编辑应重视审核医学论文中的诊断试验分析方法

张 媛 韩 锟† 石朝云
中华医学会杂志社,100710,北京

摘 要 通过实例介绍医学期刊中诊断试验的基本概念与统计分析方法,分析实际应用中存在的错误及不明确之处,并提出一些应对策略及办法。

关键词 医学期刊;诊断试验;编辑

Editors should pay attention to the audit of the diagnostic test analysis methods in medical papers//ZHANG Yuan, HAN Kun, SHI Zhaoyun

Abstract By combining with concrete examples, the paper introduces the basic concepts and statistical analysis methods of diagnostic tests in medical journals, analyzes the errors and ambiguities of practical application, and proposes some measures for reference.

Keywords medical journal; diagnostic test; editing

Authors' address Editorial Department of Chinese Journal of Laboratory Medicine, 100710, Beijing, China

随着医学技术的迅猛发展,新的诊断方法不断涌现,原有的方法也在不断改进;因此,在临床医学研究论文中会面临大量有关诊断方法的评价问题。新的诊断试验方法出来后,为了了解它们的特点,探讨其应用前景,是否有可能取代当前普遍应用的、旧的检验诊断方法,因而有必要采用一些特定的指标对其进行科学、客观的评价[1-3]。常用的诊断性试验的统计指标有敏感度、特异度、误诊率、漏诊率、正确率、Youden 指数、比数积、阳性似然比、阴性似然比、阳性预测值、阴性预测值等。这些统计指标的不足之处在于受"诊断分界点"的影响较大。克服这一缺点的方法是采用一种较新的统计分析方法,即"ROC 曲线法"。编辑在遇到这一类情况时,往往不能准确掌握其表述方式,现予介绍,供同人参考。

1 常用诊断试验评价的指标

在进行诊断性试验前,首先要明确"金标准"的概念。对人体或动物的活检,对细菌的培养和定性,内镜或手术等直视下所见的诊断,可以明确判断疾病的有无,不存在假阳性或假阴性的结果,此种诊断标准称为"金标准"[4]。此外,对很难得到形态学特征的疾病,可将根据功能性检查结果确定的诊断标准,或本专业普遍承认的标准作为金标准[5]。不存在"金标准"的试验,由于缺乏对事物真实属性能够作出客观描述的标准,对欲研究的诊断方法缺乏一把客观公正的"尺子",因而是缺乏科学性的。

一般根据诊断试验的结果和用金标准判断的结果,可以得到 4 种情况,见表 1。

表 1 诊断试验评价四格表

诊断试验	金标准		合计
	患者	非患者	
阳性	a(真阳性)	b(假阳性)	$a+b$
阴性	c(假阴性)	d(真阴性)	$c+d$
合计	$a+c$	$b+d$	$a+b+c+d$

1.1 敏感度(sensitivity) 又称"真阳性率",是实际患病者中检查为阳性者所占的比例。敏感度高的检查,患病的人几乎不会被漏掉,反映了该试验检出病例的能力。敏感度只与病例组有关,与非患者组无关。敏感度 $= a/(a+c) \times 100\%$。

1.2 特异度(specificity) 又称"真阴性率",是实际未患病的人中检查为阴性者所占的比例。特异度高的检查,几乎不会有非患者被误判为患者,反映了该试验排除非病例的能力。特异度 $= d/(b+d) \times 100\%$。

1.3 误诊率(mistake diagnostic rate) 又称假阳性率,是实际未患病的人中检查为阳性者所占的比例,即非患者被诊断为患者,反映非患者被错误诊断的可能性。误诊率 $= b/(b+d) \times 100\%$。

1.4 漏诊率(omission diagnostic rate) 又称假阴性率,是实际患病的人中检查为阴性者所占的比例,即患者被诊断为阴性,反映患者被错误诊断的可能性。漏诊率 $= c/(a+c) \times 100\%$。

1.5 正确率 又称"符合率",是指诊断试验中真阳性和真阴性的人数之和占总受检人数的比例。正确率 $= (a+d)/(a+b+c+d) \times 100\%$。

1.6 Youden 指数(Youden index) 为真阳性率与假阳性率之差,即敏感度与特异度之和减去 1,反映了诊断试验发现患者与非患者的总的能力,其取值在 -1 到 1 之间,取值越大,说明诊断试验的真实性越好。

1.7 阳性似然比 是诊断试验中阳性结果在患者中出现的概率与在非患者中出现的概率之比,即真阳性率与假阳性率之比。此值越大,说明诊断试验确诊疾病的能力越强。

1.8 阴性似然比 是诊断试验中阴性结果在患者中

† 通信作者

出现的概率与在非患者中出现的概率之比,即假阴性率与真阴性率之比。此值越小,说明诊断试验排除疾病的能力越强。

1.9 比数积(odd product) 表示患者中诊断阳性数、阴性数之比与非患者中诊断阴性数、阳性数之比的乘积。此值越大,说明诊断价值越大。

1.10 阳性预测值 指诊断为阳性者,实际为患者的概率。此值越大,说明检测方法的诊断价值越高。

1.11 阴性预测值 指诊断为阴性者,实际为非患者的概率。此值越大,说明检测方法的诊断价值越高。

2 ROC分析的方法与应用

上述指标只能表达指定某一特定诊断界点时所对应的指标,当改变诊断界点时,就会得到不同的指标值,不便于评价整个诊断系统的准确性,因此需引入ROC分析方法。ROC是受试者工作特征曲线(Receiver Operating Characteristic)的缩写,起源于20世纪50年代统计决策理论,最早用于描述雷达信号与噪声之间的关系,后来在气象、材料检验、医学检验诊断领域都有广泛的应用。它将不同诊断界点下所得到的敏感度和特异度结合起来,对整个诊断系统进行综合评价,根据所绘曲线的形状和面积,对诊断试验进行定量分析,并进一步评价检测方法诊断价值的大小[6-7]。

2.1 ROC曲线的计算与构建 例如,欲探讨某检测指标对疾病的诊断价值,需要收集患者与非患者的血标本,测定该指标的数值。选取几个诊断界点,依据这几个界点,计算各自的敏感度和特异度,并绘制ROC曲线。在以不同的诊断界值作为判断标准的情况下,都能计算出相应的敏感度和特异度,以假阳性率为横轴,以真阳性率为纵轴,横轴与纵轴的长度相等且均为1,形成正方形,在坐标系上分别标上几个诊断界值所对应的工作点,即不同的敏感度和特异度所形成的点,同时标上(0,0)和(1,1)2个点,这2个点分别对应于2种极端情况,最后将相邻2点连接起来,即构建出ROC曲线。

此曲线下的面积反映诊断系统的准确性。理论上,此面积的取值范围为0.5~1.0:完全无价值的诊断系统,其真阳性率与假阳性率相等且始终为0.5,相当于从原点到(1,1)点的对角线,这条线又称为机会线,其下面积为0.5;完善的诊断系统相当于金标准,其真阳性率始终为1,假阳性率始终为0,相当于从原点垂直上升到(0,1)点,然后水平达到(1,1)点,其下面积为1。一般认为:曲线下面积在0.50~0.70,表示诊断价值较低;在0.70~0.90表示诊断价值为中等;0.90以上表示诊断价值较高。已知ROC曲线图中从原点到右上角机会线下的面积为0.5,所以,已计算出的ROC曲线下面积是否与0.5有统计学差异,还需进一步作假设检验,以评价其诊断价值。若ROC曲线下面积与0.5之间差异有统计学意义,则说明此诊断方法的诊断效果是令人满意的。

2.2 ROC曲线下面积的比较 不同的诊断系统都能获得相应的ROC曲线,ROC曲线下面积的差别是否具有统计学意义是评价各诊断系统优劣的重要标志。若第1种诊断方法所对应的曲线下面积大于第2种诊断方法所对应的曲线下面积,且$P<0.05$,说明第1种诊断方法优于第2种诊断方法。

3 存在的问题

3.1 敏感度与灵敏度概念混淆 临床上所说的灵敏度多数是能够检出抗体、肿瘤标志物等化学物质的"检出灵敏度",即可检测的最低分析物浓度为检测系统的灵敏度或称检测限,与作为检查精密度指标的"敏感度"无任何关系。例如某论文[8]的一表中的灵敏度就应该改为敏感度(表2)。

此外,某些论文存在前后概念描述不一致问题,如某论文[9]存在这样的表述:"TIF诊断SLE的敏感性和特异性分别为46.1%和99.2%,ELISA诊断SLE的敏感度和特异度分别为51.3%和96.7%。"前一句话称为"敏感性和特异性",后一句话又改为"敏感度和特异度",即违背了科学性,又违背了编辑加工规范。正确的描述应该全文统一为"敏感度和特异度",并且应同时标注绝对数,如敏感度为73.3%(88/120)。

3.2 指标堆砌问题 某些论文在报告敏感度和特异度的同时,又给出了假阴性率(数值=1-敏感度)和假阳性率(数值=1-特异度),这就如同给出了正确率为80%,又给出了错误率为20%一样,毫无意义。

例如某论文[8]存在表3这样的描述,即给出了真阳性率(即敏感度)、真阴性率(即特异度),又给出了假阳性率和假阴性率。需要指出的是,表3中除了存在指标堆砌问题之外,还有计算错误:如果敏感度和特异度计算无误的话,假阳性率应该分别为49.28%、47.10%、50.47%、49.98%,假阴性率应该分别为72.85%、69.03%、66.70%、65.85%。其原因可能是由统计表制作不规范,诊断试验检测的阳性和阴性结果位置互换,金标准诊断结果与诊断试验结果位置互换等问题导致的。

3.3 与Kappa检验混淆 若新的诊断试验是与"非金标准方法"比较,则应作一致性(Kappa)检验,而不能计算敏感度与特异度;因为"非金标准方法"中也会有假阳性和假阴性,解释起来就会缺乏说服力。

表 2　朗迈 UriSed 尿液分析仪检测 RBC、WBC 和 CAST 的结果分析(%(例))

参数	临界值/L^{-1}	灵敏度	特异度	阳性预测值	阴性预测值	符合率
RBC	5.00	82.12(326/397)	94.99(1537/1618)	80.10(326/407)	95.58(1 537/1 608)	92.46(1 863/2 015)
WBC	9.00	93.06(402/432)	89.83(1 422/1 583)	71.40(402/563)	97.93(1 422/1 452)	90.52(1 824/2 015)
CAST	2.00	51.61(16/31)	94.86(1 882/1 984)	13.56(16/118)	99.21(1 882/1 897)	94.19(1 898/2 015)

注:表中“灵敏度”应该改为“敏感度”。

表 3　朗迈尿液分析工作站自动化尿常规分析不同设计方案的复检规则分析(%(例))

方案	真阳性率(敏感度)	假阳性率	真阴性率(特异度)	假阴性率	复检率
1	27.15(547/2 015)	15.78(318/2 015)	50.72(1 022/2 015)	6.36(128/2 015)	42.93(865/2 015)
2	30.97(624/2 015)	7.99(161/2 015)	52.90(1 066/2 015)	4.42(89/2 015)	39.70(810/2 015)
3	33.30(681/2 015)	15.09(314/2 015)	49.53(998/2 015)	1.34(27/2 015)	29.58(596/2 015)
4	34.15(688/2 015)	14.74(297/2 015)	50.02(1008/2 015)	1.04(21/2 015)	18.91(381/2 015)

注:此表中除了存在指标堆砌问题之外,还有计算错误。如果真阳性率(敏感度)和真阴性率(特异度)计算无误的话,假阳性率应该分别为 49.28%、47.10%、50.47%、49.98%,假阴性率应该分别为 72.85%、69.03%、66.70%、65.85%。

例如某论文[10]中,用硝酸盐还原试验(NRA 法)在液体培养基的基础上直接检测痰标本中的结核分枝杆菌对 4 种一线抗结核药物的耐药性,同比例法结果进行比较,以评价该方法的临床应用价值。在结果中有如下描述:“以比例法结果为判断标准,NRA 法检测结核分枝杆菌对利福平、异烟肼、链霉素、乙胺丁醇的耐药性有很好的敏感度(100%、100%、90.9%、93.8%)和特异度(97.9%、99.1%、98.1%、98.0%),一致性 Kappa 检验分析结果表明,有极佳的一致性(Kappa 值分别为 0.97、0.98、0.93、0.92)。”这里,明显的错误是没有明确“比例法”是否为金标准。纵观全文,都没有找到明确的说法,只是描述为“以比例法结果为判断标准”。如果该比例法是金标准,那么只计算敏感度和特异度就足够了,根本不用作 Kappa 检验;如果该比例法不是金标准,那么只能作 Kappa 检验,而不能计算敏感度和特异度。

3.4　案例分析　下面以文献[11]为例说明它的错误所在或不明确性。该文题名为《荧光定量 PCR 检测社区获得性肺炎患者痰标本中肺炎链球菌》,作者将常规痰培养法检测结果与荧光定量聚合酶链反应(PCR)检测肺炎链球菌的结果进行对比分析,同时结合患者临床表现及治疗反应判断荧光定量 PCR 结果,探讨荧光定量 PCR 对 CAP 病原体检出的价值和临床意义。作者在结果中列了 4 个表(表 4~7),分别用于影像学、白细胞计数、治疗反应以及痰培养结果与荧光定量 PCR 检测结果的比较,并对表 4~6 描述如下:“影像学有典型肺炎影与检测到含肺炎链球菌 DNA 负荷量≥10^4cfu/mL 敏感度、特异度为 100%,一致性为 100%;白细胞计数≥10×10^9 L^{-1}与检测到含肺炎链球菌 DNA 负荷量≥10^4 cfu/mL 敏感度 100%、特异度为 89.5%,一致性为 90.9%;治疗反应与检测到含肺炎链球菌 DNA 负荷量≥10^4 cfu/mL 敏感度 80%、特异度为 100%,一致性为 95.5%。”此处的问题除表述模糊、含义不清外,主要在于影像学、白细胞计数、治疗反应均不是诊断肺炎或检测肺炎链球菌的金标准,没有“金标准”作为计算的基础,敏感度和特异度从何而来?与这些非金标准方法比较来计算敏感度、特异度是明显错误的;但作者同时计算了一致性倒是正确的。

表 4　影像学与荧光定量 PCR 对比(例)

荧光定量 PCR	影像学		合计
	+	−	
+	10	0	10
−	0	34	34
合计	10	34	44

表 5　白细胞计数与荧光定量 PCR 对比(例)

荧光定量 PCR	白细胞计数		合计
	+	−	
+	6	4	10
−	0	34	34
合计	6	38	44

表 6　治疗反应与荧光定量 PCR 对比(例)

荧光定量 PCR	治疗反应		合计
	+	−	
+	8	2	10
−	0	34	34
合计	8	36	44

表 7　痰培养与荧光定量 PCR 对肺炎链球菌检出效果对比(例)

荧光定量 PCR	痰培养		合计
	+	−	
+	1	9	10
−	0	34	34
合计	1	43	44

而作者对表7的描述如下:“痰培养方法检测到肺炎链球菌感染例数为1例,荧光定量PCR检测肺炎链球菌感染例数为9例。统计分析痰培养与荧光定量PCR对肺炎链球菌检出例数的差异有统计学意义($P=0.004$)。”众所周知,痰培养是鉴定细菌感染的金标准,在最应该分析敏感度与特异度的表7中,作者却没有计算。若编辑能够审查出该文的缺陷或不明确所在,在稿件退修时就应该要求作者补充相关内容,也就不会出现以上问题了。

4 对策

4.1 加强学习,提高对科技论文的辨识水平 编辑要认真学习统计学知识,只有自己认为存在错误的文章,或者有疑问的文章,才会去进一步追究其正确与否。若编辑自己看不出问题,那只能带着错误出版了。例如:在遇到将“灵敏度”与“特异度”配对使用来进行诊断性试验的评价时,编辑应该将“灵敏度”改为“敏感度”;但也要视论文具体情况作具体处理,不能见了“灵敏度”就改为“敏感度”,在描述某检测系统的分析时,“灵敏度”就不能改为“敏感度”。

4.2 增强编辑责任心,认真审核稿件 对这一类论文,应注意审查统计表制作是否规范,计算方法是否正确,数据是否有误。对有疑问的文章,要与作者沟通,看作者自己是否能讲清楚,是否已经请教了相关统计学专家。作者是源头,要杜绝虚假,端正学风;编辑是闸口,要起到督促、把关作用。

4.3 合理利用同行评议,加强统计学审稿 在实际三审、五定、编辑加工的流程中,未必每个编辑部都能做到将每篇涉及统计学的稿件都送给统计学专家审稿;所以,对有疑问的文章,务必请统计学专家审核。

5 结束语

在医学诊断研究中,对诊断试验进行科学的研究和评价是正确认识该诊断试验的临床应用价值以及临床上合理选用各种诊断试验、科学地解释诊断试验各种结果的基础。但它的前提是务必采用正确的方法进行评价,在已发表的部分相关论文中确实存在大大小小的各方面的问题,这有待作者在研究设计之初即严格按照相关规定执行,同时也要求编者在编辑加工过程中严格把关,对有疑问的文章一定不能轻易放过,不可存在侥幸心理。

6 参考文献

[1] 林果为. 诊断试验的研究与评价[J]. 诊断学理论与实践,2003,2(1):1-4

[2] 陈平雁. 关于诊断试验方法的若干问题[J]. 北京大学学报:医学版,2010,42(6):764-766

[3] Zhou X,Obuchowski N A,Mcclish D K. Statistical methods in diagnostic medicine[M]. New York: Wiley,2002:111-136

[4] 熊鸿燕,易东. 医学科研方法:设计、测量与评价[M]. 重庆:西南师范大学出版社,2005:42

[5] 陈卫中,张菊英. 金标准为等级变量时诊断试验的评价及其在冠心病诊断试验中的应用[J]. 中国卫生统计,2012,29(2):172-174

[6] Nguyen P. NonbinROC: software for evaluating diagnostic accuracies with non-binary gold standards[J]. Journal of Statistical Software,2007,21(10):1-10

[7] Obuchowski N A. A ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale [J]. Statistics in Medicine,2006,25(3):481-493

[8] 马骏龙,陆玉静,李兴翠,等. 朗迈全自动尿液分析工作站复检规则制定与评价[J]. 中华检验医学杂志,2012,35(9):810-814

[9] 史晓敏,阎振林,隋宝环,等. 两种检测抗双链DNA抗体方法对系统性红斑狼疮的诊断价值比较[J]. 中华检验医学杂志,2012,35(8):742-745

[10] 田艳生,李红光,崔幸琨,等. 硝酸盐还原试验直接检测痰标本中结核分枝杆菌耐药性[J]. 中华检验医学杂志,2012,35(7):653-655

[11] 饶春梅,张中和,王勇,等. 荧光定量PCR检测社区获得性肺炎患者痰标本中肺炎链球菌[J]. 中华检验医学杂志,2012,35(4):367-369

(2013-02-18收稿;2013-03-01修回)

中国出版的6 225种期刊2012年主要计量指标的平均值统计

1)总被引频次1 089次/刊,≥1 000次的期刊共1 873种;2)影响因子0.427,≥1.000的期刊共398种;3)即年指标0.067,603种期刊为0;4)基金论文比0.229,310种期刊为0;5)海外论文比0.011,≥0.200的期刊共102种(其中英文期刊77种),3 787种期刊没有海外论文;6)他引率0.91;7)平均作者数2.17人/篇;8)平均引文数7.01条/篇;9)来源文献量357;10)地区分布数19;11)机构分布数173。

(卞 吉:摘编自《2013年版中国科技期刊引证报告(扩刊版)》第Ⅱ页)