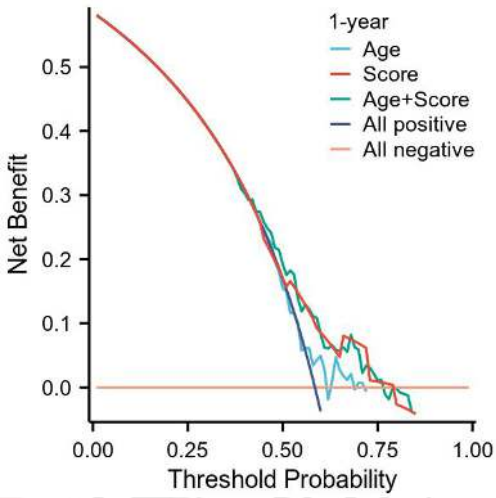
临床意义 - 预后 DCA



网址：https://www.xiantao.love

更新时间：2023.03.08

目录

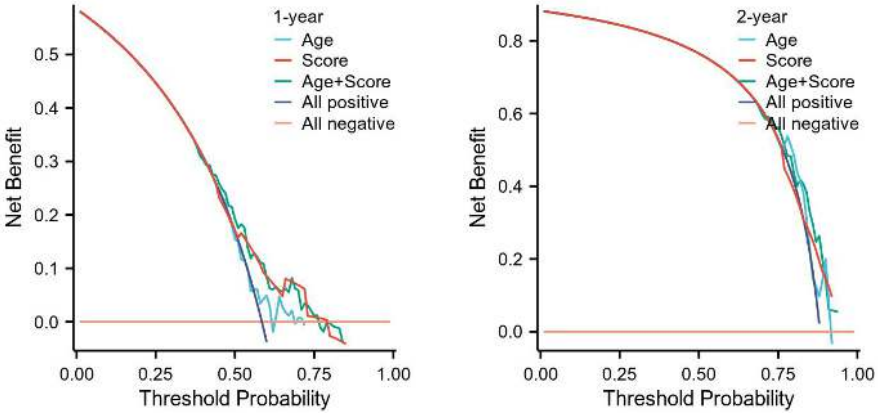
基本概念

- 决策曲线分析（decision curve analysis，DCA）：一种评估临床预测模型、 生物标志物的方法之一
- DCA 方法是纪念斯隆凯特琳癌症研究所的 AndrewVickers 博士等人提出 另一种评价方法，具体可看 https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/decision-curve-analysis
- 相对于 ROC 曲线只能通过敏感性、特异度来评估模型的好坏，只是考虑了准确度的问题，DCA 则是考虑了模型的临床效用或者患者获益方面。
- 这里的临床效用，指的是净收益。比如，要研究某个指标的预测效果，无论以这个指标具体哪个值作为阈值，均会产生假阳性和假阴性的问题，而真阳性和真阴性 对比假阳性和假阴性（经过一些转换和计算），最终产生的就是净收益，所以临床效用越大（净收益越高），也就是要尽可能减少假阳性和假阴性（人群的影响），增加对真阳性和真阴性（人群的影响），找到一个合理范围/区间的值作为阈值

应用场景

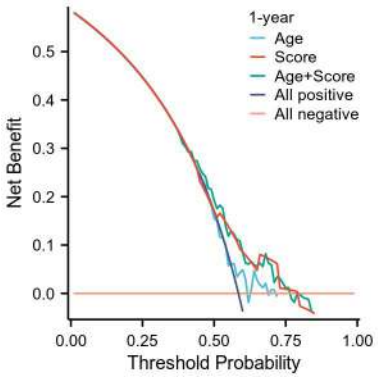
预后 DCA 图可用于评估构建预后模型或者变量在临床效用（患者获益）方面的作用（可放在预后模型的大图内）

主要结果



- 最简单对图结果好坏的判断，就是看模型的线在一定 x 值范围能稳定高于 all positive 的线和 all negative 的线
- 图中的 x 轴：代表概率阈值或者阈概率（Threshold Probability），y 轴代表净收益。阈概率可以理解成，在特定时间点（比如图中是指定了 2 年），当评价方法（模型）为某个值对应的风险作用到队列时整个队列会产生对应的死亡率情况，对死亡率设定一个阈值划分队列，这个阈值就是阈概率
  - 更简单的理解是在诊断类的模型上（不用考虑时间因素），比如一个指标的范围是 0-10，假定阈值设定为 3，则 ≥3 的人数÷总人数 就是对应的阈值概率，也就是阈概率
- 图中一般至少有 3 条线：
  - 1 条为整个队列都进入的情况（all positive, all）：指定时间点，整个队 列的死亡率随着阈概率的改变而改变（y = 队列特定时间点死亡率 × 阈值概率 /(1-阈值概率)），故起始点(x = 0.01)对应的几乎就是在特定时间点队列的死亡概率。（左图 1 年时候(x=0.01)队列死亡率是低于右图 3 年时候(x=0.01)队列死亡率）。一般这条线是作为极端情况（不作干预）的一条线，一般模型的线都会在某个 x 值开始会比这根线高
  - 1 条为整个队列都不纳入的情况（all negative, none），默认就是 0，无获益。这根也是一种极端情况的线

- 至少还有一条为模型在特定时间点下对应得到的概率 随不同阈概率划分下对应得到净获益（y 值）的情况。从图中可以看到，在 x 最开始的一段区域内，这条线几乎是和 all positive 重合的。在某个 x（阈概率）后，这条线开始比 all positive 高。这条线在特定阈概率区间与 all negative 线的差值即为净收益



- 当有多个模型的情况，那么就是看每个模型之间哪条能稳定高于另外一条，高于的部分（y 值差值）就是在特定 x 值获得的净收益。（相对来说，一个模型对应的 c-index 越高，对应的净收益也会相对高），比如图上的 Score 比 Age 在一定的 x 阈概率区间内的净收益高

数据格式

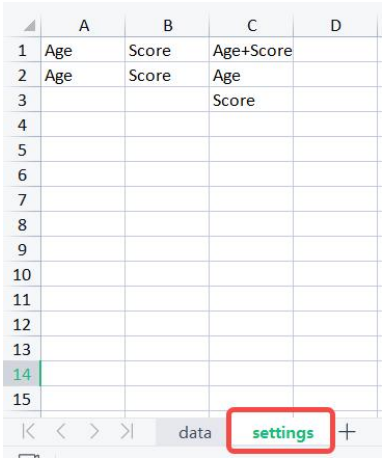
	A	B	C	D	E	F	G	H
1	event	time	Age	Weight los	Sex	Grade	Stage	Score
2	1	306	80		Male	0	Stage2	100
3	1	455	82	15	Male	0	Stage1	90
4	0	1010	42	15	Male	2	Stage1	90
5	1	210	57	11	Male	0	Stage2	60
6	1	883	60	0	Male	2	Stage1	90
7	0	1022	74	0	Male	2	Stage2	80
8	1	310	68	10	Female	0	Stage3	60
9	1	361	71	1	Female	2	Stage3	80
10	1	218	53	16	Male	1	Stage2	80
11	1	166	61	34	Male	0	Stage3	70
12	1	170	57	27	Male	1	Stage2	80
13	1	654	68	23	Female	1	Stage3	70
14	1	728	68	5	Female	0	Stage2	90
15	1	71	60	32	Male	0		70

data settings +

数据要求：excel 文件，(表 1-分析数据必须要提供)

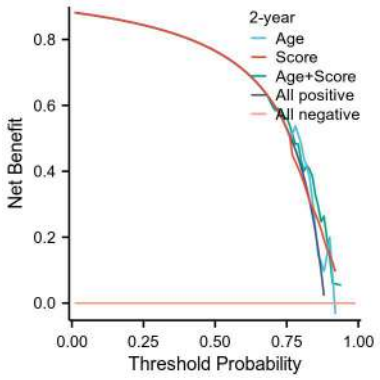
- 数据至少 3 列、20 行，最多支持 10 列（8 个预测变量）和 5000 行数据
  - 第 1 列是事件发生情况，用 0 和 1 表示，0 表示未发生事件，1 表示发生了事件。例如，事件可以定义为死亡，当受试发生了死亡，该受试的事件就定义为 1，当受试未发生死亡（删失），该受试的事件就定义为 0
  - 第 2 列是具体时间，必须以天作为单位，并且时间要长于 1 年以上
  - 第 3 列及以后为预测的变量，可以是数值类型，也可以是分类类型
    - 如果变量是数值变量，请以数值纳入。
    - 如果变量是等级变量，建议以具体的名字纳入，比如上图中的 Stage
    - 如果变量是多分类变量，可以划分成哑变量的形式纳入。可以合并一些类变成二分类变量变量纳入
    - 如果是能获得 Cox 模型的 riskscore 值，但是没有办法获得对应模型的各个变量的情况（比如利用云端数据），可以只给 riskscore，riskscore 可以代表这个模型的情况进行分析（结果和在表格中列出所有模型的变量重新构建 Cox 再分析的结果是一样的）

- 备注：二分类变量无论是否是等级，结果都是一致的



数据要求:（表 2-可以不提供）

- 每一列代表一个 Cox 模型，每一列的列名为模型的名字（会出现在 legend 部分），列的内容为第一个表的列名，这部分指定的列会被纳入作为这个模型的自变量
  - 比如图中指定了需要构建 3 个 Cox 模型，第一个模型是用第一个表的 Age 列构建一个 Cox 模型，模型的名字叫做 Age。第三个模型为用第一表的 Age 和 Score 列构建一个 Cox 名，模型的名字叫做 Age+Score。最终会得到下面的这个图的结果（指定了 2 年的时间）



当如果不提供第二个表，或者提供的第二个中的内容无法和第一个表的列名对应，则默认是以第一个表中所有的变量来构建一个 Cox 模型，以这个模型来分析 DCA 的内容。

备注：如果只能获得 Cox 模型的 riskscore 值，但是没有办法获得对应模型的各个变量的情况（比如利用云端数据），可以只给 riskscore，riskscore 可以代表这个模型的情况进行分析（结果和在表格中列出所有模型的变量重新构建 Cox 再分析的结果是一样的）。不同的模型可以分别以一列来展示，在第二个表里面指定每一列构建一个模型。

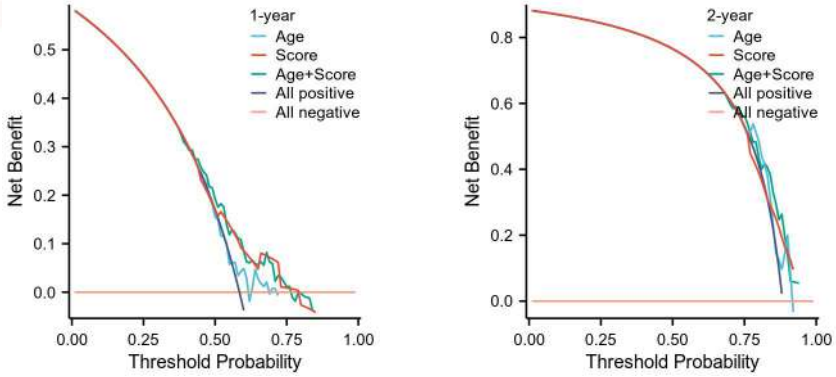
如果是 Lasso 预后模型，不建议纳入 riskscore，因为队列的的生存率和死亡率的拟合是以 Cox 的形式来拟合的，用 Lasso 预后模型的 riskscore 可能会有偏差

参数说明

（说明：标注了颜色的为常用参数。）

预测时间

- 时间 1：第 1 个时间点，数字，单位根据选择的预测时间单位
- 时间单位：可以选择上传数据预测时间列的单位，默认以年为单位，可以选择月、天为单位
  - 如下所示：左侧为预测时间为 1 年时，右侧为预测时间为 2 年时

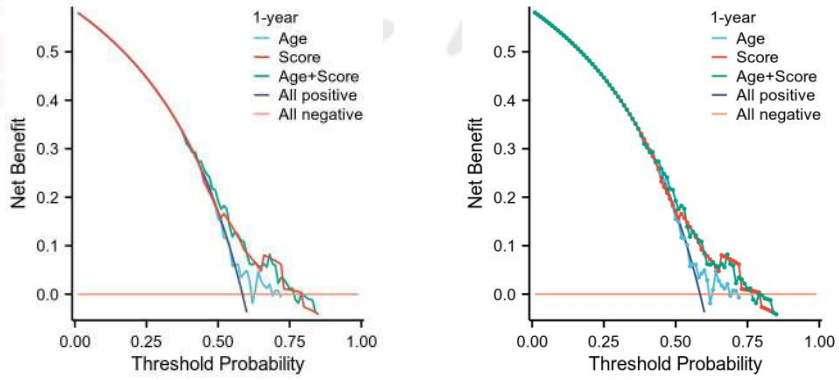


线

- 颜色：可以选择并修改图中线条，有多少条线就会取前多少的颜色。受配色方案全局性修改
- 样式：可以选择并修改图中线条的样式，默认为实线类型，还可以选择虚线
- 粗细：可以选择并修改图中线条的粗细，默认为 0.75pt
- 透明度：可以选择并修改图中线条的透明度。0 为完全透明，1 为完全不透明

点

- 展示：可以选择是否展示决策曲线各线条上的点，默认不展示，还可以选择展示，如下：左侧为不展示，右侧为展示



- 填充色：可以修改各点的填充色
- 描边颜色：可以修改各点的描边颜色
- 样式：可以选择并修改各点的形状/样式，默认为圆形，还可以选择正方形、菱形、三角形、倒三角形
- 大小：可以修改校准曲线上各点的大小，默认为 0.3
- 不透明度：可以修改各点的不透明度，默认为 1，表示完全不透明

标题文本

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

图注

- 是否展示：可以选择是否展示图注信息，默认展示（如左图），也可以不展示（如右图）
- 图注标题：可以修改图注对应的标题内容，如果需要换行可以在需要换行的位置插入“\n”
- 图注位置：可以修改图注的位置

坐标轴

- y 轴范围：可以控制 y 轴范围，需要提供 2 个值来控制范围。形如 0.1，0.3（最小值不能低于-0.5，最大值不能大于 1.5，如果调整过大可能会无作用）
- x 轴范围：可以控制 x 轴范围，需要提供 2 个值来控制范围。形如 0.1，0.3（最小值不能低于-0.2，最大值不能大于 1.2，如果调整过大可能会无作用）

风格

- 外框：是否添加外框，默认添加
- 网格：是否添加网格，默认不添加
- 文字大小：控制整体文字大小，默认为 7pt

图片

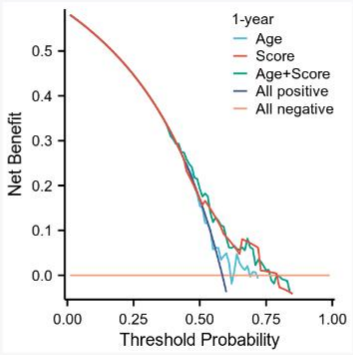
- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果:

预后DCA

预后DCA: 将决策曲线分析（decision curve analysis, DCA）运用到预后数据中评价模型的临床效用或者患者获益方面



预后DCA.pdf 预后DCA.tiff 预后DCA.pptx

· 图中的 x 轴代表 概率阈值或者阈概率（Threshold Probability），y 轴代表净收益

补充结果：变量情况统计表

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
event	数值变量	-	0	纳入	
time	数值变量	-	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	4	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 结局和时间列 中的缺失的样本(时间或者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

这里提供变量情况统计表：

- 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 结局和时间列 中的缺失的样本(时间或者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

补充结果：模型回归分析

Cox模型: Age

变量	类型	数量	系数β	HR	置信区间	p值
Age	数值变量	228	0.019543	1.020	1.002 - 1.037	0.0253

模型常数/截距(Intercept): -1.222

原始数据一共有228个, 变量信息缺失的样本有0个, 最终纳入的样本数: 228

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HR异常大或者异常小时，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导致

△ 模型全局性统计检验情况:

·· 一致性(Concordance, C-index): 0.550(0.525-0.575)

·· Likelihood ratio test= 5.13 on 1 df, p=0.0235

·· Wald test = 5.01 on 1 df, p=0.0253

·· Score (logrank) test = 5.02 on 1 df, p=0.025

Cox模型: Score

变量	类型	数量	系数β	HR	置信区间	p值
Score	数值变量	225	-0.01985	0.980	0.970 - 0.991	0.0003

模型常数/截距(Intercept): 1.5871

原始数据一共有228个, 变量信息缺失的样本有3个, 最终纳入的样本数: 225

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HR异常大或者异常小时，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导致

△ 模型全局性统计检验情况:

·· 一致性(Concordance, C-index): 0.607(0.583-0.632)

·· Likelihood ratio test= 12.47 on 1 df, p=0.000412

·· Wald test = 13.18 on 1 df, p=0.000282

·· Score (logrank) test = 13.23 on 1 df, p=0.000275

Cox模型: Age+Score

变量	类型	数量	系数β	HR	置信区间	p值
Age	数值变量	225	0.0144	1.015	0.997 - 1.032	0.1059
Score	数值变量	225	-0.018326	0.982	0.971 - 0.992	0.0008

模型常数/截距(Intercept): 0.56608

原始数据一共有228个, 变量信息缺失的样本有3个, 最终纳入的样本数: 225

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HR异常大或者异常小时，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导致

△ 模型全局性统计检验情况:

·· 一致性(Concordance, C-index): 0.615(0.590-0.640)

·· Likelihood ratio test= 15.13 on 2 df, p=0.000517

·· Wald test = 15.95 on 2 df, p=0.000344

·· Score (logrank) test = 16.01 on 2 df, p=0.000334

这里提供不同模型的 Cox 回归分析结果：

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：survival[3.3.1]，stdca.R

处理过程：

(1) 通过 survival 包拟合预后模型

(2) 使用 stdca.R 文件进行 DCA 分析

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）。

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. 只有 Cox 模型的 riskscore 值和预后资料，没有各个变量的情况，能否做 DCA?

答: 可以只提供 riskscore 和对应的预后资料（结局+时间），riskscore 可以代表这个模型的情况进行分析 (结果和在表格中列出所有模型的变量重新构建 Cox 再分析的结果是一样的）。不同的模型可以分别以一列来展示，在第二个表里面指定每一列构建一个模型