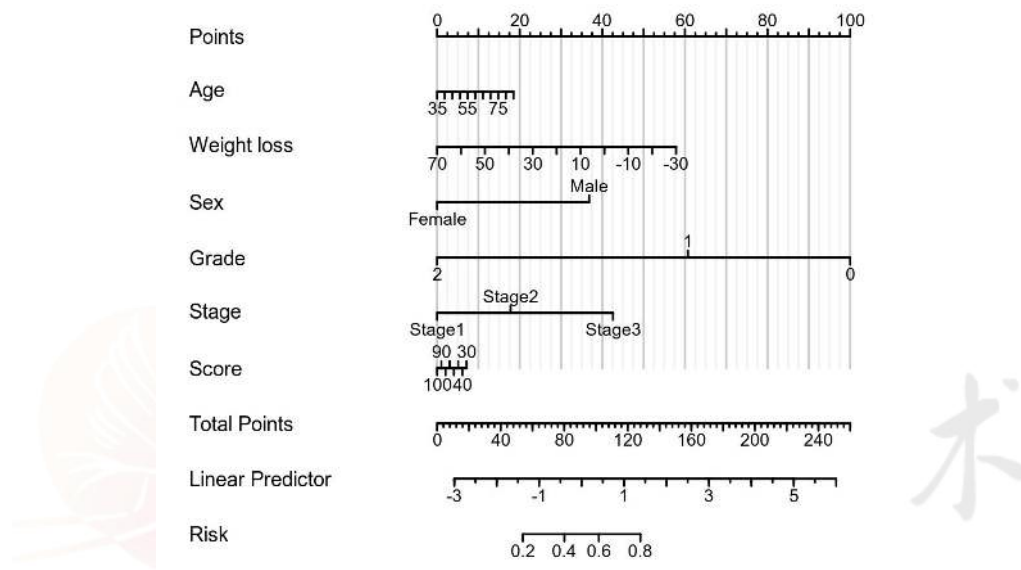


## 临床意义 - 诊断 Nomogram



网址: <https://www.xiantao.love>



更新时间: 2023.03.28

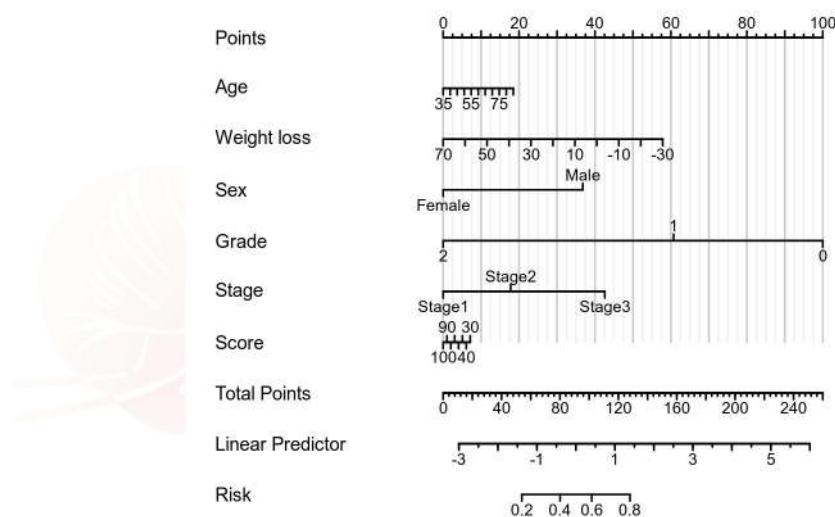
## 目录

基本概念 .....	3
应用场景 .....	3
分析流程 .....	4
结果解读 .....	5
数据格式 .....	7
参数说明 .....	10
数据处理 .....	10
置信区间 .....	10
坐标轴 .....	11
标题 .....	11
风格 .....	12
图片 .....	12
结果说明 .....	13
主要结果 .....	13
补充结果 .....	15
方法学 .....	18
如何引用 .....	19
常见问题 .....	20

## 基本概念

诊断 Nomogram 图：诺莫图或者列线图，是一种综合分析多个定量变量和定性变量以预测某特定事件发生的绘图法预测模型，模型基于 Logistic 回归模型，将其结果进行可视化的呈现。它根据模型回归系数的大小来制定评分标准，给每个自变量的每种取值赋值一个评分，对每个患者，就可计算得到一个总分，再通过得分与结局发生概率之间的转换函数来计算每个患者的结局时间发生的概率，其轴结构和风险点反映了各个变量对预测结果的影响和重要性。

### ➤ 图形构成



## 应用场景

列线图将复杂的回归方程，转变为了可视化的图形，使预测模型的结果更具有可读性，方便对患者进行评估。广泛应用在医学研究和临床实践中。

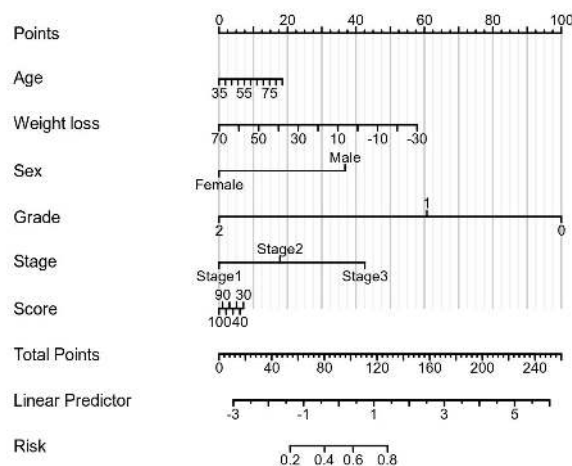
## 分析流程



注：

模型评价包括：似然比检验、C 指数评价模型的预测能力、模型拟合度评价

## 结果解读



诊断 Nomogram 图

- 图的左侧包含变量名称，右侧包含每个变量转换为分数的坐标，线段的长度则反映了该变量对结局事件的贡献大小
- Points 表示 0-100 的分数。表示模型中每个变量在不同分组/取值下所对应的单项评分情况，比如 Grade 分期中 1 组对应的单项评分为 61 分
- 预测变量行：
  - 如果变量是数值变量，则认为该变量相同数值距离下的评分差值是一样的
  - 如果变量是等级变量，则认为该变量不同等级之间的评分差值是不同的（如果是二分类变量，则这个变量以分类或者等级资料纳入的评分结果都是一样的）
- Total Points 总分对应的标尺，表示所有变量取值后对应的单项分数加起来合计的总得分
- Linear Predictor 表示线性预测值
- Risk 表示通过得分转换为结局发生的概率（XX 概率）

➤ 模型预测值表格可以用来后续的 ROC 分析

补充:

- 此图不用于评估模型的好坏, 评估模型的好坏是多个方面, 其中一个可以看一致性指数 (C-index) (在 0.5-1 之间, 越大越好), 在补充结果的模型评价中:

评价方向	评价内容	统计量	p值
模型检验	似然比检验	卡方值: 78.116	1.17e-13
区分度评价	C指数	C指数: 0.859 (0.807 - 0.912)	4.58e-42
校准度评价	拟合优度检验	卡方值: 5.6004	0.6919

1. 似然比检验: 若P值小于检验水准 ( $P < 0.05$ ), 表示本次拟合的模型中至少有一个变量的OR值有统计学意义, 即模型总体有意义

2. 模型的区分度能力采用C指数来评价, 一般而言, 0.51-0.7认为是一般的准确性, 0.71-0.9为中等的准确性, > 0.9为高度的准确性;

3. 校准度评价使用方法是Hosmer-Lemeshow Goodness of Fit 拟合优度检验, 一般而言, 若检验结果显示有统计学显著性 ( $P < 0.05$ ), 则表明模型预测值和实际观测值之间存在一定的差异。拟合校准度一般。如果  $P > 0.05$  说明预测值与实际值没有显著差异, 因此拟合校准度较好



## 数据格式

outcome	Age	Weight loss	Sex	Grade	Stage	Score
1	80		Male		0 Stage2	100
1	82	15	Male		0 Stage1	90
0	42	15	Male		2 Stage1	90
1	57	11	Male		0 Stage2	60
1	60	0	Male		2 Stage1	90
0	74	0	Male		2 Stage2	80
1	68	10	Female		0 Stage3	60
1	71	1	Female		2 Stage3	80
1	53	16	Male		1 Stage2	80
1	61	34	Male		0 Stage3	70
1	57	27	Male		1 Stage2	80
1	68	23	Female		1 Stage3	70
1	68	5	Female		0 Stage2	90
1	60	32	Male		0	70
1	57	60	Male		0 Stage2	70
1	67	15	Male		0 Stage2	90
1	70	-5	Male		1 Stage2	100
1	63	22	Male		2 Stage3	70
1	56	10	Female		0 Stage3	60
1	57		Male		0 Stage2	80
1	67	17	Male		0 Stage2	80

表格 1: 诊断数据

A	B	C
Sex	Stage	Grade
Male	Stage1	0
Female	Stage2	1
	Stage3	2
	Stage4	

表格 2: 分类变量的顺序

- 第一列因变量（**必须是二分类**），缺失值不能超过第一列长度的 85%。第 1 列中分类的前后出现的顺序会被参考的顺序，先出现的分类会被当做参考组。（影响 OR 值和置信区间计算以及图中变量中分类的顺序）
- 至少需要 2 列数据,最多不能超过 20 列,最少需要 20 行,最多不能超过 30000 行，样本量需要至少 4 倍以上变量数量，样本过少拟合模型效果相对较差。
- 第二列及以后为**预测的变量**，可以是数值类型，也可以是分类类型
  - 如果变量是数值变量，请以数值纳入，只要含有非数值（除空值）或者是无穷值外，则此列有可能没有办法纳入到分析

- 数值变量如果其分类个数  $< 5$  个（如 Grade 变量只有 0 1 2）则会按照等级变量来处理
- 如果变量是等级变量，建议以具体的名字纳入，比如上图中的 Stage，也可以（类似 Grade）以数字 0 1 2 的形式纳入，但是，如果以数字编码的形式纳入，如果种类超过 5 个，需要在 excel 的表 2 中设置等级参考顺序，否则该变量会以数值纳入（等级超过 8 个将没办法纳入）
- ◆ 等级变量在不同等级之间的 OR 是不同的，比如结果表格中的 Stage 变量，可以看到 Stage2 和 Stage1 与 Stage4 和 Stage3 之间的 OR 是不同的。尤其要注意不要随意对一个等级资料编码为 0 1 2 3，如果在上传数据进行了此类编码，则这个变量会被认为是数值变量而产生上述数值变量的效果而出现错误。如果是进行了数字编码的等级变量，比如图中 Grade 变量，假设我们设置了 Grade 变量的等级是 0 1 2，可以在表 2 中设定该变量的等级顺序
- 如果变量是分类变量，默认是以等级资料纳入。二分类变量以等级或者以分类资料或者数值纳入结果都是一样的。如果是多分类非等级资料，则需要以哑变量（暂不考虑）的形式纳入
- 数值变量

下方表格:（表 2-可以不提供）:

➤ 对应（表 1）预测变量（分类类型）中各分类的顺序

- 比如 Stage 想要设置 Stage1, Stage2, Stage3, Stage4 的顺序，就可以如上图设置。注意，设置了等级顺序后，多因素 Logistic 回归的结果都是以第一个作为参考，其他的等级顺序与第一个等级进行对比。另外，如果在表 1 中的分类变量没有设置等级顺序，则默认以在表 1 中各个分组出现的顺序作为等级顺序。此外，如果是以 0 1 2 编码的等级变量，如果没有在这个表中进行设置，则会以数值类型纳入（可见 Grade 列）




- 如果其取值跟表 1 预测变量完全一致，则会按照其顺序对上方对应的变量分类顺序进行分析。比如 Grade 变量在表 2 中各分类的顺序为 0、1、2，与表 1 的 Grade 变量中变量名还有具体值完全一样，则会按照表 2 变量法分类的顺序进行分析，如果不是则按照表 1 中变量分类的顺序进行分析。



## 参数说明

(说明：标注了颜色的为常用参数。)

## 数据处理



数据处理

缺失值处理 单因素后多因素

影响缺失处理是在单因素之前还是单因素之后。(多因素需要纳入变量均无缺失的样本)

- **缺失值处理**：默认是单因素后多因素前处理变量缺失，也可以选择单因素分析前统一删除缺失值

## 置信区间



置信区间

计算方法 Wald方法

- 计算方法：包含有：Wald 方法、profile 方法(MASS 包)、传统计算方法。其中，Wald 方法得到的置信区间是和 SPSS 是一致的，传统计算方法为

( $OR \pm 1.96 * SE$ )，传统计算方法对应原本生成置信区间的方式。建议选择 Wald 方法。

## 坐标轴

坐标轴

坐标轴与左侧文字之间距离
0.6

- 坐标轴与左侧文字之间距离：左侧文字与右侧坐标轴的距离，默认是 0.6

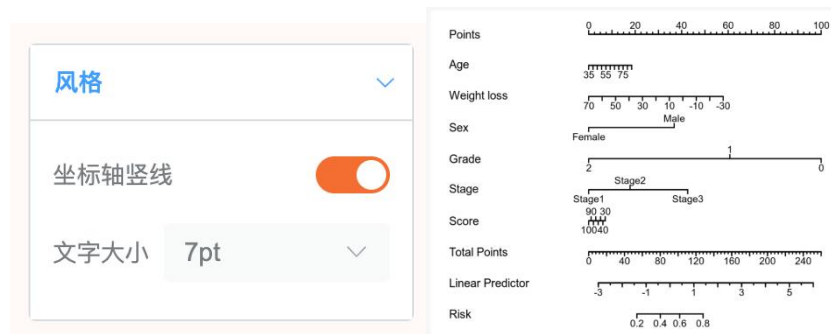
## 标题

标题

线性预测轴标题 Risk

- 线性预测轴标题：默认是 Risk

## 风格



- 坐标轴竖线：右边图展示的是不显示竖线的样子
- 文字大小：图中的文字部分的大小（包括标签文字和刻度数），默认是 7pt

## 图片

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm

## 结果说明

## 主要结果

主要结果 补充结果 方法学

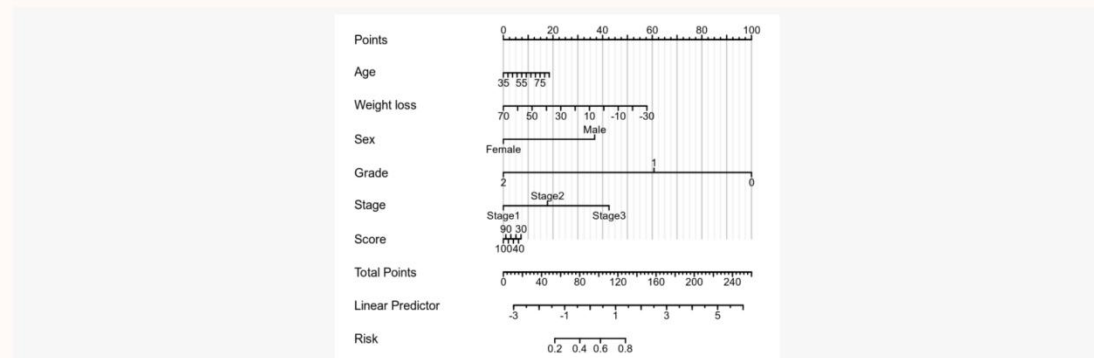
保存结果

下载整份报告

### 诊断Nomogram-二分类

诊断Nomogram: 建立在多因素回归分析的基础上, 采用带有刻度的线段将多个预测指标进行整合, 并按照一定的比例绘制在同一平面上, 从而用以表达预测模型中各个预测变量之间的相互关系

· 模型对应二分类结局: 1 vs. 0 (其中参考组: 0)[影响图中各个变量内分类的顺序]



诊断列线图.pdf

诊断列线图.tiff

诊断列线图.pptx

预测值-系数.xlsx

左侧标题文字:

· Points: 表示每个预测变量在不同取值下所对应的单项分数

左侧标题文字:

- Points: 表示每个预测变量在不同取值下所对应的单项分数
- (Variable): 表示模型中各个变量的取值和对应得分
- Total Points: 表示所有变量取值后对应的单项分数加起来合计的总得分
- Linear Predictor: 表示线性预测值

右侧坐标: 代表左侧标题文字的刻度与取值范围

主要结果格式为图片格式, 提供 PDF、TIFF、PPT 格式下载。另外还提供了模型预测系数表格的下载。

其中预测预测值-系数中提供了：

	A	B	C	D	E	F	G	H
1	outcome	Age	Sex	Grade	Stage	Score	linear_predictors	predicted_probability
2	0	42	Male	2	Stage1	90	-0.306265778	0.424026481
3	1	80	Male	0	Stage2	100	4.197987931	0.985196653
4	1	82	Male	0	Stage1	90	3.705226356	0.975995727
5	1	57	Male	0	Stage2	60	4.057732222	0.983005621
6	1	60	Male	2	Stage1	90	-0.141710681	0.464631499
7	0	74	Male	2	Stage2	80	0.532326659	0.630025605
8	1	68	Female	0	Stage3	60	3.933613065	0.980802914
9	1	74	Female	0	Stage2	80	0.532326659	0.630025605

前面的列内容对应纳入多因素的结果，linear\_predictors 代表多因素模型中每个样本对应线性预测值，对应的就是各个变量\*系数+常量后的值。predicted\_probability 代表多因素模型给每个样本预测得到的概率值。这些内容后续可以用于比如 ROC 模块中代表联合多个指标的情况。

		riskscore		多因素模型情况		+
	A	B				
1	多因素模型变	系数				
2	(Intercept)	3.113107027				
3	Age	0.00914195				
4	SexFemale	-1.25519315				
5	Grade1	-1.55359087				
6	Grade2	-3.64581414				
7	StageStage2	0.528547759				
8	StageStage3	1.559060308				
9	Score	-0.00175023				
10						

第 2 个表提供了多因素变量对应变量的系数情况（其中分类类型变量对应有除参考组以外其他分类的系数）

## 补充结果

### 1. 变量情况表：上传数据的一些基本情况和说明

变量情况					
各个变量识别出来的类型以及是否纳入进行分析					
变量	类型	分类数量	缺失数量	是否纳入分析	补充说明
outcome	分类变量	2	0	纳入	
Age	数值变量	-	0	纳入	
Weight loss	数值变量	-	14	纳入	
Sex	分类变量	2	0	纳入	
Grade	分类变量	3	0	纳入	
Stage	分类变量	3	1	纳入	
Score	数值变量	-	3	纳入	

总样本数: 228

- 如果某个分类变量的分类>10, 将无法识别为分类变量/等级变量
- 如果变量的分组是以 0 1 2此类进行编码, 如果分类数量<5, 会被识别为分类变量; 如果>5, 会被识别为数值变量
- 如果数据中含有无穷值, 无穷值会被当做缺失处理

补充说明: 单因素分析前, 会先去掉 结局列中的缺失的样本(结局缺失的样本是无法纳入进行分析的)

缺失处理策略: 单因素后多因素前处理变量缺失

### 2. 单因素 logistic 分析表格：包括变量及样本数，OR 值和 P 值

单因素Logistic					
变量	类型	数量	OR	置信区间	p值
Weight loss	数值变量	214	1.006	0.983 - 1.029	0.6088
Sex	等级变量	228			
Male		138	Reference		
Female		90	0.333	0.183 - 0.605	0.0003
Grade	等级变量	228			
0		40	Reference		
1		92	0.171	0.021 - 1.362	0.0953
2		96	0.024	0.003 - 0.179	0.0003
Stage	等级变量	227			
Stage1		63	Reference		
Stage2		113	1.859	0.970 - 3.560	0.0615
Stage3		51	5.270	1.961 - 14.163	0.0010
Score	数值变量	225	0.972	0.951 - 0.994	0.0112

表中所有变量都会纳入到多因素中

### 3. 多因素 logistic 分析表格：包括变量及样本数，OR 值和 P 值

#### 多因素 logistic

模型对应二分类结局(因变量): 1 vs. 0 (其中参考组: 0)

变量	系数 $\beta$	OR	置信区间	p值
Age	0.013846	1.014	0.972 - 1.058	0.5193
Weight loss	-0.021661	0.979	0.951 - 1.007	0.1386
Sex				
Male		Reference		
Female	-1.3773	0.252	0.116 - 0.550	0.0005
Grade				
0		Reference		
1	-1.4708	0.230	0.027 - 1.925	0.1751
2	-3.7444	0.024	0.003 - 0.191	0.0004
Stage				
Stage1		Reference		
Stage2	0.66601	1.946	0.831 - 4.559	0.1251
Stage3	1.5945	4.926	1.203 - 20.175	0.0266

#### 多因素 logistic.xlsx

模型常数/截距(Intercept): 3.1294

原始数据一共有228个, 变量信息缺失的样本有18个, 最终纳入的样本数: 210

备注: 如果出现纳入了多因素但是对应的统计量为空的情况, 说明(1)这个变量在去除变量信息缺失后某个分类数目过少(只有1个或者0个)或者是(2)存在严重共线性导致这个变量导致没办法计算。

模型常数/截距(Intercept): 3.1294

模型对应二分类结局(因变量): 1 vs. 0 (其中参考组: 0)

原始数据一共有228个, 变量信息缺失的样本有18个, 最终纳入的样本数: 210

备注: 如果出现纳入了多因素但是对应的统计量为空的情况, 说明(1)这个变量在去除变量信息缺失后某个分类数目过少(只有1个或者0个)或者是(2)存在严重共线性导致这个变量导致没办法计算。

备注: 当如果多因素中出现OR异常大或者异常小时(OR一般在0.33-3范围内), 说明这个变量的这个分类数量过少或者是存在共线性问题或者是数值类型变量变异比较小导致

(分类/等级)变量只要某个分类的p值满足阈值就会纳入到多因素中

△ 模型全局情况:

... AIC值: 194.729 (一般来说AIC越小, 模型拟合越好)

### 4. 方差膨胀因子

#### 方差膨胀因子(VIF)

方差膨胀因子可用于分析模型中的变量是否存在多重共线性问题

变量	类型	VIF
Weight loss	数值变量	1.1024
Sex	等级变量	
Male		Reference
Female		1.1259
Grade	等级变量	
0		Reference
1		7.4091
2		7.4974
Stage	等级变量	
Stage1		Reference
Stage2		1.3374
Stage3		1.7668
Score	数值变量	1.3094

一般认为, 当  $0 < VIF < 10$ , 不存在多重共线性(补充: 也有认为  $VIF > 4$  就存在多重共线性); 当  $10 \leq VIF < 100$ , 存在较强的多重共线性; 当  $VIF \geq 100$  或者是出现NaN, 多重共线性非常严重



## 5. 模型评价

### 模型评价

评价方向	评价内容	统计量	p值
模型检验	似然比检验	卡方值: 78.116	1.17e-13
区分度评价	C指数	C指数: 0.859 (0.807 - 0.912)	4.58e-42
校准度评价	拟合优度检验	卡方值: 5.6004	0.6919

1. 似然比检验: 若P值小于检验水准 ( $P < 0.05$ ), 表示本次拟合的模型中至少有一个变量的OR值有统计学意义, 即模型总体有意义
2. 模型的区分度能力采用C指数来评价, 一般而言, 0.51-0.7认为是一般的准确性, 0.71-0.9为中等的准确性,  $> 0.9$ 为高度的准确性;
3. 校准度评价使用方法是Hosmer-Lemeshow Goodness of Fit 拟合优度检验, 一般而言, 若检验结果显示有统计学显著性( $P < 0.05$ ), 则表明模型预测值和实际观测值之间存在一定的差异, 模型校准度一般; 如果 $P > 0.05$ , 说明预测值与观测值没有显著差异, 因此模型拟合度较好



## 方法学

统计分析和可视化均在 R 4.2.1 版本中进行

涉及的 R 包:

rms[6.4-0]用于模型拟合和计算每个变量的风险分数,

ResourceSelection[0.3-5]用于模型校准度衡量。



## 如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分 (Acknowledge) 致谢仙桃学术 ([www.xiantao love](http://www.xiantao love))。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。



## 常见问题

### 1. 为什么所有的变量都进行了单因素分析和多因素分析？

答：一般情况下，是通过对变量进行单因素分析，在对其结果进行筛选，选择单因素变量统计学  $p$  值大于 0.1（常用）作为筛选条件，满足则对这些变量进行多因素分析，不满足的这就不分析。但是不能避免有时候上传的数据所有变量都不满足（或条件太过于苛刻）导致无法分析，所以就不进行筛选，直接通过单因素和多因素分析进行计较就行。如果有需要筛选，需要自己剔除掉一些不想纳入的变量后再上传数据。

### 2. 为什么现在的 logistic 回归结果的 OR 值的置信区间和原来的不一样了？

答：

目前仙桃 logistic 相关的计算 OR 值的置信区间已经从原来的传统方法 `exp(summary(model)$coefficients[2,1]+1.96*summary(model)$coefficients[2,2])` 替换成了 Wald 方法来计算置信区间，并且提供了置信区间计算参数的方法（当前默认会选择 Wald 方法）（Wald 方法计算得到的结果和 SPSS 中 logistic 计算得到的置信区间是一致的），如果计算得到旧的置信区间，可以在置信区间参数中选择传统方法。

### 3. 如何更改结局二分类中的顺序（参考组）？

主要结果    补充结果    方法学

#### 诊断Nomogram-二分类

**诊断Nomogram:** 建立在多因素回归分析的基础上，采用带有刻度的线段将多个预测指标度量之间的相互关系

· 模型对应二分类结局: 1 vs. 0 (其中参考组: 0)[影响图中各个变量内分类的顺序]

答：

结局变量二分类的顺序按照数据中第 1 列中分类的出现顺序来，比如下面的数据先出现的是 0，然后是 1，所以参考组就是取的第 1 个出现的组。

	A	B	C	D	E	F	G	H
1	outcome	Age	Weight loss	Sex	Grade	Stage	Score	
2	0	42	15	Male	2	Stage1	90	
3	1	80		Male	0	Stage2	100	
4	1	82	15	Male	0	Stage1	90	
5	1	57	11	Male	0	Stage2	60	
6	1	60	0	Male	2	Stage1	90	
7	0	74	0	Male	2	Stage2	80	

如果要调整这个顺序，可以自己在上传数据里面把行换一下，然后再上传数据即可：

	A	B	C	D	E	F	G	H
1	outcome	Age	Weight loss	Sex	Grade	Stage	Score	
2	1	80		Male	0	Stage2	100	
3	0	42	15	Male	2	Stage1	90	
4	1	82	15	Male	0	Stage1	90	
5	1	57	11	Male	0	Stage2	60	
6	1	60	0	Male	2	Stage1	90	
7	0	74	0	Male	2	Stage2	80	