·临床研究与循证医学·

诊断试验准确性研究设计及临床应用

张丽帆[1,2,3]，刘晓清[1,2,3]

中国医学科学院 北京协和医学院 北京协和医院 [1] 感染内科 [2] 临床流行病学教研室，北京 100730
[3] 国际临床流行病学网临床流行病学单位，北京 100730

通信作者：刘晓清 电话：010-69155087，E-mail：liuxq@pumch.cn

【摘要】新的诊断方法在临床开展之前，必须经由严格设计的诊断试验准确性研究进行评价。诊断试验准确性研究设计包括应用 PICOS（P：Patient；I：Intervention；C：Comparison；O：Outcome；S：Study design）原则构建研究问题、确定诊断金标准、选择具有代表性的研究对象、估算样本量、同步盲法比较诊断试验与金标准结果、确立最佳截点值、评价诊断准确性以及遵循诊断准确性研究报告规范进行论文报告 8 个方面。诊断试验的准确性指标包括灵敏度、特异度、预测值和似然比。其中，诊断试验的似然比可帮助医生从验前概率获得验后概率。当医疗环境与研究环境相似、收治患者符合研究入组标准时，应用诊断试验研究的似然比有助于对目标疾病进行诊断与鉴别诊断。

【关键词】诊断试验；研究设计；临床应用
【中图分类号】R-1 【文献标志码】A 【文章编号】1674-9081(2020)01-0096-06
DOI：10.3969/j.issn.1674-9081.20190276

Study Design and Clinical Practice of Diagnostic Accucary Test

ZHANG Li-fan[1,2,3], LIU Xiao-qing[1,2,3]

[1]Department of Infectious Diseases, [2]Department of Clinical Epidemiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China
[3]Clinical Epideniology Unit, International Clinical Epidemiology Network, Beijing 100730, China

Corresponding author: LIU Xiao-qing Tel: 86-10-69155087, E-mail: liuxq@pumch.cn

【**Abstract**】New diagnostic methods must be evaluated by rigorously designed diagnostic accuracy studies before clinical implementation. Designing a diagnostic accuracy study includes 8 procedures: constructing the research question with the PICOS (P: Patient; I: Intervention; C: Comparison; O: Outcome; S: Study design) framework, identifying an appropriate gold standard, choosing a representative patient sample, estimating the sample size, interpreting results of diagnostic tests and the gold standard blind to the other, setting up the optimal threshold, evaluating the diagnostic accuracy, and finally drafting a report according to the standards for reporting diagnostic accuracy. The accuracy of diagnostic tests includes sensitivity, specificity, predictive value (PV), and likelihood ratio (LR). The LR estimated by diagnostic tests can move clinicians from the pretest probability to a posttest probability. If the clinical setting is similar to that of the study and the patient meets all eligibility criteria of the study, the LR may facilitate the diagnostic process in clinical practice.

【**Key words**】diagnostic test; study design; clinical practice

Med J PUMCH, 2020,11(1):96-101

基金项目：北京协和医学院青年教师培养项目（2014zlgc0742）；北京协和医学院研究生教育教学改革项目（10023201600109）
利益冲突：无

准确及时地诊断，是有效治疗的前提。诊断试验可为疾病正确诊断及鉴别诊断提供重要证据。广义的诊断试验涉及以下内容：（1）临床资料，如病史、症状、体征；（2）实验室检查，如生化、免疫学、病原学、病理学检查等；（3）影像学检查，如X线、超声、CT、MRI等；（4）特殊器械检查，如心电图、内镜等。

随着医学技术的发展，新的诊断方法不断涌现。理想的诊断方法除具备精确性和准确性之外，还应快速、简便、安全、经济。任何新的诊断方法在临床开展之前，必须经由科学设计的诊断试验准确性研究进行严格评价。此外，如何解读诊断试验准确性研究的结果，并应用于疾病的辅助诊断，亦是临床医生关注的问题。本文将介绍如何进行诊断试验准确性研究设计，以及在临床实践中如何合理应用诊断试验准确性研究证据。

1 诊断试验准确性研究设计

1.1 构建研究问题

诊断试验准确性研究的问题来源于临床，其结果也将应用于临床，为临床实践提供证据。在提出临床问题时，可采用PICOS（P：Patient；I：Intervention；C：Comparison；O：Outcome；S：Study design）原则将其转化为科学问题。诊断试验准确性研究中，P为疑诊某病的患者；I为待评价的诊断试验；C为诊断金标准；O为诊断准确性评价，包括灵敏度、特异度[1]、预测值（predictive value，PV）[2]和似然比（likelihood ratio，LR）[3]。

诊断试验准确性研究初期，可采用病例对照研究设计，以确诊某病的患者作为病例组，排除某病的患者作为对照组[4]。值得注意的是，Meta分析显示，病例对照研究设计可能高估诊断试验的准确性[5]，因而评价其临床应用价值时，应采用横断面或队列研究设计，同期纳入疑诊某病的连续病例或按比例抽样的随机样本（图1）。相对于横断面或队列研究设计，病例对照研究易于开展，成本较低，其结果可提示我们该诊断试验是否值得进一步研究，避免造成资源浪费。

1.2 确定诊断金标准

诊断金标准亦称标准诊断，是指目前临床医学界公认的最为准确可靠的诊断方法，其确立应结合临床具体情况。常用的金标准包括：（1）实验室检查、细菌培养（病原学诊断）等；（2）手术探查、组织活检、尸体解剖（病理学诊断）等；（3）特殊影像诊断；（4）公认的综合诊断标准，如系统性红斑狼疮等；（5）长期随访的肯定诊断，如慢性胰腺炎等；（6）权威医疗机构颁布的诊断标准，如重症急性呼吸综合征诊断标准等。

应用金标准的目的是将疑诊某病的患者准确地区分为“有病”或“无病”，在同期同条件下进行待评价的诊断试验检测，并与金标准比较，评价诊断试验的准确性。

需注意的是，临床研究中的金标准仅是目前“公认”的，随着对疾病的认识和医疗技术的发展，也可能随之变化[6]。此外，金标准不可包括待评价的诊断试验，否则可增加金标准与诊断试验的一致性，导致加和偏倚（incorporation bias）[7]。

1.3 选择研究对象

诊断试验的价值在于能否在具有相似临床表现的疑诊患者中，正确识别出目标疾病患者。临床诊疗过程中，医生所接诊的患者可能包含目标疾病的各种类型，如不同病情严重程度（轻、中、重）、不同病程阶段（早、中、晚）、不同症状和体征（典型、不典型）、是否经过治疗、有无并发症等。而需要与之鉴别的患者，往往具有相似的临床特征，易与目标疾病

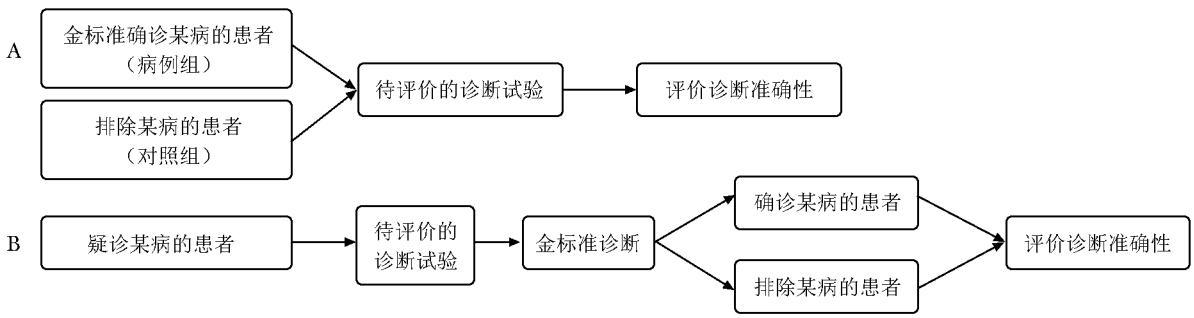

图1 诊断试验准确性研究设计模式图

A. 病例对照研究设计；B. 横断面或队列研究设计

混淆[8]。因此，选择研究对象时，应包括上述所有患者，以保证足够的代表性。当纳入患者不具有代表性时，可导致疾病谱偏倚（spectrum bias）[9]。

选择诊断明确的患者和健康人作为研究对象仅适用于诊断试验准确性评价的初期，一方面，诊断试验识别疾病晚期或病情严重患者的效力可能优于疾病早期或病情轻微的患者；另一方面，医生几乎无须用诊断试验区分健康人与已确诊的严重疾病患者。因此，如果选择严重疾病患者和健康人作为研究对象构成病例对照研究设计，会高估诊断试验的准确性[10]。在临床工作中，纳入连续疾病谱的患者对获得准确的灵敏度和特异度估计极其重要，而这一点非常容易被研究者忽略。

研究对象纳入和排除标准的确定应结合临床实际，根据构建的研究问题定义目标总体的主要特征，注意外推性的同时兼顾可行性。

1.4 估算样本量

诊断试验准确性研究样本量的大小与下列参数有关：（1）显著性水平 α，α 值越小，所需样本量越大。α 通常取 0.05；（2）容许误差 δ，δ 值越小，所需样本量越大，δ 通常取 0.05~0.10；（3）灵敏度或特异度的估计值，用灵敏度的估计值计算病例组样本量，用特异度的估计值计算对照组样本量。

样本量的计算公式：$n=U_{\alpha}^{2}P(1-P)/\delta^{2}$

公式中 U_{α} 为正态分布中累积概率为 α/2 时的 U 值（$U_{0.05}=1.960$，$U_{0.01}=2.576$），δ 为容许误差，P 为灵敏度或特异度的估计值。此外，诊断试验样本量还可通过 LR、受试者工作特征（receiver operator characteristic，ROC）曲线下面积等参数进行估算[11-12]。

1.5 同步盲法比较诊断试验与金标准结果

进行诊断试验准确性研究时，所有研究对象均应接受金标准诊断和待评价试验检测，与金标准的结果应在同样的病例中获得，且尽可能同步进行。如果二者间隔时间过长，则病例的状态可能会发生改变。此外应使用盲法独立评价诊断试验与金标准的结果，以预防偏倚、先入为主以及检验以外的其他信息对判断的影响。诊断试验准确性研究的盲法是指待评价诊断试验的结果判断者不应知道金标准结果，即不应知道研究对象是“有病”还是“无病”；金标准结果判断者不应知道待评价诊断试验的结果。研究显示，未使用盲法可能高估准确性[13]。此外，为避免测量偏倚，诊断试验与金标准的判断者应对其他临床信息或检测结果不知情[9]。例如，评价胸片诊断肺部结节的准确性，若读片者事先看到了患者胸部 CT 上的结节影，可能会先入为主，读片更加仔细，甚至在同一部位发现之前忽略的结节影。

1.6 确立最佳截点值

评价诊断试验准确性时，需将试验结果按照阳性和阴性进行分类，故需要一个判断标准。许多诊断试验，尤其是实验室检测，其测量结果多为连续性变量。对于连续性变量，需要选择区分正常与异常的截点值（cut-off point），即界值。诊断试验中确定最佳截点值的方法包括：（1）均数±标准差法：当测量值为正态分布时，双侧正常值范围常用“均数±1.96 标准差”界定；单侧则用“均数+1.64 标准差”或“均数-1.64 标准差”界定。（2）百分位数法：当测量值为偏态分布或分布类型尚不能确定时，双侧正常值范围常用“$P_{2.5}$~$P_{97.5}$”界定；单侧用“P_{95}”或“P_{5}”界定。（3）ROC 曲线法：诊断试验的结果为连续性变量时，依照不同截点值可分别计算出灵敏度和特异度，以诊断试验的灵敏度为纵坐标、以 1-特异度为横坐标，绘制成连续曲线，即为 ROC 曲线（图 2）。ROC 曲线上最靠近左上方的点对应的截点值即为最佳截点值。ROC 曲线下面积反映了诊断试验的准确性，取值范围在0.5~1.0。ROC 曲线越向左上偏，曲线下面积越大，该诊断试验的准确性越高。因此，除可用于确立截点值外，ROC 曲线还可比较两个或以上独立诊断试验的准确性，如图 2 中诊断试验 1 的准确性优于诊断试验 2。ROC 曲线简单、直观，是确定诊断试验截点值较为常用的方法。（4）结合专业实际进行临床判断：按照大量临床观察或系列追踪观察某些致病因素对健康损害的阈值，作为诊断正常水平的界值。

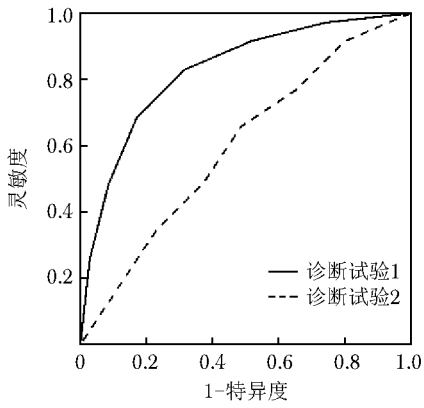

图 2 受试者工作特征曲线

1.7 绘制四格表，评价诊断准确性

依据金标准诊断可将研究对象划分为“有病”或“无病”；依据待评价诊断试验的结果可将研究对象划分为检测“阳性”或检测“阴性”。以金标准诊断为列，待评价的诊断试验结果为行，可绘制四格表（表1）。

表1 诊断试验四格表

诊断试验	金标准诊断		合计
	有病	无病	
阳性	真阳性（a）	假阳性（b）	a+b
阴性	假阴性（c）	真阴性（d）	c+d
合计	a+c	b+d	a+b+c+d

a. 真阳性，指金标准诊断为“有病”且诊断试验结果是“阳性”的例数；b. 假阳性，指金标准诊断为“无病”但诊断试验结果是“阳性”的例数；c. 假阴性，指金标准诊断为“有病”但诊断试验结果是“阴性”的例数；d. 真阴性，指金标准诊断为“无病”且诊断试验结果是“阴性”的例数

灵敏度（以Sen表示），即真阳性率，是金标准诊断为“有病”的研究对象中，诊断试验结果是“阳性”的比例，反映了诊断试验识别疾病的能力。灵敏度只与病例组有关，Sen = a/（a+c）。特异度（以Spe表示），即真阴性率，是金标准诊断为“无病”的研究对象中，诊断试验结果是“阴性”的比例，反映识别无病的能力。特异度只与对照组有关，Spe=d/（b+d）。一项汇总了23项Meta分析的研究显示，诊断试验的灵敏度和特异度会随疾病患病率而变化，特异度会随着患病率的升高而降低[14]。灵敏度和特异度是诊断试验的重要指标，但无法帮助临床医生估计单个患者的疾病概率[15]。

PV，是应用诊断试验的结果来估计研究对象有病或无病可能性的大小。阳性预测值（positive PV，PPV）是诊断试验结果为阳性者中“有病”者所占的比例，PPV = a/（a+b）；阴性预测值（negative PV，NPV）是诊断试验结果为阴性者中“无病”者的比例，NPV=d/（c+d）。预测值可用于估计疾病的概率，但会随患病率的变化而变化。因此，当临床医生面临的患者群体与已发表文献中研究对象的患病率不同时，不可将文献中的预测值数据直接应用于自己的患者[15]。

LR，是诊断试验的某种结果（阳性或阴性）在“有病”组中出现的概率与“无病”组中出现的概率之比。是患者“有病”与“无病”概率的比值。阳性似然比（positive LR，PLR）是真阳性率和假阳性率的比值，PLR = Sen/（1−Spe）= ［a/（a+c）］/［b/（b+d）］；阴性似然比（negative LR，NLR）是假阴性率和真阴性率的比值，NLR =（1−Sen）/Spe =［c/（a+c）］/［d/（b+d）］。似然比利用了诊断试验的全部信息，不受患病率影响，可用于估计单个患者的疾病概率[16]。

1.8 论文报告

诊断试验的结果解释应结合临床实际，结论要客观真实。推荐遵循诊断准确性研究报告规范（Standards for Reporting of Diagnostic Accuracy，STARD）进行论文报告。STARD于2003年发表，旨在提高诊断试验的报告质量[17]；2015年发布了更新版本，对2003版STARD的清单条目和流程图进行了修订增补[18]。其中文译文和相关解读也已发表[19-21]。

2 诊断试验结果的临床应用

对于临床医生而言，非常重要的问题是：如何将某项诊断试验准确性研究的结果应用于自己的患者？回答这个问题之前，需要明确两点：（1）该诊断试验的结果是否准确可靠？如研究问题明确、设计科学严谨、金标准和研究对象选择合理、采用盲法、检测结果稳定可重复，则较为准确可靠。可应用诊断试验准确性研究的质量评价工具（Quality Assessment of Diagnostic Accuracy Studies，QUADAS）对偏倚风险进行评估[22-23]。（2）该诊断试验是否适用于自己的患者？如所处的医疗环境与该诊断试验实施的环境相似，且患者符合该研究的纳入标准，则较为适用。

2.1 由验前概率获得验后概率

当医生接诊一例患者，综合病史、体格检查以及已有的化验结果会形成初步诊断，此时临床分析估计所得的疾病概率称之为验前概率（pretest probability）[25]，在此基础上，医生进行某项诊断试验，检测结果可能会提高或降低初步诊断的可能性，此时的疾病概率称之为验后概率（posttest probability）。诊断试验的LR[3]帮助医生从验前概率得到验后概率，LR的大小表明某个诊断试验的结果将会提高或降低目标疾病验前概率的程度。

应用LR由验前概率获得验后概率包括以下几种方法：（1）计算法：验前比值=验前概率/（1−验前概率），验后比值=验前比值×似然比，验后概率=验后比值/（1+验后比值）；（2）诺模图法[26]：左栏代表验前概率，中间栏代表LR，右栏代表验

后概率，将验前概率和 LR 对应的数值连线并延长，即可得到验后概率（图 3）；（3）软件法：通过网址 http://meta. cche. net/clint/templates/calculators/lr_nomogram. asp，输入验前概率及 LR 的数值，即可获得验后概率。

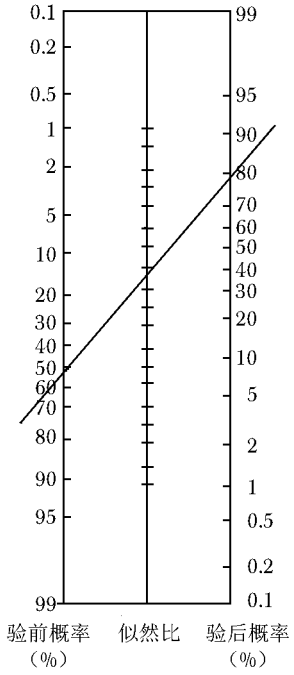

图 3 诺模图
将验前概率与似然比对应的数值连线并延长，即可得到验后概率

2.2 临床场景模拟

患者女性，68 岁，发热 3 个月，峰值体温39 ℃，体温多波动在 37~38. 5 ℃，伴盗汗，午后及夜间为著。伴咳嗽，干咳为主，偶有少量白色泡沫样痰。当地医院查红细胞沉降率为 65 mm/h，C 反应蛋白为 13. 7 mg/L；胸片正常；EB 病毒、巨细胞病毒、支原体、衣原体等相关检测无明显异常。诊断“病毒感染”，予以莫西沙星、感冒清热冲剂等药物治疗，效果不佳。2 年前诊断糖尿病。父亲患陈旧性肺结核。

分析患者病情：老年女性，午后低热伴盗汗，炎症指标升高，既往糖尿病史，父亲患陈旧性肺结核。怀疑患者为结核菌感染，但未发现明确的结核病灶。依据现有证据，估计患者患结核病的概率为 50%。但仍需与其他感染、肿瘤及自身免疫性疾病相鉴别。进一步完善相关检查，结核感染 T 细胞检测的结果为每百万个外周血单个核细胞中存在 680 个斑点形成细胞，提示存在结核感染。除此之外，无其他阳性发现。

患者符合经典型不明原因发热（fever of unknown origin，FUO）定义，总结临床问题：结核感染 T 细胞检测对经典型 FUO 患者的诊断准确性如何？文献检索到 2016 年发表的一项研究——“结核感染 T 细胞检测在结核高流行区对 FUO 病因诊断价值”[27]。通过仔细阅读文献，充分评估偏倚风险，认为该研究结果可信，且研究场所、临床情况、研究人群的年龄性别等描述均与患者相符，研究结果可用于该患者。此项研究中，未获得病原学证据的临床诊断结核病患者结核感染 T 细胞检测的 PLR 为 4. 24。应用诺模图法，获得诊断结核病的验后概率为 81%（图 3）。据此，加用诊断性抗结核治疗，1 个月后患者体温正常，无其他不适，复查显示红细胞沉降率和 C 反应蛋白逐渐降至正常。继续规范抗结核治疗，总疗程 1 年。最终诊断：结核菌感染（部位未明确）。

3 小结

诊断试验准确性研究遵循通用的临床研究设计理念，如 PICOS 原则构建研究问题、选择有代表性的研究对象、估算样本量、采用盲法、依规范进行论文报告等，但也有其独特之处，如确定诊断金标准、确立最佳截点值以及计算诊断准确性参数等。在研究设计过程中，应注意控制偏倚，使得研究结果准确且能够外推。需特别注意的是，在应用 LR 帮助临床诊断时，除评价证据质量外，应充分评估研究结果是否适用于接诊患者，否则可能误导诊断。

参 考 文 献

[1] Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques [J]. Public Health Rep, 1947, 62: 1432-1449.

[2] Vecchio TJ. Predictive value of a single diagnostic test in unselected populations [J]. N Engl J Med, 1966, 274: 1171-1173.

[3] Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios [J]. Lancet, 2005, 365: 1500-1505.

[4] Sackett DL, Haynes RB. The architecture of diagnostic research [J]. BMJ, 2002, 324: 539-541.

[5] Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests [J]. JAMA, 1999, 282: 1061-1066.

[6] Glasziou P, Irwig L, Deeks JJ. When should a new test be-

come the current reference standard? [J]. Ann Intern Med, 2008, 149: 816-822.

[7] Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias [J]. CJEM, 2008, 10: 174-175.

[8] Weiss NS. Control definition in case-control studies of the efficacy of screening and diagnostic testing [J]. Am J Epidemiol, 1983, 118: 457-460.

[9] Whiting PF, Rutjes AW, Westwood ME, et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies [J]. J Clin Epidemiol, 2013, 66: 1093-1104.

[10] Rutjes AW, Reitsma JB, Di Nisio M, et al. Evidence of bias and variation in diagnostic accuracy studies [J]. CMAJ, 2006, 174: 469-476.

[11] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics [J]. J Biomed Inform, 2014, 48: 193-204.

[12] Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies [J]. J Clin Epidemiol, 1991, 44: 763-770.

[13] Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review [J]. Ann Intern Med, 2004, 140: 189-202.

[14] Leeflang MM, Rutjes AW, Reitsma JB, et al. Variation of a test's sensitivity and specificity with disease prevalence [J]. CMAJ, 2013, 185: E537-E544.

[15] Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values [J]. Acta Paediatr, 2007, 96: 338-341.

[16] Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice [J]. Acta Paediatr, 2007, 96: 487-491.

[17] Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative [J]. BMJ, 2003, 326: 41-44.

[18] Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies [J]. BMJ, 2015, 351: h5527.

[19] 王波, 詹思延. 如何撰写高质量的流行病学研究论文第三讲诊断试验准确性研究的报告规范——STARD 介绍 [J]. 中华流行病学杂志, 2006, 27: 909-912.

[20] 朱一丹, 李会娟, 武阳丰. 诊断准确性研究报告规范 (STARD) 2015 介绍与解读 [J]. 中国循证医学杂志, 2016, 16: 730-735.

[21] 孙凤. 医学研究报告规范解读 [M]. 北京: 北京大学医学出版社, 2015: 181-188.

[22] Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews [J]. BMC Med Res Methodol, 2003, 3: 25.

[23] Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies [J]. Ann Intern Med, 2011, 155: 529-536.

[24] Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies [J]. BMJ, 2008, 336: 1106-1110.

[25] Richardson WS. Where do pretest probabilities come from? [J]. Evid Based Med, 1999, 4: 68-69.

[26] Fagan TJ. Letter: Nomogram for Bayes theorem [J]. N Engl J Med, 1975, 293: 257.

[27] Shi X, Zhang L, Zhang Y, et al. Utility of T-Cell Interferon-gamma Release Assays for Etiological Diagnosis of Classic Fever of Unknown Origin in a High Tuberculosis Endemic Area—a pilot prospective cohort [J]. PLoS One, 2016, 11: e0146879.

(收稿日期: 2019-12-11)