·循证医学中的医学统计学问题·

基于EXCEL软件的诊断性Meta分析中缺失数据的提取方法

瞿 振[1]， 尹长青[2]， 胡翠苹[3]

（1. 广东食品药品职业学院护理学院，广州 510520；
2. 武汉大学中南医院检验科，武汉 430071；3. 武汉大学药学院，武汉 430071）

［摘要］ 诊断性Meta分析逐渐成为循证医学研究中的一种重要分析方法，但是，实际上因许多文献中数据提供不全，常导致文献无法纳入而降低诊断性Meta分析的效能。本文总结几种常见的诊断性Meta分析中数据不全的情形，应用实例数据，以Excel软件为平台，提供一种纳入文献数据给出不全的解决方案。

［关键词］ 诊断性Meta分析；Excel软件；数据提取
［中图分类号］ R195.1 ［文献标识码］ A DOI：10.12019/j.issn.1671-5144.2017.02.012

A Method of Data Extraction Based on Excel for Incomplete Data in Diagnostic Meta-Analysis

QU Zhen[1], YING Chang-qing[2], HU Cui-ping[3]

(1. College School of Nursing, Guang Dong Food and Drug Vocational, Guangzhou 510520, China; 2. Zhongnan Hospital, Wuhan University, Wuhan 430071, China; 3. School of Pharmaceutical Sciences, Wuhan University, Wuhan 430071, China)

Abstract: Diagnostic meta-analysis has become an important method of evidence-based medicine study. However, most of the original study did not offer the comprehensive data, which led to the study cannot be incorporated in meta-analysis and reduce the effectiveness of diagnostic. We describe several situations of insufficient data and the principle of the data extraction in diagnostic analysis. Excel was exploited to calculate the extracted data by application data analysis when complete data are unavailable in diagnostic meta-analysis.

Key words: diagnostic meta-analysis; Excel software; data extract

诊断性Meta分析是一种重要的Meta分析类型，它涉及到诊断标志物的敏感性、特异性以及对受试者工作特征曲线（receiver operating characteristic curve，ROC）的综合性评估[1-2]。ROC曲线是一种重要的处理诊断性Meta分析数据的方法，通过ROC曲线，我们可以获得许多与诊断性相关的数据。之前曾介绍一种运用软件在ROC曲线上提取敏感性和特异性数据的方法[3]，很好地解决了在Meta分析实践中从图片数据中获取实验数据的问题。然而，我们也常常会遇到这样的情况，由于文章的侧重点不同，有些数据在文章中既没有以图片形式给出，也没有完整的数据形式，或多或少缺乏一些我们想要的关键数据，这为我们进行诊断性Meta分析带来了不小的困难。本文就诊断性分析相关的文献中常见的一些数据给出形式进行简述，并且提供一种简单方便的方法，在数据给出不全的情况下用来提取相关数据，为诊断性Meta分析过程中的完整数据提取提供便利。

［作者简介］ 瞿振（1989-），男，湖北洪湖人，讲师，硕士研究生，研究方向为分子免疫学和临床检验诊断。

1 诊断性分析中几种常见数据的提供方式

诊断性研究常常以信息丰富的ROC曲线来表示诊断效能情况，诊断界值的选取并不会改变ROC曲线对该种疾病的诊断性能的评价[4]。通过选取合适的cut-off值，诊断的效能就反应在此时cut-off值对应的各种诊断指标，如真阳性（true positive，TP）、假阳性（false positive，FP）、假阴性（false negative，FN）和真阴性（true negative，TN）中，常用于诊断性分析的四格表如表1所示。

在诊断性Meta分析实践中，研究结果最直接的信息呈现方式为TP、FP、FN和TN，当给出了TP、FP、FN和TN等对应的诊断数据后，通过表2所给的公式，分别计算出各种诊断性能评价的指标如敏感性、特异性、阳性预测值以及阴性预测值，这是文章中常见的数据给出形式。此外，还有一些数据形式要通过简单计算才能得出，如患者人数和健康人数，一些数据给出情况如表2所示。

表1 受试诊断研究数据报告四格表

受试诊断	金标准诊断	
	阳性	阴性
阳性	TP	FP
阴性	FN	TN

表2 常见的与诊断性研究相关的指标和公式

缩略名	名称	表达式	表示意义
TP	真阳性例数	a	被正确地识别为疾病的患病对象数，即金标准阳性，受试阳性
FP	假阳性例数	b	被错误地识别为病变的健康受试者数，即金标准阴性，受试阳性
FN	假阴性例数	c	被错误地识别为健康的患病对象数，即金标准阳性，受试阴性
TN	真阴性例数	d	被正确地识别为正常的健康受试者数，即金标准阴性，受试阴性
Sen	敏感性	a/(a+c)	受试方法诊断患者的能力
Spe	特异性	d/(b+d)	受试方法确定非患者的能力
PPV	阳性预测值	a/(a+b)	受试方法阳性者患病的可能性
NPV	阴性预测值	d/(c+d)	受试方法阳性者不患病的可能性
n1	患者数	a+c	金标准诊断为病变的数目
n2	健康数	b+d	金标准诊断为健康的数目
N	总人数	a+b+c+d	参与受试的总人数

2 诊断性分析中几种常见数据的计算方法

在做诊断性Meta分析时，我们优先想得到与诊断直接相关参数，或者用表2中的公式通过简单计算转换即可得到的数据。但在有些情况下，在我们所纳入的文献中并未提供以上完整的数据，如一种在文章中较常见的数据给出形式为仅提供Sen、Spe以及N值，缺少能完整计算出TP、FP、FN和TN的一个或两个参数，甚至有时作者会对其研究中非重点的诊断数据干脆仅以ROC曲线的图表展示[5]，在这种情况下，尽可能地得到或者转换为有用信息的方法非常重要。

解决这种问题的一个思路就是求解上述方程组，由于在诊断性研究中，TP、FP、FN和TN值代表的是各种研究结果的例数，均为正整数，因此通过文章中提供的不完整信息组成的方程组，通过求其正整数解，进一步得出TP、FP、FN和TN值也就成为可能。据此，我们总结了常见的几种数据提取的情形和方法（如表3所示）。

表3 几种数据的给出形式和计算求解方法

数据给出形式	计算公式	求解方式
（1）直接给出TP、FP、FN和TN	a=TP，b=FP，c=FN，d=TN（公式一）	直接给出
（2）给出Sen、Spe和n1、n2	a/(a+c)=Sen；d/(b+d)=Spe；a+c=n1；b+d=n2（公式二）	简单计算
（3）给出Sen、Spe、PPV、NPV中任意2个和n1、n2、N中任意1个	例如a/(a+c)=Sen；d/(b+d)=Spe；a+b+c+d=N（公式三）；例如a/(a+b)=PPV；d/(c+d)=NPV；a+b+c+d=N（公式四）	四元方程组的正整数解
（4）给出Sen、Spe、PPV、NPV中任意1个和n1、n2、N中任意2个	例如a/(a+c)=Sen；d/(b+d)=Spe；a+c=n1（公式五）；例如a/(a+c)=Sen；a+c=n1；b+d=n2（公式六）	四元方程组的正整数解

3 Excel软件在诊断性Meta分析数据提取中的应用

Excel具有十分强大的功能，其宏命令在许多研究中应用广泛[6-7]，本文根据表3对应的公式和原则，通过Excel 2013平台来说明在缺少部分数据的情况下，有关TP、FP、FN和TN值的计算。

3.1 Excel中诊断性研究数据的引用和处理

为了方便理解和操作，以Lasko等[2]文章中的数据为例来说明，例子中有10名糖尿病患者和10名健康人（共20人）接受口服葡萄糖耐量试验[2]，已分别计算出TN、FP、TP、FN值，我们利用文章中给出的血糖浓度值，在Excel中很容易计算出Sen、Spe、PPV、NPV、约登指数和到点（0，1）距离等数据（如表4所示）。

表4 接受口服葡萄糖耐量试验的实例数据

血糖浓度		金标准(−)		金标准(+)		Sen	Spe	PPV	NPV	约登指数	距离
患者（n1=10）	健康者（n2=10）	TN（d）	FP（b）	TP（a）	FN（c）						
		0	10	10	0	1.000	0.000	0.500		0.000	1.000
	4.86	1	9	10	0	1.000	0.100	0.526	1.000	0.100	0.900
	5.69	2	8	10	0	1.000	0.200	0.556	1.000	0.200	0.800
	6.01	3	7	10	0	1.000	0.300	0.588	1.000	0.300	0.700
	6.06	4	6	10	0	1.000	0.400	0.625	1.000	0.400	0.600
	6.27	5	5	10	0	1.000	0.500	0.667	1.000	0.500	0.500
	6.37	6	4	10	0	1.000	0.600	0.714	1.000	0.600	0.400
	6.55	7	3	10	0	1.000	0.700	0.769	1.000	0.700	0.300
7.29	7.29	8	2	9	1	0.900	0.800	0.818	0.889	0.700	0.224
	7.82	9	1	9	1	0.900	0.900	0.900	0.900	0.800	0.141
9.22		9	1	8	2	0.800	0.900	0.889	0.818	0.700	0.224
9.79		9	1	7	3	0.700	0.900	0.875	0.750	0.600	0.316
11.28		9	1	6	4	0.600	0.900	0.857	0.692	0.500	0.412
11.83		9	1	5	5	0.500	0.900	0.833	0.643	0.400	0.510
	12.06	10	0	5	5	0.500	1.000	1.000	0.667	0.500	0.500
18.48		10	0	4	6	0.400	1.000	1.000	0.625	0.400	0.600
18.50		10	0	3	7	0.300	1.000	1.000	0.588	0.300	0.700
20.49		10	0	2	8	0.200	1.000	1.000	0.556	0.200	0.800
22.66		10	0	1	9	0.100	1.000	1.000	0.526	0.100	0.900
26.01		10	0	0	10	0.000	1.000		0.500	0.000	1.000

3.2 Excel绘制ROC曲线的解读

通过Excel绘制出的ROC曲线图可知，在最大约登指数和最小距离点处所取的Sen和Spe分别为0.90和0.90，PPV和NPV分别为0.90和0.90，这些均与文章中报道的相一致（如图1所示）。

3.3 Excel宏文件提取不完整数据

根据上述运算思路，以制作的Excel宏程序进行运算，为方便理解，在此以常见的一种数据给出形式（3）中的公式三为例（如表3）。已知Sen=0.9、Spe=0.9和N=20，根据上述求解正整数思路和算法，可以很方便求出对应的值：TP=9、FP=1、FN=1和TN=9，利用这种思路计算的结果与原文一致（如图2所示）。

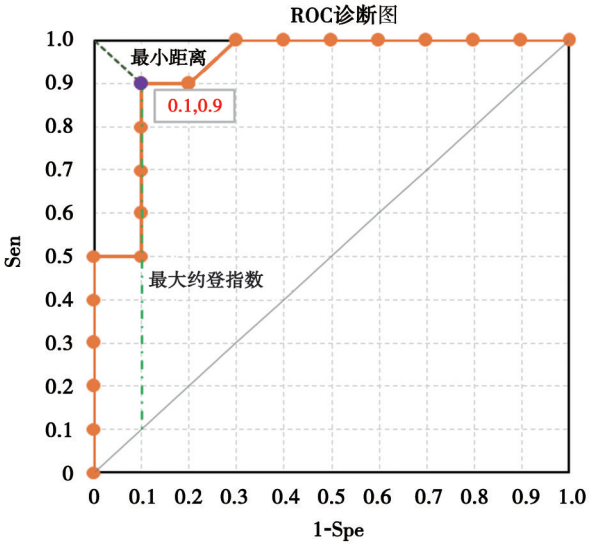

图1 口服葡萄糖耐量试验数据确定的ROC曲线

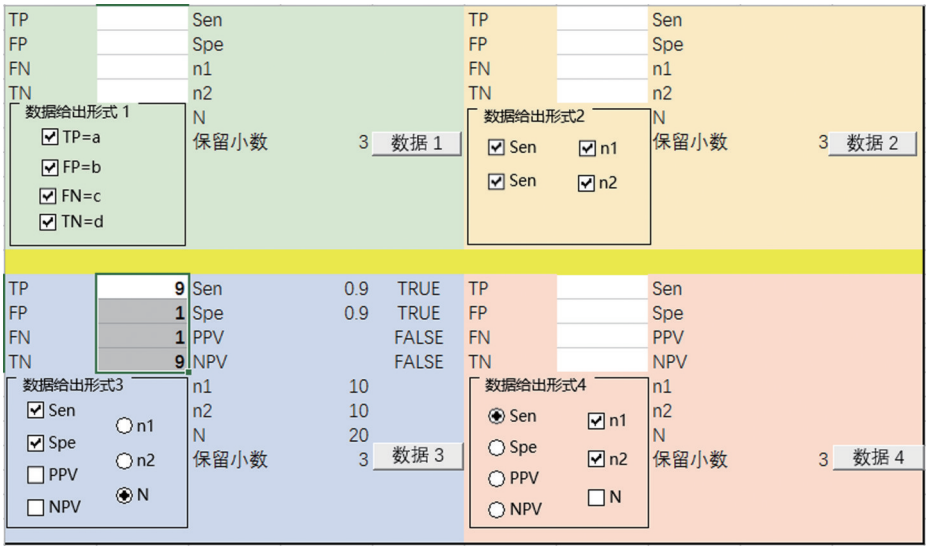

图2 EXCEL宏文件进行数据计算

3.4 提取数据的验证

为了进一步验证计算的准确性，在此采用一种数学矩阵计算软件MATLAB进行验证（如图3所示），运算出的结果与Excel宏计算出来的结果一致，接下来我们对其他数据给出形式也进行了验算，计算结果也与预期一致（数据省略）。

```
>> %a+b+c+d=20;
%a/(a+c)=0.9
%d/(b+d)=0.9
index = 0;
cxd = zeros(20,4);
    for cxdd = 1:20
        for cxdc = 1:20
            for cxdb = 1:20
                for cxda = 1:20
                    check1 = (cxda + cxdb + cxdc + cxdd == 20);
                    check2 = (cxda/(cxda +cxdc ) == 0.9);
                    check3 = (cxdd/(cxdb +cxdd ) == 0.9);
                    if check1&check2&check3;
                        index = index+1;
                        cxd(index,1) = cxda;
                        cxd(index,2) = cxdb;
                        cxd(index,3) = cxdc;
                        cxd(index,4) = cxdd;
                    end
                end
            end
        end
    end
disp(cxd(1:index,:))
    9     1     1     9
fx >>
```

分别对应a, b, c, d值

图3 MATLAB对计算结果进行验证

4 讨 论

Excel作为一种常用于数据计算和整理的软件，有强大的功能，被应用于生存曲线中数据的提取[8-9]和ROC曲线的绘制[10]。ROC曲线被广泛地应用于生物标志性分子对某种疾病的诊断性分析中，通过一系列原则，选择最合适的cut-off值，确定某分子对某种疾病诊断的敏感性和特异性值，以及ROC曲线下面积（area under the curve，AUC）的大小[11]。在先前的研究中，我们介绍了一种方法，能很好地从ROC曲线图中提取敏感性和特异性值，这种方法解决了我们在做Meta分析过程中，仅给出ROC曲线而无具体数据时导致文章无法纳入的问题[3]。本文进一步提出几种诊断性Meta分析过程中数据给出不完整时的情形（如表3所示），并提供一种不完整数据提取的解决方案。

以Lasko文章中报道的数据为例，通过Excel软件对数据进行简单的处理，以绘制的ROC曲线图和计算表格，清晰地得出最大约登指数和对应的敏感性和特异性值。我们以一种常见的数据给出形式为例，基于Excel软件，准确地计算出了TP、FP、FN和TN值。根据所提出的计算思路和方法，我们在MATLAB软件中进行验证，结果表明这些思路和方法具有较高的可靠性。

我们在此提供的方案是基于表2中诊断性分析相关的基本公式进行的，实质是利用计算机软件优势，对所需提取的TP、FP、FN和TN值，求方程组的正整数解过程。这种方法虽然很方便，但是也存在自身的局限性：一方面，在进行初始值运算时，可能会遇到找不到整数解的情况，这时可以尝试调整一下保留小数点的精确度值，再重新求解就可以了。另一方面，这种求整数解的过程是在

（下转第124页）

循证医学 http://www.jebm.cn

derived ferrichrome inhibits colon cancer progression via JNK-mediated apoptosis[J]. Nat Commun, 2016, 7:12365.

[29] KIM Y, LEE D, KIM D, et al. Inhibition of proliferation in colon cancer cell lines and harmful enzyme activity of colon bacteria by Bifidobacterium adolescentis SPM0212[J]. Arch Pharm Res, 2008, 31(4):468-473.

[30] CHEN Z F, AI L Y, WANG J L, et al. Probiotics Clostridium butyricum and Bacillus subtilis ameliorate intestinal tumorigenesis[J]. Future Microbiol, 2015, 10(9): 1433-1445.

[31] WALIA S, KAMAL R, KANWAR S S, et al. Cyclooxygenase as a target in chemoprevention by probiotics during 1, 2-dimethylhydrazine induced colon carcinogenesis in rats[J]. Nutr Cancer, 2015, 67(4):603-611.

[32] KAHOULI I, TOMARO-DUCHESNEAU C, PRAKASH S. Probiotics in colorectal cancer (CRC) with emphasis on mechanisms of action and current perspectives[J]. J Med Microbiol, 2013, 62(Pt 8):1107-1123.

[33] GIANOTTI L, MORELLI L, GALBIATI F, et al. A randomized double-blind trial on perioperative administration of probiotics in colorectal cancer patients[J]. World J Gastroenterol, 2010, 16(2):167-175.

[34] YANG Y, XIA Y, CHEN H, et al. The effect of perioperative probiotics treatment for colorectal cancer: Short-term outcomes of a randomized controlled trial[J]. Oncotarget, 2016, 7(7): 8432-8440.

[35] LIU Z H, HUANG M J, ZHANG X W, et al. The effects of perioperative probiotic treatment on serum zonulin concentration and subsequent postoperative infectious complications after colorectal cancer surgery: A double-center and double-blind randomized clinical trial[J]. Am J Clin Nutr, 2013, 97(1):117-126.

[收稿日期] 2017-03-07

(上接第119页)

限定的人数范围内,虽然常规情况下都可得出唯一的解,但是也有少数情况下可得出多组解,当遇到这种情况时,可以结合文章中提供的其他信息得出真正的解。

[参 考 文 献]

[1] ZWEIG M H, CAMPBELL G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine[J]. Clin Chem, 1993, 39(4): 561-577.

[2] LASKO T A, BHAGWAT J G, ZOU K H, et al. The use of receiver operating characteristic curves in biomedical informatics[J]. J Biomed Inform, 2005, 38(5): 404-415.

[3] 瞿振, 胡翠苹. 一种ROC曲线数据提取方法并用于miRNA-122诊断慢性病毒型肝炎的Meta分析[J]. 循证医学, 2016, 16(3): 159-164.

[4] 宋花玲, 贺佳, 虞慧婷, 等. 应用ROC曲线下面积对两相关诊断试验进行评价和比较[J]. 第二军医大学学报, 2006, 27(5): 562-563.

[5] van der MEER A J, FARID W R, SONNEVELD M J, et al. Sensitive detection of hepatocellular injury in chronic hepatitis C patients with circulating hepatocyte-derived microRNA-122[J]. J Viral Hepat, 2013, 20(3): 158-166.

[6] 鲍祥生, 梁兵, 周海燕, 等. VBA和EXCEL函数结合编程在数据处理中的应用[J]. 石油工业计算机应用, 2009, (4): 9-12.

[7] 苏进. 探究如何在EXCEL中使用VBA编程处理数据[J]. 数字技术与应用, 2016, (1): 250.

[8] TIERNEY J F, STEWART L A, GHERSI D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis[J]. Trials, 2007, 8: 16.

[9] 周支瑞, 张天嵩, 李博, 等. 生存曲线中Meta分析适宜数据的提取与转换[J]. 中国循证心血管医学杂志, 2014, 9(3): 243-247.

[10] 杨浏, 黎增文. 用Excel制作ROC曲线[J]. 现代检验医学杂志, 2005, 20(4): 81.

[11] XIA J, BROADHURST D I, WILSON M, et al. Translational biomarker discovery in clinical metabolomics: An introductory tutorial[J]. Metabolomics, 2013, 9(2): 280-299.

[收稿日期] 2016-06-30