# 风险因素的识别和结局的预测大不同（二）

作者：曾琳 赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/1447.html



上次文章中提到风险因素识别和结局发生的预测最为常见的不同之处是，通常风险因素识别着重于发现“新”的风险因素，因此在研究设计阶段，往往采用匹配与研究组相似的对照组来进行比较，从而发现别的研究者没有发现的风险因素。而进行结局发生预测则主要是采用队列研究设计，在一个同质性较好的人群中构建模型来探讨结局发生的几率。同时以logistic回归的结果说明了如果要进行结局事件发生的预测除了关心各影响因素本身在模型中的权重外，模型中的常量也非常重要。今天让我们继续来聊聊风险因素识别和结局发生的预测的其它不同。

**风险因素识别和结局发生的预测另一大类不同是下结论时主要考察的指标是不同的**。

对于风险因素的识别，从统计的角度来说，判断某个我们关心的因素是否风险因素通常会基于假设检验的结论。也就是，通过单因素或多因素分析中出现了结局的人和没有出现结局的人这些因素分布差异有统计学意义来识别风险因素;或者通过比较暴露于这些因素的人和不暴露于这些因素的人结局事件发生频率的差异是否有统计学意义来进行判断。这时，统计学意义往往通过p值的大小来体现，一般会认为$p \leq 0.05$某个因素是结局发生的独立影响因素。总结下来就是，在进行风险因素识别的时候，下结论主要基于该因素假设检验中p值的大小。但是细心的小伙伴一定会说了，我们学流行病的时候老师给我们讲过病因推断的方法，除了统计学上有差异以外，应该还有其它条件。说得没错，在探索风险因素的时候，我们目前更倾向利用假设检验的结果，但是在确认某些因素是否风险因素时，则需要考虑因果推断中的其它条件了，比如去除或控制这个因素是否会影响事件的发生。显然为了验证这一点需要开展一个干预性研究。对因果推断感兴趣的小伙伴可以查找我们之前的文章，里面有相关的内容介绍哦。

再来说说结局发生的预测，显然在这个研究中我们更为关心的指标是预测准确性。也就是对一个结局未知的个体，如果我们收集到他/她的人口学、遗传、临床信息是否能对他/她的结局做出准确的判断?我想不用我说大家也知道了，与风险因素识别不同，在结局预测的统计模型构建中，我们希望尽量全的把对结局发生有贡献的因素都纳入，尽量提高预测的准确率。何为“准确”？其实就是上面说的那样，对于结局未知的个体，我们通过模型对其结局的判断与其实际情况越相符则准确度越高。说到这儿大家是不是觉得和诊断试验有莫名的相似呢?是的，对每个个体来说，我们既不希望误判也不希望漏判，所以对于Logistic预测模型的好坏，我们常常会像做诊断试验一样绘制ROC曲线来判断预测的准确性。不同的是，这时我们用于绘制ROC曲线所用的指标不再是临床上的某个检测结果了，而是用Logistic回归模型的预测概率。根据预测概率的大小判断事件是否发生，我们会得到一连串的灵敏度和特异度用于绘制ROC曲线，这样帮助我们来了解构建的预测模型是不是能准确预测事件的发生。

说到这儿不得不补充一下，这个准确性不仅仅指的是在构建预测模型的数据中的准确性，更强调

外部准确性。“ 窝里横 ”不是真英雄，是骡子是马总要拉出来溜溜。如果构建这个模型数据以外的个体，通过模型的预测也能得到较一致的结果，那么我们才能说预测模型的外部准确性好。也就是，模型的准确性需要进行外部验证。大家肯定会说外部验证的数据太难获得了，如果能获得更多的患者信息，我们肯定也会用于构建更稳定的预测模型啊，怎么会浪费作为外部数据呢?说得很有道理，不过在数据量有限的情况下，我们也可以做所谓的“ 外部验证 ”。请大家跟我一起默念：交叉验证大法好!没错，就是交叉验证，在我们获得的数据中抽取一部分作为验证样本，其它作为训练样本，训练样本用于构建模型，验证样本用于验证模型预测准确性。这个过程可以在我们的数据中反复运行多次，从而寻找到准确性最高的预测模型。关于交叉验证，我们公众号里也有相关的介绍，感兴趣的小伙伴可以通过检索复习一下哦。

进行到风险因素的识别和结局的预测大不同的第二篇了，我们这次主要强调了这两者用于下结论所参考的指标是不同的，风险因素的识别主要关心假设检验的结果，而结局的预测则更重视预测的准确性。同时还引导大家去复习因果推断和交叉验证的内容。