医学统计报告中要注意p值和置信区间

作者：医咖会 来源：医咖会

本文原地址：https://www.iikx.com/news/statistics/5754.html

医学统计报告中要注意p值和置信区间。2019年3月，EuropeanUrology杂志(IF=17.298)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。

这份统计报告指南对其他医学专科领域的研究同样具有借鉴意义，下面让我们一起浏览该指南的主要推荐意见，看看我们用的统计分析方法是否符合规范。


EUROPEAN UROLOGY 75 (2019) 358–367

available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Opinion – Editor's Choice

Guidelines for Reporting of Statistics for Clinical Research in Urology

1. 统计推断和p值

1.1 不要写接受无效假设

在统计检验中，无效假设只能被拒绝或不被拒绝。如果p≥0.05，研究者应避免得出诸如“药物无效”、“组间无差异”或“反应率未受影响”等结论。相反，应使用“我们没有看到药物作用的证据”、“我们无法证明两组之间的差异”、或“反应率的差异没有统计学意义”。

1.2 p值略高于0.05，不是一种“趋势”

对于p=0.07这种情况，避免说“有达到统计学差异的趋势”，或“接近统计显著性”，因为p值不是在移动的。可以说，尽管我们看到一些证据表明接受新手术患者的反应率有所改善，但两组

间的差异并未达到传统的统计学显著性水平。

1.3 p值和95% CI不能量化假设的概率

p=0.03，并不意味着结果是由偶然机遇导致的可能性是3%;同样，95% CI也不应被解释为真实参数值在95% CI范围内的可能性为95%。p值的正确解释为当原假设为真时所得到的样本观察结果或更极端结果出现的概率;而95% CI，意味着如果用同样的步骤去选样本，那么100次这样的独立过程，有95%的概率计算出来的区间会包含真实参数值。

1.4 不要使用置信区间来检验假设

当OR值的95% CI不包含1时，研究者可能会称组间差异具有统计学意义，这其实是有问题的：因为置信区间与估计值有关，与推断无关;而且，计算置信区间的数学方法可能不同于计算p值的方法。即使p<0.05，95% CI也可能不包括两组之间的差异在内，反之也是如此。例如，一项纳入100例患者的研究，两组事件的发生率为70%和50%，采用Fisher精确检验计算的p值为0.066，而OR值的95% CI却为1.03-5.26。

1.5 报告多个p值时，需要注意合理解读结果

当你对5个独立的无效假设报告p值时，那么至少有一个错误拒绝无效假设的概率不是5%，而是大于20%。在某些特定的情况下，如基因组学研究，对p值进行调整是合适的;更常见的方法是在多重检验时对p值进行简单解释。

1.6 善于使用交互项

对于一个假设而分别进行检验的错误，常常发生在干预被证明对一个亚组有效而其他组别无效时。更合适的方法是在统计模型中使用交互项。例如，为了确定一种药物是否在女性比在男性身上更能减轻疼痛，可以建立模型如下：

最终疼痛评分= β 0+ β 1(基线疼痛评分)+ β 2(药物)+ β 3(性别)+ β 4(药物)×(性别)

2. 报告研究估计值

2.1 使用适当的精确级别

研究者应该仔细考虑报告的每一个数字，而不是简单地从统计软件中复制和粘贴出来。当然，根据报告数值类型的不同，小数点精确的位数也是有差异的。

2.2 避免描述中的冗余统计

对于描述性统计分析结果，研究者应该适当取舍。例如，没有必要说男性占40%，女性占60%，二者取一即可。

2.3 报告主要研究问题的估计值

一项临床研究通常聚焦于少数几个科学问题上，研究者通常应对每个问题提供估计值。例如，在

两组比较时，应该提供两组差异大小的估计值，避免仅单独给出每组的数据，或者简单地说差异有或无统计学意义。在对预后因素的研究中，应给出预后因素的影响强度大小，如OR值或HR值，并且报告p值。

2.4 报告主要估计值的置信区间

作者应报告与主要研究问题有关的估计值的95%CI。例如，在比较两种手术方法的研究中，作者可能会报告10%和15%的不良事件率;然而，这个研究关键是想看两组之间的差异，因此，差异大小5%还应给出95%CI(比如1%-9%)。对于平均年龄、性别比等统计量则没必要给出置信区间。

2.5 不要把分类变量视为连续变量

像Gleason分级的变量得分为1-5分，但是3分和4分之间的差异并不是2分和4分之间差异的一半。因此，Gleason分级这个变量应该以百分比的形式来报告(如第1级占40%)，而不是当成连续变量。同样地，在多因素回归模型中，Gleason分级也应该当成多分类变量放入模型。

2.6 如果没有令人信服的理由，避免将连续变量进行分类

对于年龄这类变量，比较常见的做法是根据年龄大小将患者分组(如老年人定义为年龄≥60岁)，然后将年龄作为分类变量进行分析。在流行病学研究中，将变量按照四分位数进行分组，报告各组与对照组相比的HR值也比较常见。

然而，这也可能带来问题，因为我们假设了每个类别中变量的所有值都是相同的。一般来说，最好将原本的变量保持连续变量的形式，同时也可以适当进行非线性的转换。

2.7 连续型预测因子与结局之间的关系可以用图片来说明，尤其是建立非线性模型

在研究年龄和并发症发生率的研究中，研究者可以分别在X轴和Y轴上绘制年龄和并发症的发生风险，并显示带有95% CI的回归线。非线性模型通常也很有用，因为它并没有假设一个线性关系，可以允许研究者确定是否风险在某个年龄以后开始不成比例地增加。

2.8 不要忽视meta分析中的异质性

通俗来讲，meta分析中异质性检验的目的是检查各个独立研究的结果是否具有可合并性。如果存在异质性，不仅需要报告p值，而且要关注随机效应的估计值。研究者应调查异质性的来源，并确定导致研究结果差异的因素。

2.9 对于生存分析，报告终点事件数，而不是比例

举例来说，“ 60名患者中，10人(17%)死亡 ”。由于患者在不同的时间进入研究，并且随访的时间段不同，因此报告17%的比例没有意义。对于生存分析来说，标准的统计方法是计算生存概率，例如报告5年内死亡风险为60%，或者中位生存时间为52个月。

2.10 对于生存分析，报告未发生终点事件患者的中位随访时间，或者给定时间内未发生终点事件的患者数

以1970年到2010年治疗的1000名儿童癌症患者队列数据为例，如果治愈率仅为40%，所有患者的中位随访时间可能仅有几年;然而，存活患者的中位随访可能为40年，后面这个数据可能对于了解队列的随访时长更有帮助。假设在2009年，又有2000名患者加入了研究。幸存者的中位随访时间为一年左右，这又是一个误导。同时，我们也可以这样报告：“ 至少35年来，312名患者没有发生任何终点事件 ”。

2.11 对于生存分析，确保所有预测因子在零时已知，或者考虑界标(Landmark)分析或时间依赖协变量等方法

许多情况下，感兴趣的变量会随时间发生变化。比如，当我们想看看PSA速度是否可以预测前列腺癌患者在积极监测下的疾病进展时间。问题是PSA在诊断后的不同时间点进行检测的，研究者很可能会用距离诊断的时间放入Kaplan-
Meier或Cox回归模型中，而不是使用根据随访时间计算出来的PSA速度。

通常有两种方法来解决这个问题：界标分析可用于当感兴趣的变量在短而明确的时间段内已知时(如辅助治疗或化疗反应)。简言之，研究者在一个固定的“ 界标 ”开始计时(如手术后6个月)。或者，研究者也可以采用时间依赖变量的方法：每当有关于变量的新信息出现时，将“ 重置时间 ”。这是目前最常用于PSA速度和进展研究的方法。

文献来源：Assel M, Sjoberg D, Elders A,et al. Guidelines for Reporting of Statistics for Clinical Research in Urology. Eur Urol, 2019, 75(3): 358-367.