临床科研论文中统计报告注意事项：多因素模型和诊断试验

作者：杨超 鲁艺斐 李延龙 张耀文 来源：医咖会

本文原地址：https://www.iikx.com/news/statistics/11346.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！

临床科研论文中统计报告注意事项：多因素模型和诊断试验。2019年3月，European Urology杂志(IF 17.581)发表了泌尿外科临床研究领域的统计报告指南《Guidelines for Reporting of Statistics for Clinical Research in Urology》，目的在于提升人们的统计学知识，改善论文质量。



这篇文章关注
多因素模型和诊断试验中统计报告的注意事项，以及结论和解释部分要注意的问题。

1. 多因素模型和诊断试验.

1.1 多变量回归、倾向性评分和工具变量并不是一根魔棒.

有研究者认为多变量调整可以“消除混杂”、“使两组相似”或“模拟随机试验”。但这个说法有两个问题：首先，记录在数据集中的变量值通常是近似值，因此可能会掩盖组间差异。

例如，当使用临床分期作为协变量，比较不同治疗方案对局限性前列腺癌的预后时，如果一组临床分期为T2c 期患者多于另一组，那么该组的预后可能会更差。但是T2c 期也具有不同表型。多变量调整T2c 时，并不能保证T2c 的不同表型也完全相同。

其次，模型只针对少数测量的协变量进行调整，并不能排除未测量(甚至不可测量)的协变量存在重要作用的可能性。通常情况下，倾向性评分匹配比传统的多变量回归能更好地调整混杂因素。工具变量法的效果则主要取决于工具变量的质量。在许多情况下，工具变量与干预没有很强的关联，这导致了95% CI的大幅增加;而在某些情况下，则会低估治疗效果。

1.2 避免完全依赖“ 逐步选择 ”.

研究者通常根据单因素分析的结果来选择多因素模型中要纳入哪些变量;或者，先将所有变量都纳入到一个模型中，然后删除那些统计学意义不显著的变量。然而，这种完全依赖现有数据的变量选择方法在回归模型的建立过程中并不可取，可能会增加过度拟合的风险，并使许多统计量(如95% CI)受到高度质疑。逐步选择法应该限定于特定情况，比如在模型建立的初期，对哪些变量可能是预测变量知之甚少的时候。

1.3 在检验干预措施的效果时，避免报告协变量的OR值或HR值.

一个研究应该只回答一个科学问题。比如，研究者在探讨两种不同根治性前列腺切除术对疾病复发的影响，应当着重报告两种术式的OR值或HR值。尽管模型可能会调整诸如临床分期、分级以及前列腺特异抗原(PSA)等因素(协变量)，但报告这些因素(协变量)的OR值或HR值对主要研究问题反而没有帮助，反倒会干扰读者对研究主要结果的关注。

1.4 将连续变量分段，使结果的解释更有意义.

有的连续变量取值范围很大，作为影响因素预测结局时，结果无法解释。例如某研究显示，年龄每增加1岁时，癌症风险OR值为1.02(95% CI 1.01-1.02)。其中点估计值与95%CI上限均为1.02，结果难以解释，也没有临床意义。为了解决这一问题，我们可以将年龄以10岁为单位分段，这样模型得到的OR值，即年龄每增加10岁时，增加的癌症风险OR的值是多少。

1.5 避免根据效应大小对预测因素进行排序.

有的研究者在模型中会对预测因子进行排序，例如，声称“ 新的标志物是复发的最强预测因子 ”。多数情况下，这种排序是基于对OR值和HR值大小的比较，这是一种错误的思想。不同变量间的OR或HR值，会受到变量的编码方式，变量单位的选择的影响。此外，比较模型中分类变量和连续变量的系数，也是很困难的。

最后，报告分类预测因子的暴露比例也很重要，这一点从疾病防治的角度尤为重要。例如，去除OR值为2.0、占比为50%的危险因素，相较于去除OR值为3.5但占比为0.1%的危险因素，可以避免更多的疾病结局。因为第一个因素尽管OR值小，但是在总人群中占比高，所以如果去除第一个因素，那么在总人群中，改变疾病结局发生人数的绝对值要更多。

1.6 谨慎考虑模型的区分度和校准度.

将基于某个特定人群建立的模型应用于其他人群时，应当谨慎考虑模型的区分度。例如，一个预测勃起功能障碍的模型(包括年龄因素)在普通成年男性群体中比老年男性群体中区分度更高，因为普通成年男性群体年龄变异更大。因此，将该模型外推应用于老年男性人群中时，是否能够很好的区分勃起功能障碍，应当谨慎考虑。

校准度也是统计模型的一个重要组成部分。校准度的意义是模型给出的风险是否接近其真实风险。如果某个特定人群的数据中建立的一个模型，则该模型用于其他人群时，应当报告校准度，或者以校准图的方式展示校准度。

1.7 报告模型的临床价值.

除了灵敏度、特异度或ROC等统计量，作者应考虑报告模型的临床价值。例如，在接受活检的患者中测量某标记物，可以报告如果采用该标记物的某个水平作为cutoff值，来确定患者是否需要活检，照此执行将会进行多少活检，又有多少的癌症被发现和遗漏。

2. 结论和解释.

2.1 下结论时，不要只是重复结果.

例如，“ BMI与疾病预后之间存在统计学上的显著关联 ”这并不是一个结论，作者需要陈述结果对研究和临床实践的影响。结论部分需要指引未来的研究去探索BMI与疾病预后之间是否真的存在因果关系，或者建议对BMI较高的患者进行更为积极的治疗。

2.2 避免使用“ 可能 ”等字眼.

事实上，治疗方案“ 可能 ”会有效，这应该是研究前提出的基本假设。在结论中使用may或might等词语，就相当于陈述“ 我们在本研究结束时所知道的并不比开始时所知道的多 ”。

2.3 有统计学意义的p值并不意味着有临床意义.

一个小的p值仅意味着无效假设被拒绝了，并不意味着一定会对临床实践造成影响。例如，两种治疗方案之间差异具有统计学意义，不一定意味着哪种方法一定更加优越。作者需要仔细分析研究的临床意义来证明临床建议的合理性。

2.4 避免“ 小样本 ”、“ 回顾性分析 ”等伪局限性;考虑潜在偏差的来源及其对结果的影响机制.

作者通常将研究的局限性浅显地描为“ 我们的研究样本量小，或者基于回顾性分析 ”等。但如果研究结果明确，例如，某治疗方案或预测因子的OR值很大，样本量大小可能无关紧要。类似地，在具有长期随访的标记物研究中，回顾性的研究设计可能完全合适。

2.5 考虑缺失数据发生的可能性及其影响.

一项研究中，很难做到数据集完整没有缺失。例如一项纳入200例参与者的研究，可能有42例参与者缺失研究相关变量。又或者，许多研究仅包含一个相对较小的患者子集。例如，在500例接受手术治疗的患者中，仅纳入50例术前接受影像学检查的患者子集。在这两种情况下，有必要考虑和分析有缺失数据或者未入组患者，与入组患者之间的差异。

2.6. 对治疗方案的效果解释不应受亚组结局影响：预后较差的亚组可能仍然适合接受某种治疗.

有些研究者会比较同一种治疗方式在不同亚组中的结局，并错误的得出结论：对该种治疗方案预后较差的亚组应该接受别的治疗。这种错误解释将个体的差异性与治疗效果的差异性混为一谈。

例如，肿瘤组织块大的癌症患者术后复发率相较于肿瘤组织块小的患者更高，但是我们不能由此得出结论，肿瘤块大的患者不应该接受手术治疗。

2.7 注意因果推断：相关并不代表因果.

作者在下结论时往往会忽视这一点。比如，引言和方法部分可能会标明本研究目的是确定治疗频率和治疗反应之间是否相关，但下结论可能会说更频繁的治疗将提高反应率这种具有因果推断性的结论。

3.使用和解释p值.

目前p值已经被广泛滥用和误用。普遍的问题是p值在解读研究结果时被赋予更多的权重，更有学者主张完全禁止使用p值。我们在此特别强调，p值只是一个有助于解释研究的统计量，它不能最终决定我们对研究结果的解读。从临床研究中得到结论，需要仔细评估研究方法的优缺点、已发表文献中的其他相关数据结果、生物学合理性和效应大小等。

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转发