# 医学论文统计学报告水平评价量表的研制及其意义*

刘清海[1) ] 方积乾[2)]

(1)中山大学学报编辑部/编辑学与出版研究中心;2)中山大学公共卫生学院流行病与统计学系:510080,广州)

**摘 要** 作为实时、内部评价医学论文科学水平的一部分,在系列研究的基础上研制出含27个项目的医学论文统计学报告水平评价量表。经过信度和效度的外考核,该量表在不同人之间评价信度较好,对不同级别杂志的区分效度也好,可用于医学论文的评价。本量表适用于医学编辑对未发表论文科学水平的实时内部评价,也适用于同行或管理部门对已发表论文的评价,其适用对象是有对照的临床研究论文。

**关键词** 医学论文;统计学报告水平;评价量表;内部评价;论文评价;科学水平

**Development and significance of scale of medical article's statistical reporting level**//LIU Qinghai, FANG Jiqian

**Abstract** The authors of this paper developed a Scale of Medical Article's Statistical Reporting Level which contained 27 items based on the series research. After the external assessment, the reliability of the scale among different evaluators is good enough, and the distinction ability of the scale between different reporting levels of medical journals is also satisfactory. It can be used for the evaluation of medical papers. The Scale can be applied to the medical editors or peer reviewers of unpublished medical papers in real-time evaluation mode, also can be applied to the evaluation of the published papers by administration departments. The applicable targets are controlled clinical research papers.

**Key words** medical article; statistical reporting level; evaluation scale; internal assessment; article evaluation; scientific level

**Author's address** Editorial Office of Journal of SUN Yatsen University (Medical Sciences), 510080, Guangzhou, China

上世纪末,对科技期刊的评价研究方兴未艾。从文献计量学的角度对科技期刊进行评价的主要方法有核心期刊、引用情况等,同时,以科技期刊的质量代替科技论文质量而行简单管理的做法开始引起非议[1];于是,近年来,对科技论文的评价研究开始热闹起来,评价科技期刊的某些指标也移植到科技论文的评价中。以期刊的影响因子代替论文质量存在重大缺陷[1]。应明确期刊的质量并不完全代表论文的质量[2],而且文献计量学指标都是从论文的外部进行的评价,是间接的、事后的评价,而非从论文内部进行的直接的、实时的评价[3]。对于医学论文,统计学是评价它们的重要方面[4],应将统计学评审纳入医学期刊的评价内容,加强医学编辑对论文统计学问题的鉴审工作[5]。广义的统计学包括科研设计、数据收集、统计分析方法、对统计结果的推断与解释及对整个过程的报告等,可以说是论文科学水平的主要体现[6]。要提高医学期刊的学术水平和规范化水平,就必须正确表述论文的统计学问题[7]。为此,笔者通过国内40余位各医学相关学科的专家得出一份广义统计学项目列表,并从中精选出一份医学论文统计学报告水平的评价量表,以供医学编辑、医学科研管理工作者及论文作者参考。

## 1 调查方法

我们的前期工作得到一份临床试验论文统计学项目自查清单(即必须报告项目)和含35个项目的初步评价量表[8-9]。我们认为,要对各医学论文的统计学报告水平进行评价,其评价项目必须能区分出各等级的医学论文,而那些所有论文都报告的统计学项目不能拉开各等级医学论文的差距,因此,应从评价表中予以清除。依照这种思想,我们初定《British Medical Journal》(BMJ)、《中华医学杂志》和《临床医学》作为统计学报告水平的3个层次,以初步评价量表对其中有对照的临床研究进行评价预试验,结果是初步评价量表能区分出3种层次的杂志,于是,我们将其扩大应用到国外的《BMJ》《Lancet》《Ann Intern Med》共3种权威医学杂志,国内的《中华医学杂志》《中华内科杂志》《中华外科杂志》及任意选定非中华医学系列非学报系列的其他杂志的有对照的临床研究论文,共45篇,以方便在应用中发现该评价量表的不足之处及进一步精简其中的项目。

## 2 最终评价量表的形成

我们对初步评价量表扩大应用于45篇论文后所有项目的总方差、标准化方差及中文文献的方差进行了计算与排序,将其中有2种排序皆排在了后10位的9项删去,而I 2项在中文文献排序中排在了后10位,对外文文献来说都很少报道,因此也删去。对剩余项目再仔细分析,重新描述,其中"各组样本量是否清楚"单独出一项。另外,中文文献中纯粹的RCT研究较少,但分组研究却不少,因此,将"分组方法是否清楚"补充进去。于是最终的评价量表变成了27项(见

*广东省软科学项目(2005B70101121)

表 1,项目代号已重排)。

**表 1　随机对照临床研究论文统计学报告项目评分表**

| 项目编号与描述 | | 评分 |
|---|---|---|
| M1 | 目标人群描述是否清楚(如年龄、地理、转诊) | 1,0 |
| M2 | 是否有明确的诊断标准 | 1,0 |
| M3 | 是否包括入选标准与排除标准(有 1 项即得分) | 1,0 |
| M4 | 有无说明确定样本量的理由 | 1,0 |
| M5 | 有无说明分组的具体方法(仅"随机"2 字可计 1 分,但请在旁注 * 号) | 1,0 |
| M6 | 若盲法,则有无说明谁"盲"对什么因素(未采用盲法亦应如是说) | 1,0 |
| M7 | 有无说明数据收集的方法(手工、仪器或计算机皆可) | 1,0 |
| M8 | 有无定义观察或研究终点(个体或整体皆可) | 1,0 |
| R1 | 研究与随访的起止时间是否清楚(有 1 项不明,则不得分) | 1,0 |
| R2 | 志愿对象例数或符合入选标准数是否描述清楚(实际分组各组例数不在此项) | 1,0 |
| R3 | 是否说明失访数及原因(缺一不可,若无失访应如是说) | 1,0 |
| R4 | 分析的数据集是否清楚(全集或符合方案集等) | 1,0 |
| R5 | 是否说明依从实验计划或违反计划治疗例数 | 1,0 |
| R6 | 有无描述干预前主要指标的集散趋势(缺一不可) | 1,0 |
| R7 | 有无描述干预后主要指标的集散趋势(缺一不可) | 1,0 |
| R8 | 有无描述干预效果(效应大小)的指标(如差值、比值或改善率等) | 1,0 |
| R9 | 主要指标的各组样本量是否清楚 | 1,0 |
| R10 | 主要指标统计检验的实际方法是否清楚 | 1,0 |
| R11 | 有无提供主要指标检验的统计量值(如 $t$、$F$、$\chi^2$) | 1,0 |
| R12 | 有无提供主要指标检验的确切 $P$ 值(而非大于或小于某界值)或置信区间 | 1,0 |
| R13 | 在相应 $P$ 值(或置信区间)前后做出具实践意义的决定 | 1,0 |
| R14 | 有无分组报告负性反应或事件的人次与程度(未分组不得分,讨论中不计分) | 1,0 |
| D1 | 讨论首段有无说明本研究的性质以指导讨论 | 1,0 |
| D2 | 有无对设计中可能存在的偏倚或研究的不足进行说明 | 1,0 |
| D3 | 有无综合比较干预措施的利弊从而得出概括性结论(仅有结论没有比较得 0 分) | 1,0 |
| D4 | 有无对该研究结论的适用性(外推性)进行说明 | 1,0 |
| D5 | 有无结合其他文献加强或平衡本文结论(纯解释性不计分) | 1,0 |

请画圈评分。答案肯定者得 1 分,否则得 0 分。M 为材料与方法部分;R 为结果部分;D 为讨论部分。R1 ~5 也可在方法部分。全集:即所有实行分组者皆纳入分析,符合方案集即扣除违反方案者再行分组分析。

## 3　最终量表的效度和评价的一致性考核

首先确定了《BMJ》和《JAMA》、《中华医学杂志》和《中华内科杂志》、《临床医学》和《河北医学》作为 3 个层次的医学期刊,查出其近 5 年的所有随机对照研究或至少是有对照的临床研究论文,随机抽取每种杂志各 9 篇文献(以 2 级中文杂志的文献得分差距计算出的样本量),请 4 位流行病学或统计学硕士研究生各随机选择每种杂志的 2 篇论文,以最终量表评价论文统计学报告水平得分,第一作者也参与评价各杂志剩余的 1 篇论文。结果是,不同评价者之间的一致性较好(表 2)。该量表对不同级别统计学报告水平的区分效度也很好,虽然项目比初表精简了许多,但仍能将 3 种层次的杂志完全区分开来(表 3)。由于量表终表的总分为 27 分,在实际应用时,按本研究的 95% 可信限看,级别的区分可为小于 12 分、12 ~17 分、18 分以上 3 级。

**表 2　不同评分者用量表终表评价的一致性**　(得分)

| 评分者 | 论文数 | 均数 | 标准差 | 95% CI | 最小值 | 最大值 |
|---|---|---|---|---|---|---|
| 1 | 12 | 17 | 5 | (14,21) | 10 | 25 |
| 2 | 12 | 12 | 5 | (9,15) | 6 | 21 |
| 3 | 12 | 16 | 6 | (12,19) | 8 | 24 |
| 4 | 12 | 16 | 6 | (13,20) | 6 | 23 |
| 5 | 6 | 14 | 6 | (7,20) | 8 | 22 |
| 总计 | 54 | 15 | 5 | (14, 17) | 6 | 25 |

注:$F = 1.650$,$P = 0.177$。

**表 3　量表终表对不同级别杂志论文评价的效度**(得分)

| 杂志级别 | 论文数 | 均数 | 标准差 | 95% CI | 最小值 | 最大值 |
|---|---|---|---|---|---|---|
| 外文文献 | 18 | 20.9 | 2.8 | (19.6,22.3) | 15 | 25 |
| 中华医学系列杂志 | 18 | 14.4 | 4.1 | (12.4,16.4) | 6 | 20 |
| 国内一般医学杂志 | 18 | 10.3 | 2.5 | (9.1,11.6) | 6 | 14 |
| 总　计 | 54 | 15.2 | 5.4 | (13.7,16.7) | 6 | 25 |

注:$F = 49.73$,$P < 0.001$,经 SNK 检验,3 个级别之间能在 0.05 的水平两两区别开来。

## 4　评价量表的意义与适用性

前些年业界比较注重对科技期刊的评价,近些年则开始关注对科技论文的评价。对论文的评价可分为外部评价和内部评价。受对期刊评价的影响,对论文的评价也比较注重从外部进行评价。本研究的评价量表属于实时、内部评价的一部分。对于医学论文,内部评价一般都是从先进性、实用性和科学性来进行的。科学性评价没有一个统一的标准。实际上,科学性问题基本上就是广义的统计学问题,因此,我们的系列研究先研究出重要的统计学项目,然后精选出必须报告的统计学项目作为作者投稿时的自查清单。本研究进一步研究出本评价量表,并对其效度和评价一致性进行了考核,发现其效度和评价一致性都好,可用于医学论文的统计学报告水平的评价。

本研究的评价量表不但适用于医学编辑对未发表论文的实时评价,而且适用于同行或管理部门对发表后的论文进行评价。杨扬等[5]提出,应将统计学内容纳入期刊的评价内容。我们也认为,现行的科技期刊评价系统不甚完善,可以单设一些专项评价项目[5]。具体到本研究,鉴于医学论文的统计学缺陷太常见,而其中报告缺陷是重要内容,因此,可以单设一个医学论文统计学报告水平的评价项目。评价时当然不用选中期刊的所有文献,只需有代表性地选择近期某些合适