Newcombe-Wilson 得分方法在有效率为 100% 非劣效临床试验中的应用

唐欣然，黄耀华，王杨，李卫

【摘要】 **目的**　探讨试验组和对照组实际有效率为 100% 的非劣效临床试验的设计和推断方法。**方法**　分别介绍获得率差的传统正态近似方法和 Newcombe-Wilson 得分方法的基本理论，再通过实例计算两组率为 100% 时的率差及合理的试验设计。**结果**　Newcombe-Wilson 得分方法解决了两组有效率都为 100% 率差置信区间的计算问题；实例采用传统近似正态进行试验设计，保守估计成功率为 98%，非劣效界值取 10% 时，每组样本为 33 例，当实际成功率都为 100%，两组率差点估计和 95% 置信区间估计为 0.0%（-10.4%，10.4%），试验失败。**结论**　Newcombe-Wilson 得分方法能够计算非劣效临床试验中试验组和对照组有效率都为 100% 率差的置信区间；在高成功率非劣效试验设计中还应考虑 Newcombe-Wilson 得分方法进行试验设计。

【关键词】 临床试验；效率；流行病学方法

【中图分类号】R446.1；R181　　【文献标识码】A　　【文章编号】1674-3679(2015)02-0190-03

DOI:10.16462/j.cnki.zhjbkz.2015.02.022

Application of Newcombe-Wilson score method in non-inferiority clinical trials with full rates of success

TANG Xin-ran, HUANG Yao-hua, WANG Yang, LI Wei. *State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing* 100037, *China*

【**Abstract**】 **Objective**　To explore methods of design and evaluation in non-inferiority clinical trials with success rates of 100% in both treatment and control groups. **Methods**　Firstly, basic principal of Normal approximation interval and Wilson score methods were introduced. Then an example about calculation of proportion difference in non-inferiority clinical trials with success rates of 100% was showed to illustrate these two methods. **Results**　Wilson score solved the issues of calculating intervals in trials with success rates of 100%; In the given example, the success rates were conservatively accessed as 98%, the non-inferiority margin was 10%, when using traditional approximately normal method, 33 subjects would be required in each group. When the success rates of both two groups were 100%, the point estimate and 95% confidence interval of rate difference were 0.0% (-10.4%, 10.4%). The trial failed. **Conclusions**　Intervals of proportions difference could be calculated using Wilson score method in clinical trials with success rate of 100%. In designing the non-inferiority clinical trials with probable success rate of 100%, Wilson score design method should be considered besides the traditional Normal approximation method.

【**Key words**】 Clinical Trials; Efficiency; Epidemiologic methods

(*Chin J Dis Control Prev* 2015, **19**(2): 190-192)

在医疗器械临床试验中，某些产品（如骨科的骨钉、骨板）常常由于技术相对成熟，成功率和有效率往往能达到 100%。这类产品的研究设计如果为非劣效随机对照试验，研究者往往会保守地估计产品的成功率，如 98%，99%，并且结合临床上认可的非劣效界值，计算所需要的样本量。但如果临床试验最终的结果未出现无效的病例，即本次试验中试验组和对照组成功率都为 100%，传统的 CMH（cochran-mantal haenszel）卡方因无法估计两组率差的标准误而无法计算率差的置信区间，进而无法判断率差置信区间与预先设定的非劣效界值的关系[1,2]。因此，现阶段急需一种评价方法用于计算两组有效率都为 100% 的率差[3]。

1　对象与方法

1.1　研究对象　非劣效（non-inferiority）临床试验，是指通过与阳性对照的比较评价试验产品的有效性和安全性[4]，已广泛应用于药物和医疗器械临床试

【作者单位】中国医学科学院，北京协和医学院，国家心血管病中心，阜外心血管病医院，心血管疾病国家重点实验室，医学统计部，北京　100037

【作者简介】唐欣然（1988-），女，河北沧州人，实习研究员，学士。主要研究方向：生物统计与流行病学。

【通讯作者】李卫，E-mail：liwei@mrbc-nccd.com

验中。对于定性指标，其统计思想可以理解为试验组和对照组率差和设定的非劣效界值的比较研究。假定某非劣效临床试验的主要终点为越大越好的高优指标，比如成功率，则其假设为 H_0: P_T-$P_C \leqslant \triangle$，H_1: $P_T - P_C > \triangle$，P_T为试验组的成功率，Pc 为对照组的成功率，△为预先制定的非劣效界值，即试验组相对于对照组在临床可接受范围内的差值，由临床专家和统计专家共同决定[5]。

1.2 传统正态近似方法 在试验的设计阶段，非劣效临床试验往往采用正态近似方法估算样本量，其计算的公式[2,4,6]为：

$$n = \frac{[z_{1-\alpha/2}\sqrt{2\bar{P}(1-\bar{P})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}]^2}{[\triangle - (p_1-p_2)]^2}$$，$\bar{p}$为两组有效率的平均值，△为非劣效界值。

在试验的评价阶段，正态近似方法也常用于估算组间的率差，其公式为 $D \pm Z_{1-\alpha/2}SE(D)$，$z$ 表示标准正态分布的分位数，α 为检验水准，常取双侧 5%，$SE(D)$ 为两组合并的标准误，计算公式为 $SE(D) = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$。SAS 程序中的 CMH 卡方能直接输出两组率差的点估计和置信区间，然而在两组率都为 100% 情况下，正态近似方法将无法估算两组有效率的合并方差，从而无法获得两组率差的置信区间。

1.3 Newcombe-Wilson 得分法 Newcombe-Wilson 得分方法由 Newcombe 在 1998 年率先提出，在比较 11 种率差的计算方法之后，Newcombe 从保守性和可行性方面推荐使用 Wilson 得分方法计算率差的置信区间[1,3,7]。目前该方法已被 FDA 指南推荐使用，作为事件率可能为 0% 或 100% 率差置信区间计算方法的首选[1]。其计算方法是通过 Wilson 得分单样本计算公式得到两单样本率的可信区间上下限[3]。利用渐进方差 $p(1-p)/n$ 求解，令 $(z^2_{1-\alpha/2}+n)\pi^2 - (z^2_{1-\alpha/2}+2np)\pi + np^2 = 0$，其中 π 的两解分别为置信区间的上下限，$n$ 表示单组的样本量，p 表示单组的有效率。其置信区间的计算公式为 $[2np + z^2 \pm z\sqrt{(z^2+4npq)}]/2(n+z^2)$，$q$ 为 $(1-p)$。在分别计算两组有效率的置信区间后，再通过公式 $D_l = D - \sqrt{(p_1-l_1)^2+(u_2-l_2)^2}$，$D_u = D + \sqrt{(p_1-l_1)^2+(u_2-l_2)^2}$分别计算率差的可信区间上下限，其中 l_2，u_1，l_2，u_2 分别为两组率 Newcombe-Wilson 得分方法计算得到的置信区间上下限。

连续校正 Newcombe-Wilson 得分方法在所有计算率差方法中为最保守的一种方法，率差计算公式同 Newcombe-Wilson 得分方法，区别在于单组置信区间的计算公式有所调整，下限的计算公式[1,3]为 $[2np + z^2 - 1 - z\sqrt{(z^2-2-1/n+4p(qn-1))}]/(2n+z^2)$，上限的计算公式 $[2np + z^2 + 1 + z\sqrt{(z^2+2-1/n+4p(qn-1))}]/2(n+z^2)$。连续校正方法因单组率计算的调整而增加可信区间的宽度，从而更加保守地估计组间差异。

Newcombe-Wilson 得分方法和连续校正的 Newcombe-Wilson 得分方法计算过程相对简单，通过简单的 SAS 或 Excel 编程即可实现，非常易于使用。

2 实例分析

某项用于骨折患者的骨板临床试验为一非劣效试验，欲证明试验产品非劣于对照产品。在前期的预试验中，产品的有效率为 100%。在本次试验设计中，为保守估计产品的疗效，假设试验组和对照组应用于骨折患者术后 6 个月的有效率为 98%，临床上认可的非劣效界值为 10%，通过正态近似的样本量计算公式获得每组至少 31 例受试者，并考虑 6 个月随访时 5% 脱落率，则每组入选 33 例，两组共入选 66 例受试者进入试验[8,9]。

试验的最终结果为试验组和对照组在 6 个月内没有发生失访，所有病人按照规定随访时间完成随访，最终纳入 FAS(full analysis set)、PPS(per-protocol set) 受试者人数都为试验组 33 例，对照组 33 例。FAS、PPS 数据集中，试验组有效率为 100%，对照组有效率为 100%，根据 Newcombe-Wilson 得分方法计算两组有效率差值的点估计和置信区间估计，为 0.0%(-10.4%，10.4%)。通过与预先设定的非劣效界值 -10% 相比，置信区间的下限已经超出非劣效界值，试验失败。其原因是试验设计阶段采用的是近似正态的样本量估计方法，而评价阶段由于近似正态方法无法计算率差的置信区间而采用的是 Newcombe-Wilson 得分方法，设计和评价的方法不一致，导致即使所有的病例都成功，试验却失败的结果。

因此在试验设计阶段，当产品的有效率可能为 100% 时，研究者不仅要保守估计产品的有效率，还要考虑产品有效率一旦达到 100% 的试验设计。因此上述例子中，除了常规计算样本量 31 例之外，还需设计有效率为 100% 的情况，通过 Newcombe-Wilson 得分方法进行研究设计。即假设产品最终有效率为 100%，非劣效界值取 10%，通过 Newcombe-Wilson 得分方法反推的样本量计算公式求得每组至少 35 例，考虑 5% 脱落，每组 37 例，两组共需 74 例受试者。取两次样本量最大的结果作为最终的样

本量，即最终需要 74 例受试者才能证明非劣效的研究假设。表 1 列出了有效率为 100% 不同非劣效界值下 Newcombe-Wilson 得分方法设计对应的每组样本量[10]。

表 1　不同非劣效界值下 Newcombe-Wilson 得分方法设计对应的每组样本量

Table 1　Sample sizes in each group using Newcombe-Wilson Score Method with different Non-inferiority Margins

非劣效界值(%)	每组样本量(NO CC)	每组样本量(CC)
20.0	16	21
19.0	17	22
18.0	18	23
17.0	19	25
16.0	21	27
15.0	22	29
14.0	24	31
13.0	26	33
12.0	29	37
11.0	32	40
10.0	35	45
9.5	37	47
9.0	39	50
8.5	42	53
8.0	45	56
7.5	48	60
7.0	52	65
6.5	56	70
6.0	61	76
5.5	67	84
5.0	73	92
4.5	82	103
4.0	93	116
3.5	106	133
3.0	125	156

注：NO CC 非连续校正；CC：连续校正。

3　讨论

以率为评价指标的非劣效临床试验已广泛应用于药物和医疗器械临床试验中，研究者在试验设计阶段通过近似正态方法进行样本量估算，在评价阶段采用 CMH 卡方计算两组率差的点估计和置信区间，结合非劣效界值判断试验产品相对于对照产品非劣效是否成立[2,4,11]。然而，当两组率同时为 0% 或 100% 时，CMH 卡方将无法实现试验的评价过程，给研究者评价产品性能带来一定的困难。Newcombe -Wilson 得分方法解决两组率同时为 0% 或 100% 率差计算问题，为试验结果的评价提供了更多选择。

但如果设计方法和评价方法不一致，研究则有可能出现即使试验结果已经达到最优整个试验却失败的情况。譬如实例分析中，即使两组所有产品都为成功，有效率为 100%，两组率差的 95% 置信区间下限仍然超过非劣效界值。因此，无论用什么方法或者用软件估计样本量，需要用统计检验方法对应起来，不同的检验方法对应不同的样本量估计。

本文对试验组和对照组有效率都为 100% 的非劣效临床试验设计和评价进行了探讨，但对于采用近似正态非劣效试验设计，不同的评价方法检验效能是否不同，是否会随着有效率的变化而发生变化，这些方面有待进一步深入挖掘和研究。

参 考 文 献

[1] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods [J]. Stat Med, 1998, 17(8): 873-890.

[2] 王杨，李卫，成小如，等. 随机模拟法验证非劣效临床试验样本量计算公式 [J]. 中国卫生统计，2008，25(1)：26-28.

[3] Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods [J]. Stat Med, 1998, 17(8): 857-872.

[4] CCTS 工作小组，夏结来. 非劣效临床试验的统计学考虑 [J]. 中国卫生统计，2012，29(2)：270-274.

[5] 刘玉秀，陈峰，苏炳华，等. 非劣性/等效性试验的样本含量估计及把握度分析 [J]. 中国卫生统计，2004，21(1)：31-35.

[6] 刘沛. 四种方法计算总体率可信区间的比较研究 [J]. 中国卫生统计，2005，22(6)：354-358.

[7] Miettinen O, Nurminen M. Comparative analysis of two rates [J] Stat Med, 1985, 4(2): 213-226.

[8] 唐欣然，黄耀华，王杨，等. 单组目标值试验样本量计算方法的比较研究 [J]. 中华疾病控制杂志，2013，17(11)：993-996.

[9] 王杨，胡泊，陈涛，等. 抽样调查法和单组目标值法对诊断试验样本量计算差异的分析 [J]. 中华流行病学杂志，2010，31(12)：1403-1405.

[10] 刘沛. 总体率可信区间计算的一次近似法及其特征 [J]. 中国卫生统计，2004，21(5)：297-299.

[11] 李卫. 医疗器械临床试验统计方法 [M]. 北京：人民军医出版社，2012.

(收稿日期：2014-09-22)
(修回日期：2014-12-17)
(陈双双校)