·基础理论与方法·

逆概率权重法在诊断试验评价证实偏倚中的应用

康乐妮 张韶凯 赵方辉 乔友林

【导读】 如何估计和校正筛查或诊断试验中存在的证实偏倚，文中通过宫颈癌筛查实例，采用逆概率权重法R软件CompareTests校正其灵敏度和特异度，利用随机抽样方法生成新数据，将逆概率权重法计算的灵敏度和特异度，与传统计算方法以及最大似然估计方法计算得到的灵敏度和特异度进行比较。结果表明HPV自检法的真实灵敏度和特异度分别为83.53%(95%*CI*：74.23～89.93)和85.86%(95%*CI*：84.23～87.36)。随机抽样结果显示，传统方法计算的灵敏度和特异度分别为90.48%(95%*CI*：80.74～95.56)和71.96%(95%*CI*：68.71～75.00)，采用逆概率权重法校正后的灵敏度和特异度分别为82.25%(95%*CI*：63.11～92.62)和85.80%(95%*CI*：85.09～86.47)；采用最大似然估计法校正后的灵敏度和特异度分别为80.13%(95%*CI*：66.81～93.46)和85.80%(95%*CI*：84.20～87.41)。表明在复杂抽样情况下，逆概率权重法能够有效校正存在证实偏倚的灵敏度和特异度。

【关键词】 逆概率权重法；证实偏倚；灵敏度；特异度；最大似然估计

Implication of inverse-probability weighting method in the evaluation of diagnostic test with verification bias *Kang Leni, Zhang Shaokai, Zhao Fanghui, Qiao Youlin. Department of Cancer Epidemiology, Cancer Institute of Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing* 100021, China

Corresponding author: Qiao Youlin, Email: qiaoy@cicams.ac.cn

【Introduction】 To evaluate and adjust the verification bias existed in the screening or diagnostic tests. Inverse-probability weighting method was used to adjust the sensitivity and specificity of the diagnostic tests, with an example of cervical cancer screening used to introduce the Compare Tests package in R software which could be implemented. Sensitivity and specificity calculated from the traditional method and maximum likelihood estimation method were compared to the results from Inverse-probability weighting method in the random-sampled example. The true sensitivity and specificity of the HPV self-sampling test were 83.53% (95%*CI*: 74.23-89.93) and 85.86% (95%*CI*: 84.23-87.36). In the analysis of data with randomly missing verification by gold standard, the sensitivity and specificity calculated by traditional method were 90.48% (95%*CI*: 80.74-95.56) and 71.96% (95%*CI*: 68.71-75.00), respectively. The adjusted sensitivity and specificity under the use of Inverse-probability weighting method were 82.25% (95% *CI*: 63.11-92.62) and 85.80% (95% *CI*: 85.09-86.47), respectively, whereas they were 80.13% (95%*CI*: 66.81-93.46) and 85.80% (95%*CI*: 84.20-87.41) under the maximum likelihood estimation method. The inverse-probability weighting method could effectively adjust the sensitivity and specificity of a diagnostic test when verification bias existed, especially when complex sampling appeared.

【Key words】 Inverse-probability weighting method; Verification bias; Sensitivity; Specificity; Maximum likelihood estimation

疾病筛查或诊断过程中，常使用生物标志物[1,2]。而评价其检测效果，通常采用金标准验证，但由于价格昂贵，且一些方法具有创伤(如组织活检)，存在一定风险，因此需要一定的抽样策略而非所有标本均进行金标准检测。一般而言，生物标志物检测阳性的个体采用金标准检测验证的比例要大于阴性个体。如果仅将经过金标准验证的个体用来评价生物标志物的效果，并以此计算灵敏度和特异度，可能产生证实偏倚(verification bias)，影响对生物标志物的效果评价[3-5]。为此介绍一种对灵敏度和特异度计算过程中出现证实偏倚的校正方法—逆概率权重法，同时采用传统的校正证实偏倚方法—最大似然

DOI：10.3760/cma.j.issn.0254-6450.2014.03.025

作者单位：100021 北京协和医学院/中国医学科学院肿瘤医院肿瘤研究所流行病室

通信作者：乔友林，Email：qiaoy@cicams.ac.cn

估计法进行灵敏度和特异度校正，并将其结果与逆概率权重法的结果比对。

基本原理

假定诊断试验中，所有个体均进行A检测（待评价的检测方法）和B检测（金标准检测方法），且两检测方法均包含I个检测分类结果。根据一定的人群特征（例如A检测结果的信息），总人群可划分为S个抽样层。如果试验中每个个体均经过A、B两种检测，则每个可能出现的检测结果个数为N_{ijs}，i，$j=1,\cdots,I$；$S=1,\cdots,S$，其中i代表A检测结果，j代表B检测结果。如果将S个抽样层的检测结果合并，高维列联表$I\times I\times S$将转化为$I\times I$表，其中每个单元格的个数$N_{ij}=N_{ij+}$。因此诊断试验的灵敏度$S_e=N_{22}/N_{2+}$，特异度$S_p=N_{11}/N_{1+}$。

在二次抽样存在的条件下，假定列联表的每个格子数为n_{ijs}。由于所有个体均经过A检测，所以经过A检测的个体数目可以表示为N_{+js}。然而，仅有一部分抽取到的个体做了B检测，因此N_{i+s}难以观测。在特定抽样层S内，每列j均包含一个特定抽样率π_{js}，即N_{js}中被抽中做B检测的比例。此时，N_{ijs}的估计值$\hat{N}_{ijs}=n_{ijs}w_{js}$，其中$w_{js}=1/\pi_{js}$。将$S$个抽样层合并后，可以得到每个单元格$N_{ij}$的估计值$\hat{N}_{ij}=\sum_{S=1}^{S}\hat{N}_{ijs}$，而B检测的边际估计值$\hat{N}_{i+}=\sum_{j=1}^{I}\hat{N}_{ij}$。

在列联表中，每个单元格$\hat{N}_{ij}$的方差可表示为

$$Var(\hat{N}_{ij})=\sum_{S=1}^{S}Var(\hat{N}_{ijs})=\sum_{S=1}^{S}Var(n_{ijs})w_{js}^2$$

式中，n_{ijs}服从多项式$(n_{+js},p_{1s},\cdots,p_{Is})$，其中$\hat{p}_{is}=n_{ijs}/n_{+js}$。故$Var(n_{ijs})=n_{+js}p_{is}(1-p_{is})$。由于不同列之间相互独立，因此协方差$Cov(\hat{N}_{ij},\hat{N}_{ij'})=0$。在同一抽样层的相同列之间，数值具有共变性，其协方差为

$$\begin{aligned}Cov(\hat{N}_{ij},\hat{N}_{i'j})&=Cov(\sum_{S=1}^{S}\hat{N}_{ijs},\sum_{S=1}^{S}\hat{N}_{i'js})\\&=\sum_{S=1}^{S}Cov(\hat{N}_{ijs},\hat{N}_{i'js})=\sum_{S=1}^{S}-n_{+js}p_{is}p_{i's}w_{js}^2\end{aligned}$$

据此推导可得出

$$\begin{aligned}Cov(\hat{N}_{i+},\hat{N}_{i'+})&=Cov(\sum_{j=1}^{I}\hat{N}_{ij},\sum_{j'=1}^{I}\hat{N}_{i'j'})\\&=\sum_{j=1}^{I}Cov(\hat{N}_{ij},\hat{N}_{i'j})\end{aligned}$$

或

$$Var(\hat{N}_{i+})=\sum_{j=1}^{I}Var(\hat{N}_{ij})$$

使用delta方法[6]，在结局标量为2分类的条件下，可得到灵敏度和特异度的方差分别为

$$Var(sen)=\frac{[(1-sen)^2Var(\hat{N}_{22})+(sen)^2Var(\hat{N}_{21})]}{\hat{N}_{2+}^2}$$

$$Var(spe)=\frac{[(1-spe)^2Var(\hat{N}_{11})+(spe)^2Var(\hat{N}_{12})]}{\hat{N}_{1+}^2}$$

由此，可以计算出逆概率权重法校正后的灵敏度、特异度及其95%CI，计算过程可借助R软件CompareTests完成。

实例分析

采用1999年中国医学科学院肿瘤医院开展的一项多种方法联合筛查宫颈癌项目的数据[7]，即1 988名研究对象均有金标准证实结果（表1），故计算HPV自检结果完整人群的真实灵敏度和特异度分别为83.53%（95%CI：74.23～89.93）和85.86%（95%CI：84.23～87.36）。本文选择其中醋酸染色后肉眼观察（VIA）和HPV自检两种筛查方法举例说明，将两种方法并联（即任何一种筛查方法阳性者视为阳性人群，两种筛查方法均为阴性者视为阴性人群），按照阳性人群80%的回访比例，与阴性人群20%的抽样比例进行金标准验证，随机生成数据。表2为随机抽样下HPV自检法与疾病状态的关系。本文以下分析均基于此次抽样结果产生的数据。

表1　HPV自检和病理诊断结果（真实情况）

HPV自检	病理诊断		
	≥CIN2	<CIN2	合计
+	71	269	340
-	14	1 634	1 648
合　计	85	1 903	1 988

注：CIN宫颈上皮细胞非典型增生

表2　HPV自检结果与疾病状态（随机抽样）

证实情况	疾病状态	诊断试验		
		HPV自检+	HPV自检-	合计
金标准确认	≥CIN2	57	6	63
	<CIN2	219	562	781
	计	276	568	844
未经金标准确认		64	1 080	1 144
合　计		340	1 648	1 988

1. 传统计算方法：根据传统计算方法灵敏度为90.48%（95%CI：80.74～95.56），计算特异度采用两种处理方式：①仅分析经金标准确认的个体，其特异度为71.96%（95%CI：68.71～75.00）；②将未经金标准确认的检测结果阴性的个体视为真阴性，由此计

算得到的特异度为88.23%(95%*CI*:86.69~89.62)。

2. 逆概率权重法:应用R软件CompareTests进行灵敏度和特异度计算[8]。该软件是针对存在证实偏倚的诊断试验进行评价,得到基于逆概率权重法校正后的诊断试验准确性评价指标,如灵敏度和特异度等。CompareTests的使用语句:CompareTest(stdtest, sampledtest, strata=NA, goldstd="sampledtest")中,stdtest是指初筛时使用的检测方法,即待评价的方法;sampledtest是指对于二次抽样标本的检测,即金标准检测,如某个体未进行金标准检测,结果记录为NA;strata是指研究中所使用的抽样层(以此描述抽样方法,可根据实际需要设定不同的抽样比例),如无抽样层,则设置为NA;goldstd用于指定何种方法为金标准检测。因此整理数据库需要生成3个变量:待评价方法检测结果、金标准检测结果和抽样层。

2种检测方法共有4种不同组合,每种组合的抽样比例和检出病例情况不同,本实例计算中将数据划分为了4个抽样层,每层抽样比例和≥CIN2例数见表3。

表3 不同抽样层的比例和≥CIN2例数

VIA	HPV自检	实际人数	抽样人数	抽样比例	≥CIN2例数	
					观测时	校正后
-	-	1 246	250	0.20	2	10
-	+	193	155	0.80	13	16
+	-	402	318	0.79	4	5
+	+	147	121	0.82	44	54
合计		1 988	844	0.42	63	85

使用R程序计算:

```
a<-read.csv('hpv.csv')
library("CompareTests")
a.com<-CompareTests(a$stdtest, a$sampledtest, a$strata, goldstd="sampledtest")
```

根据计算结果可知校正后的灵敏度和特异度分别为82.25%(95% *CI*: 63.11~92.62)和85.80%(95%*CI*:85.09~86.47)。R给出经过校正后的列联表见表4。

表4 经逆概率权重法校正后的列联表

HPV自检	病理诊断		
	≥CIN2	<CIN2	合计
+	70	270	340
-	15	1 633	1 648
合计	85	1 903	1 988

3. 最大似然估计法:国内外学者对存在证实偏倚条件下,采用最大似然估计对诊断试验准确性评价指标进行校正的方法进行研究。Litde和Rubin[9]提出随机缺失(missing at random, MAR)概念,即二次验证检测(常为金标准检测)人群的选择仅依赖于初次诊断试验的结果。Zhou[10]给出了灵敏度和特异度最大似然估计值及其对应的方差。计算公式见文献[11-13]。根据最大似然估计计算结果,实例数据校正后的灵敏度和特异度分别为80.13%(95%*CI*:66.81~93.46)和85.80%(95%*CI*:84.20~87.41)。

讨 论

证实偏倚与二次抽样试验人群的选择有关,人群选择与初次诊断结果的关联性越大,证实偏倚就越大[14]。因此为避免证实偏倚,首先应确定合理的抽样层。设定抽样层应充分利用诊断试验中可能影响其效果的因素[15-17]。如某项诊断试验中研究对象年龄跨度较大,且某些特定年龄段人数较少。在二次抽样时,除考虑初次诊断结果外,还应将研究对象按年龄划分为不同抽样层,其优点:提高研究对象中人数分布较多年龄段的抽样率,以保证占据信息优势的群体尽量多进入二次验证;人数分布较少的年龄段也能保证进入二次验证,使样本更具代表性。

逆概率权重法正是基于抽样层的考虑,为评价和校正证实偏倚提供的新方法。本文实例分析中真实的灵敏度和特异度分别为83.53%和85.86%。而存在随机缺失和抽样的情况下,采用传统方法计算的灵敏度和特异度为90.48%和71.96%;经逆概率权重法校正后,灵敏度为82.25%,特异度为85.80%,均更接近于真实值。说明存在证实偏倚时,采用传统计算方法对灵敏度和特异度的估计存在一定偏差,可能影响对检测方法的正确评价,而采用逆概率权重法校正后,灵敏度和特异度与真实值之前的偏差降低。本文还使用最大似然估计法校正实例数据中存在的证实偏倚,校正后的灵敏度和特异度分别为80.13%和85.80%,与逆概率权重法类似,与真实值的偏差也低于传统计算方法。因此这两种方法均可用于对证实偏倚进行校正,但由于最大似然估计法在校正时未考虑抽样层的因素,故逆概率权重法更适用于评价多种方法联合检测研究中的证实偏倚。如本文实例中将逆概率权重法的分层变量仅按照HPV自检法的结果分层(即分为阳性和阴性两层,不考虑VIA的影响)重新计算,得到校正的灵敏度和特异度分别为80.13%(95% *CI*: 63.78~90.23)和85.80%(95%*CI*:85.05~86.53),均与最大似然估计法计算结果相似。有研究表明,最大似然估计法其

计算效率高，而在权重变异范围太大时不宜采用逆概率权重法[8]。在抽样层较为复杂的条件下，最大似然估计法需要同时估计较多的参数值，而运用逆概率权重法相对适宜。但应注意，应用最大似然估计法需要满足随机缺失机制，反之计算结果可能有偏[5]。而逆概率权重法主要基于抽样层进行估计，因此实际应用中需要选择合适的抽样层，且每层抽样比例不能过低，否则可能产生偏性影响结果。

总之，逆概率权重法能够有效校正证实偏倚下的灵敏度和特异度，且简便易行，值得推广应用。

参 考 文 献

[1] Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer[J]. J Natl Cancer Inst, 2001, 93(14): 1054-1061.

[2] Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests[J]. J Natl Cancer Inst, 2009, 101(16): 1116-1119.

[3] Begg CB. Biases in the assessment of diagnostic tests[J]. Stat Med, 1987, 6(4): 411-423.

[4] Zhou XH. Effect of verification bias on positive and negative predictive values[J]. Stat Med, 1994, 13(17): 1737-1745.

[5] Alonzo TA. Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests[J]. Stat Med, 2005, 24(3): 403-417.

[6] Korn EL, Graubard BI. Analysis of health surveys[M]. John Wiley & Sons, 1999.

[7] Qiao YL, Zhang WH, Li L, et al. A cross-sectional comparative trial of multiple techniques to detect cervical neoplasia[D]. Acta Acad Med Sin, 2002, 24(1): 50-53. (in Chinese)
乔友林，章文华，李凌，等. 子宫颈癌筛查方法的横断面比较研究[D]. 中国医学科学院学报，2002，24(1)：50-53.

[8] Katki HA, Li Y, Edelstein DW, et al. Estimating the agreement and diagnostic accuracy of two diagnostic tests when one test is conducted on only a subsample of specimens[J]. Stat Med, 2012, 31(5): 436-438.

[9] Litde R, Rubin D. Statistical analysis with missing data[M]. New York: John Wiley and Sons, 1987.

[10] Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias[J]. Communications in statistics-Theory and Methods, 1993, 22: 3177-3198.

[11] Su CJ, Min J, Liu P, et al. Comparison the correcting for verification bias in studies of a diagnostic test[J]. Chin J Health Stat, 2007, 24(2): 143-145. (in Chinese)
苏春娟，闵捷，刘沛，等. 诊断试验证实偏倚校正方法的比较研究[J]. 中国卫生统计，2007，24(2)：143-145.

[12] Xie HB, Tong Y, Peng X, et al. The accuracy of lung cancer diagnosed by PET/CT and the study of verification bias correction [J]. Chin J Med Imag, 2010, 18(6): 529-531. (in Chinese)
谢和宾，童瑶，彭翔，等. PET/CT诊断肺癌的准确度及其证实偏倚校正的研究[J]. 中国医学影像学杂志，2010，18(6)：529-531.

[13] Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients[J]. Invest Radiol, 1985, 20(7): 751-756.

[14] Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy[J]. Stat Methods Med Res, 1998, 7(4): 337-353.

[15] Pickles A, Dunn G, Vzquez-Barquero JL. Screening for stratification in two-phase('two-stage') epidemiological surveys [J]. Stat Methods Med Res, 1995, 4(1): 73-89.

[16] Irwig L, Glasziou PP, Berry G, et al. Efficient study designs to assess the accuracy of screening tests[J]. Am J Epidemiol, 1994, 140(8): 759-769.

[17] McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value[J]. Stat Med, 2002, 21(23): 3609-3625.

（收稿日期：2013-09-29）
（本文编辑：张林东）

·书讯·

《社会心理流行病学》现已出版

由山西医科大学曲成毅教授主编的《社会心理流行病学》一书已由人民卫生出版社出版发行。本书是我国首部关于社会心理流行病学研究方法的专著，分上下两篇。上篇重点介绍基本原理和方法，包括人群调查研究设计（现场观察、现场调查、档案研究、分析性研究、实验研究等）、问卷及量表的编制技术、心理现象的测量及生物学指标（分子生物学、电生理、代谢、内分泌等）、资料分析方法（通径分析、潜变量交互作用分析、支持向量机模型分析等）；下篇介绍具有代表性的社会心理流行病学研究成果，包括突发事件与心理应激、认知障碍、抑郁、自杀、成瘾、暴力、性心理、职业压力等。力图以通俗简明的方式突出这一学科领域思维和方法的特色，实际工作者带着这本书就能在现场开展有关社会心理流行病学领域的研究。本书的读者应当具有基本流行病学理论知识，主要读者群是医学、心理学、社会学、人口学及其他相关学科教学、科研人员及基层实践人员，相关领域的研究生和本科生会对本书更感兴趣。

本刊编辑部