ICS 11.100
C 50

WS

中华人民共和国卫生行业标准

WS/T 505—2017

定性测定性能评价指南

Guideline for evaluation of qualitative test performance

2017-09-06 发布　　2018-03-01 实施

中华人民共和国国家卫生和计划生育委员会　　发布

目　　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准的制定参照美国国家临床实验室标准化委员会(NCCLS)EP12-A Vol.22 No.14 User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline 和 EP12-A2 Vol.28 No.3 User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline-Second Edition。

本标准起草单位:国家卫生计生委临床检验中心、广东省中医院、首都医科大学附属北京朝阳医院。

本标准主要起草人:李金明、张括、陈曲波、王露楠、张瑞、谢洁红、王清涛。

引　言

定性测定是临床实验室最为常用的方法之一，广泛作为各种疾病的筛查、诊断和处理手段，但在实际工作中，常因为不同厂家试剂、不同方法甚至不同实验室使用相同试剂或方法得到不一样的结果，影响定性检验结果的临床应用。为保证日常检验结果的一致性和可比性，临床实验室在将相应的定性检验试剂、方法或检测系统用于常规检验前，需对试剂、方法或检测系统进行性能验证或方法学比较评价。由于每个实验室在实验设计、数据分析或结果解释等各方面的侧重点不同，定性测定的方法学评价多种多样。为使实验室的定性测定性能验证或评价有规则可循，从而保证实验室定性检测的质量，特制定本标准。

定性测定性能评价指南

1 范围

本标准规定了定性测定性能的评价方法。本标准描述的定性测定仅限于只有两种检测结果(例如:阳性/阴性、有反应的/无反应的)的方法。对于结果报告为阴性、+1、+2、+3、+4 或滴度的半定量方法,本标准所涉及的精密度和方法学比较试验仅可用于其临界值。

本标准适用于开展各种类型定性检测的实验室,以及需对定性诊断试剂盒检测性能进行研究或描述的生产厂家、监管机构和实验室检查人员。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

NCCLS GP10-A 使用接受者操作特征曲线评价实验室检测临床准确性(Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Charateristics (ROC) plots: Approved Guideline,1995)

3 术语和定义

下列术语和定义适用于本文件。

3.1

偏倚 bias

当两方法或两仪器进行比对实验时,被测方法的测量观测平均值(在重复条件下的测量)与参比方法或参比仪器获得的测定值之间的差异。

3.2

C_5～C_{95}区间 C_5～C_{95} interval

临界值附近的分析物浓度,可认为此浓度区间之外的分析物检测结果始终为阴性(浓度<C_5)或始终为阳性(浓度>C_{95})。

注 1:不精密度的存在使得这一区间之内的检测结果并非始终一致。

注 2:字母 C 是浓度的缩写,下标(5,50 或 95)表示阳性结果的百分率。

注 3:C_5 即仅有 5%被检样品可被判定为阳性时的分析物浓度,C_{95} 即有 95%被检样品可被判定为阳性时的分析物浓度。

3.3

50%浓度 C_{50}

接近临界值的分析物浓度,多次重复检测此浓度的单一样本时将获得 50%的阳性结果和 50%的阴性结果。

3.4

准确度 accuracy

被分析物质的测定结果与真实结果之间的接近程度。

3.5

关注状况　condition of interest

某个对象被关注的某种特定疾病、疾病阶段、健康状况或其他任何可鉴别的状况或特征，例如对临床措施（治疗的起始、改进和终止）有指导意义的某一已知疾病或健康状况的分级。

3.6

诊断准确度　diagnostic accuracy

待评价方法的检测信息和诊断准确度标准的符合程度。

注 1：诊断准确度可以多种形式表达，包括敏感性-特异性配对、似然比配对和接受者工作特征曲线（ROC 曲线）下的面积。

注 2：诊断准确度需在关注状况下，结合特定标准与使用的方法进行阐述。

注 3：诊断准确度不等同于准确度。

3.7

诊断准确度标准　diagnostic accuracy criteria

使用一种方法或联合多种方法，包括实验室检测、影像学检测、病理和随访信息在内的临床信息，来界定状况、事件和关注特征有无的标准。

注 1：诊断准确度标准可随着分析系统的进步而改变，或者在特定的情况下真实诊断可能与管理或权威机构的测定不同。

注 2：诊断准确度标准并不考虑待评价方法（新检测方法）的结果。诊断准确度标准可为一种指定某个选择或为一套方法进行排序运算法则，从而以不同的结果组合来确定最终的阳性/阴性分类。

3.8

分析物　analyte

实验室试验所检测的物质或成分。

3.9

敏感性　sensitivity

真阳性率　true positive rate

在患有明确临床疾病的患者中，诊断性试验检测为阳性或超过决定限例数的比例，反映新试验正确判断是否罹患某种疾病的能力。

该临床疾病应由不依赖于被评价试验的标准来定义。本指标可用于评价测定方法的临床应用价值，理想的测定方法临床诊断敏感性应为 100%。

3.10

特异性　specificity

真阴性率　true negative rate

在没有特定临床疾病的患者中，诊断性试验结果为阴性或在决定限范围内的比例，特异性反映新试验正确排除某病的能力。

注：本指标是用于评价测定方法的临床应用价值，理想的测定方法临床诊断特异性应为 100%。

3.11

质控品　control material

用于质量控制过程的装置、溶液或冻干品。

注 1：质控品中待分析物的预期反应性或浓度在制备和使用过程中确保控制在已知的限值之内。

注 2：质控品通常不用于校准它们本身作为对照的同一试验过程。

3.12

质量控制　quality control；QC

履行质量要求的质量管理部分。

注 1：包括用于满足总质量要求的操作技术和活动。

注 2：在医疗检测中，指监控测定方法和结果以确保检测系统性能的一系列步骤。

3.13

临界点或值　cut off

在定性试验中，临界点是指检测反应的某一点，低于此检测反应点的定性检测结果被判定为阴性，而高于此点则被判定为阳性。

3.14

定性测定　qualitative test

只提供两种反应结果的检测方法(即阳性/阴性或者是/否)。

注：真正的定性检测基于唯一的医学判定值；另外，某些所谓的定性检测来源于二分类定量或者顺序等级。

3.15

假阴性　false negative；FN

一份阳性的样本或一个阳性的患者，在诊断性试验中待检测的成分检测结果为阴性。

3.16

假阳性　false positive；FP

一份阴性的样本或一个阴性的患者，在诊断性试验中待检测的成分检测结果为阳性。

3.17

金标准　gold standard

一个能得到的最接近真实情况的过程或材料。

注：慎用该术语。

3.18

被测对象　measurand

一个特定的待测量的量值对象。

注：这一术语或定义包括所有的量，例如“物质”的浓度就是一个可能与某个特定分析物有关的量。

3.19

不精密度　imprecision

特定条件下得到的独立测定结果的分散程度。数值上可用标准差和变异系数表示。

3.20

真阴性　true negative；TN

一个未患疾病或未处于某种疾病状态的被检对象得到阴性的检测结果。

3.21

真阳性　true positive；TP

一个患有疾病或处于某种健康状态的被检对象得到阳性的检测结果。

3.22

阴性预测值　negative predictive value；NPV

检测为阴性的个体确实未患特定疾病的可能性。理想测定方法的阴性预测值应为 100%，即没有假阴性。

3.23

阳性预测值　positive predictive value；PPV

检测为阳性的个体确实患有特定疾病的可能性。理想测定方法的阳性预测值应为 100%，即没有假阳性。

3.24

流行率　prevalence

与特定的人群中总的成员数相比，患病或处于某种特定健康状态下的人数所占的比例。

3.25

似然比　likelihood ratio

同时反映敏感性和特异性的复合指标。即患病者中得出某一筛查试验结果的概率与未患病者得出这一概率的比值。该指标全面反映筛查试验的诊断价值，且非常稳定。似然比的计算只涉及敏感性和特异性，不受患病率的影响。阳性似然比为真阳性率与假阳性率之比，阴性似然比为假阴性率与真阴性率之比。

3.26

重复性　reproducibility

同一被测对象在不同状态下多次检测的结果符合程度的接近性。

注 1：不同的状态可能包括：检测的原理或方法、观察者、测试仪器、检测地点、操作条件和时间。

注 2：重复性也可能被用于描述定量实验中结果的离散程度。

4　定性测定方法

4.1　概述

临床上，定性测定方法可用于疾病筛查、诊断、确认或者监测。类似于定量方法，定性测定方法的敏感性、特异性、预测值以及疾病或者症状在被检测人群中的流行率决定了其在临床中的应用。

4.2　筛查试验

临床上，筛查方法通常用于检测整个人群（或者人群中的特定的一部分）中特定待测物或因子的存在情况。如粪便隐血检测或性病研究实验室（VDRL）梅毒血清学试验。用于筛查的定性试验应具有高敏感性以确保真正罹患某种疾病的患者被检出。与诊断试验或确认试验相比，筛查试验会产生更多的假阳性结果。筛查试验的低特异性可通过特异性较好的确认试验加以弥补。

4.3　诊断试验

定性试验也用于临床怀疑某种特定疾病或状况是否存在的诊断。如各种微生物培养就是用于检查感染情况的诊断试验。临床上要求对患者疾病进行及时合理的处理，因此，诊断试验需具有良好的敏感性和特异性。诊断试验后如需进行确认试验，对诊断试验的特异性要求可以稍微降低。

4.4　确认试验

确认试验用于验证筛查试验或者诊断试验结果。如果确认试验证实了之前的检测结果，临床医生即可依其做出诊断。可通过设计使确认试验具有较高的特异性（有时甚至以牺牲敏感性为代价）以及高阳性预测值。例如，梅毒密螺旋体抗体荧光吸收试验（FTA-ABS）就是一种用于 VDRL、快速血浆反应素环状卡片试验（rapid plasma reagin test，RPR）、甲苯胺红不加热血清试验（toluidine red unheated serum test，TRUST）、梅毒血清学试验等筛查试验之后的确认试验。

5　性能验证的时机

5.1　实验室在下述情况下，应对测定试剂或系统进行性能验证：

——使用新的检测试剂或系统；

——更换检测试剂或系统；

——检测试剂或系统出现重大改变时，如仪器设备故障维修后、试剂制备用原材料来源改变等。

5.2　对一常规检测项目，实验室一旦使用了一种检测试剂或系统后，不应随意更换。更换通常发生在

下述情况：

——新的检测试剂或系统更易于操作；

——新的检测试剂或系统有更高的检测性能；

——新的检测试剂或系统更经济；

——新的检测试剂或系统更能满足实验室测定要求。

6 性能验证的准备

6.1 性能验证前，应对实验室技术人员进行必要的培训。培训内容包括：

——试剂和仪器设备的熟悉；

——样本的处理和保存；

——试剂的处理和保存；

——合理的检测程序；

——对结果合理的解释；

——系统的质量控制。

在阶段性培训过程中，经过实验演示和实际操作后，每个检测操作者应在下一阶段培训前，证明其对于这一检测试剂方法或系统业已熟练。在继续培训前应对每个操作者是否达到熟练程度，进行相应的书面和/或实验考核确认。

6.2 性能验证前，应制订质量保证计划。为保证性能验证检测过程中结果的一致性，应采用合适的质控品进行质量控制。质控品使用原则如下：

——使用试剂或检测系统的生产厂家提供的质控品时，应遵照相应的使用说明书进行；

——在确保没有基质效应的前提下，也可使用其他稳定的商品化质控品或源自临床标本的自制质控品；

——进行多种试剂方法或检测系统的比较实验时，所有试剂方法或检测系统所使用的质控品应相同；

——日常工作中，多数定性检测试剂或系统使用一个阳性和一个阴性质控品即可，但某些定性检测方法可能需要检测多个质控品；

——某些定性检测提供了量化的结果，评估其检测效果时可采用与定量方法的质控品检测类似的方法；

——质控品的选择可参照临界点附近的检测信息；

注：使用接近临界点的阳性或阴性质控品比极高或极低的质控品更易检测出错误，然而，处于或非常接近临界点的质控品则可能会导致没有显著错误的检测批次因失控而无效。

——如果质控品检测时没有得到预期的结果，则应采取纠正措施。

6.3 性能验证前，应确定用于性能验证的患者样本的数量、阴阳性及用来比较的试剂方法或检测系统等。

7 样本的采集和处理

用于性能验证的临床样本的采集应按照试剂或检测系统生产厂家的说明书中关于样本采集的说明进行。应注意样本采集容器对检测结果可能造成的影响。如特定的试剂方法或系统要求使用新鲜的样本，则样本采集后应尽快进行检测，以减少样本质量对检测结果的影响。如样本运送要求有严格的条件，则应采用适当的运送系统。某些情况下，如对样本的采集时间要求较为严格，则在性能验证过程中应始终保持一致。

因为感染性未知，所有患者或实验室样本均应按照有生物传染危险性样本对待。

8 重复性研究

8.1 概述

与定量检测相似，定性检测同样应考虑偏倚（系统误差）和不精密度（随机误差）。

评价定性检测试剂或系统精密度时，应采用浓度接近临界值的分析物作为检测材料，不宜采用阴性低值或强阳性样本来评价定性检测方法的不精密度。定性测定的临界值由试剂生产厂家依据阳性或阴性样本结果确定。

定性检测的不精密度曲线有助于理解由于不精密度的存在，对同一样本进行多次重复检测可能得到并不完全一致的结果（如阳性或阴性、正值或负值、有或无）这一现象。可用 C_5～C_{95} 区间描述分析物浓度接近 C_{50} 的样本重复检测结果的不一致性（不精密度）。

8.2 精密度偏差来源

待评价方法不同，影响浓度接近临界值样本检测结果的精密度偏差来源也不相同，包括：

——样本种类；

——运输和贮存条件；

——操作人员；

——环境温湿度的差异；

——光照条件；

——批间变异；

——实验安排；

——温育、干燥或判读结果的时间等。

在设计重复性研究方案时需考虑这些偏差来源，必要时可对上述影响检测的实验条件进行明确规定。

8.3 临界点（cut-off）附近的分析物浓度

试剂生产厂家根据检测目的及临床敏感性和特异性建立 cut-off 浓度，cut-off 一旦确立，用户不可随意更改，若检测结果低于 cut-off 则判定为阴性，高于 cut-off 则判定为阳性。

在理想条件下对恰好为 cut-off 浓度的样本进行重复性检测，阴性结果和阳性结果各为 50%。实际操作中，此理想条件不易达到，因此接近 cut-off 的分析物浓度，即出现 50/50 分界点的浓度，每个实验室会有些许差异，但是均称为 C_{50}。以 C_{50} 浓度为基础，逐步增加待测物的浓度并检测，理应获得相应逐步增大的阳性结果百分比和更小的阴性结果百分比。同理，逐步降低待测物浓度，则应得到相反的结果。如果待测物的浓度接近 C_{50}，测定结果将具有不确定性，同一份样本的多次检测结果（阳性或阴性、有或无、有反应或无反应）不可能保持一致。

理想条件下，如果使用浓度恰好等于 cut-off 的样本进行重复性研究，C_{50} 应恰好等于厂家建立的 cut-off。在实际重复性研究中，厂家定义的 cut-off 和方法评价时估计的 C_{50} 之间的差异会导致定性测定的偏差。

8.4 分析物浓度接近 cut-off 时的不精密度曲线

可用“不精密度曲线”表明在规定条件下进行一系列检测时，接近 C_{50} 的分析物得到的阳性和阴性结果百分比是如何随着实际浓度改变而改变的（见图 1）。随着分析物浓度接近 C_{50}，阳性和阴性结果的百分比发生变化。分析物浓度的增加（浓度横坐标右移），阳性结果的百分比升高，而阴性结果的百分比

下降。分析物浓度降低(浓度横坐标左移),阳性结果的百分比降低,而阴性结果百分比升高。

方法不同、实验室不同或在同一实验室采用同一方法但检测条件不同均会影响不精密度曲线的实际形状和陡峭程度。图 2 为两种测定方法的不精密度曲线比较。两种方法的 C_{50} 浓度相同,因此两种方法间不存在系统误差。方法 1 在接近 C_{50} 处的曲线陡峭,因此浓度向任何一个方向稍有改变,将产生全阳性或全阴性结果。方法 2 在近 C_{50} 处比较平滑,所以改变相同浓度将产生更多的阳性和阴性的混合结果。要得到和方法 1 类似的一致阳性或阴性结果,方法 2 需要更大的浓度增量。因此方法 1 在接近 C_{50} 处的精密度高于方法 2。

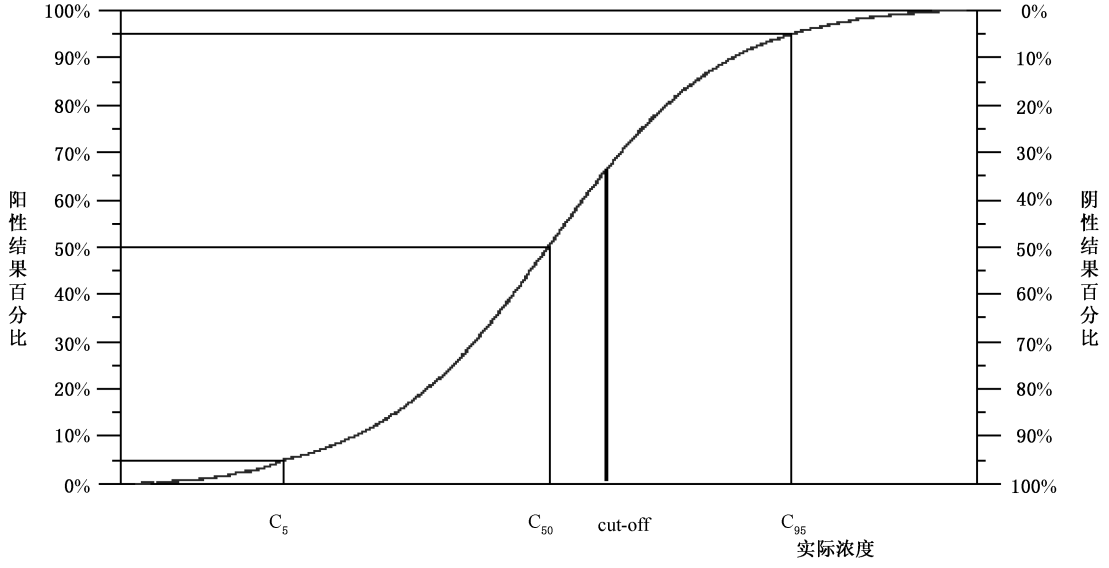

图 1　分析物浓度接近临界值时的不精密度曲线

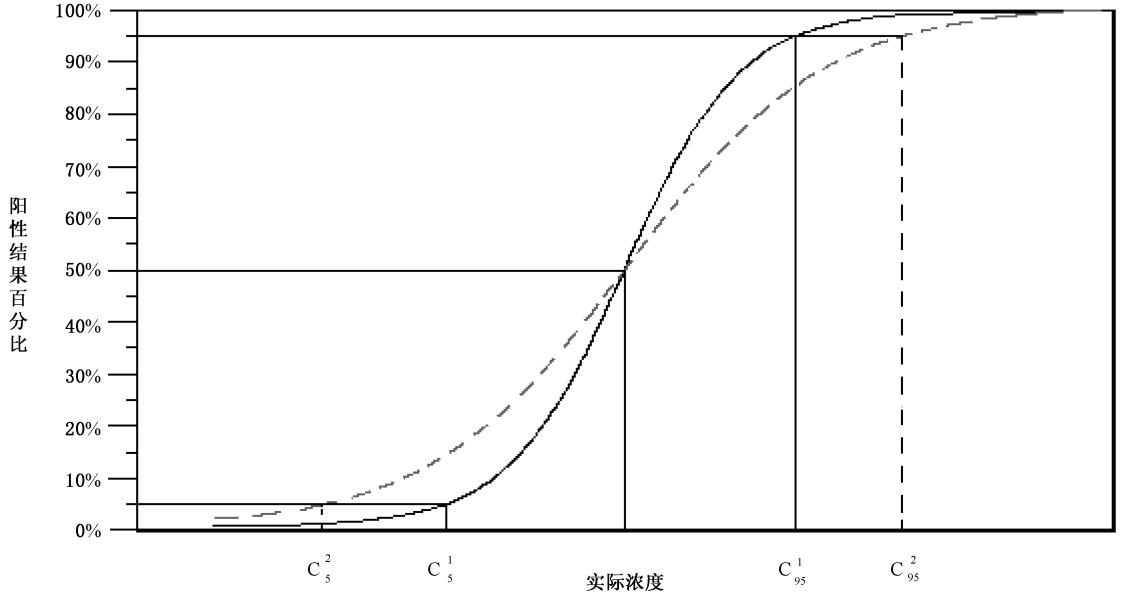

图 2　两种测定方法的不精密度曲线

8.5　C_5～C_{95} 区间

类似于 C_{50} 的定义,检测浓度为 C_5 和 C_{95} 的分析物时将分别产生 5%和 95%的阳性结果。用浓度＜C_5 的样本进行重复性检测时,将持续得到阴性结果,用浓度＞C_{95} 的样本进行重复性检测时,将持续得到阳性结果。而结果的真阳性或真阴性则取决于候选方法的诊断准确性,而诊断准确性则以候选方法

的临床敏感性和临床特异性为特征。

分析物浓度位于 C_5～C_{95} 区间之外(小于 C_5 或大于 C_{95})时,候选方法对同一样本的重复性检测将得到相同结果。C_5～C_{95} 区间越窄,检测方法越好。C_5～C_{95} 区间(≥C_5 且≤C_{95})和同一样本重复检测可获得一致结果时的浓度范围,在使用相同的分析物但采用不同方法检测时可能存在差异。而区分这种差异的能力正是评价定性测定方法性能的一个有用的工具。由于 C_5～C_{95} 区间反映了重复检测可能获得不完全一致结果的浓度范围,因此 C_5～C_{95} 区间的宽度提供了定性检测精密度的相关信息,从图 2 可以看出,方法 2 的 C_5～C_{95} 区间大于方法 1 的 C_5～C_{95} 区间。

注:C_5～C_{95} 之间的浓度范围称为方法的"95%区间",切勿将该术语和 95%置信区间混淆。

8.6 性能验证中试剂盒阴、阳性对照的作用

性能验证评价过程中,每一批次实验都应加入试剂盒的阴性和阳性对照品同时进行检测,只有阴性和阳性对照品的检测结果符合试剂盒的预期要求,才可认定实验数据有效。如方法学比较研究在10 d内完成,则每批次应重复测定对照品,共提供 20 次重复检测结果。如方法比较研究超过 20 d,则每批次应对每份对照品进行单次检测,总计也提供 20 次重复检测结果。

如果阴性和/或阳性对照品未获得预期结果,则这一批次检测结果必须作废,应在当天或另外一天安排新批次的检测以替换不合格的批次。并且,所在实验室应分析造成不合格质控结果的原因。如不合格批次不止一次,所在实验室应中止检测工作,查找原因并采取相应纠正措施。必要时应咨询试剂生产厂家。

8.7 分析物浓度接近 C_{50} 的定性方法精密度试验

8.7.1 概述

进行重复性研究时,理想情况下需绘制在规定条件下候选方法的整个不精密度曲线,然而具体操作时需要检测的样本数量较大,因此,可使用一个简单的方法,即为待评价的检测方法建立分析物的临界浓度,并判定某一特定浓度范围,例如 C_{50} ±20%,是否包含了 C_5～C_{95} 区间。若−20%～+20%浓度范围包含了 C_5～C_{95} 区间,那么 20%或距离 C_{50} 更远浓度的样本将得到一致的检测结果,即在 C_5～C_{95} 区间之外的样本检测结果可认为是精密的,浓度>C_{95},将持续得到阳性结果,浓度<C_5,将持续得到阴性结果。实验室也可根据检验目的和可接受的精密度选择±10%或±30%。

8.7.2 试验步骤

8.7.2.1 确定临界值浓度

特定检测试剂或系统的说明书有可能会注明该试剂或系统的临界值浓度,此时,可用该值作为 C_{50} 的近似值。如不能由此或由其他方法获得临界浓度,可将阳性样本进行系列倍比稀释,然后对其重复检测,以确定能够获得 50%阳性和 50%阴性结果的那个稀释度。处于这一稀释度的待测物浓度即为 C_{50}。

8.7.2.2 制备评价用样本

制备 3 份样本。一份浓度为 C_{50},一份为 C_{50} +20%,一份为 C_{50} −20%。每份样本的体积需保证 40 次或更多次重复检测的需要。

8.7.2.3 评价方法

每份样本检测 40 次,确定每一份样本结果为阳性和阴性的百分比。

注:如检测次数达不到 40 次,结果的统计学意义有限。

8.7.2.4 结果判断

8.7.2.4.1 C_{50}是否准确的判断

根据浓度为C_{50}的样本在40次检测中得到阳性结果的次数判断C_{50}是否准确，见表1。

表1 C_{50}是否准确判断标准

类型	检测次数	阳性结果次数	阳性结果所占百分比	C_{50}准确性判定
1	40	≤13次	≤32.5%	不准确或不可信(统计学的错误率为5%)
		≥27次	≥67.5%	
2	40	14～26次	35%～65%	准确或可信
注：C_{50}可信度取决于实际检测结果以及检测的样本数量。				

如果C_{50}准确，浓度为C_{50}的样本重复检测应获得50%的阳性和50%的阴性结果。可根据检测次数和阳性结果次数的双侧95%可信区间提示阳性结果的真正百分比，得知样本的实际浓度，见表2。

表2 重复性检测总次数与样本的实际浓度

重复性检测总次数	阳性结果			样本的实际浓度
	次数	百分比	真正百分比	
20	10	50%	30%～70%	C_{30}～C_{70}
40	20	50%	35%～65%	C_{35}～C_{65}
100	50	50%	40%～60%	C_{40}～C_{60}

8.7.2.4.2 候选方法的－20%～＋20%浓度范围是否包含了C_5～C_{95}区间

阴性或阳性结果所占比例不同，结论也不同，共4种不同类型，见表3。示例参见附录A。

表3 －20%～＋20%浓度范围是否包含C_5～C_{95}区间

类型	样本浓度	阴性或阳性结果所占比例	结论
1	＋20%	阳性结果≤87.5%(35/40)	－20%～＋20%浓度范围在C_5～C_{95}区间之内； 用该方法检测，浓度超过C_{50}±20%的样本检测结果不一致； 此结论错误率5%，需使用更宽浓度范围的样本(如±30%)进行另外的试验
	－20%	阴性结果≤87.5%(35/40)	
2	＋20%	阳性结果≥90%(36/40)	－20%～＋20%浓度范围包含了C_5～C_{95}区间； 用该方法检测，浓度超过C_{50}±20%的样本检测结果一致
	－20%	阴性结果≥90%(36/40)	

表 3（续）

类型	样本浓度	阴性或阳性结果所占比例	结论
3	+20%	阳性结果≥90%(36/40)	−20%～+20%浓度范围只是部分地在 C_5～C_{95} 区间内(+20%包含了 C_5～C_{95} 区间,但−20%浓度的样本在 C_5～C_{95} 区间内); 用该方法检测,C_{50} +20%的样本检测结果一致,C_{50} −20%的样本不一定能得到一致结果; 需要用低于 C_{50} 更大百分率浓度的样本(如−30%)进行补充试验
	−20%	阴性结果≤87.5%(35/40)	
4	+20%	阳性结果≤87.5%(35/40)	−20%～+20%浓度范围只是部分地在 C_5～C_{95} 区间内(+20%在 C_5～C_{95} 区间内,但−20%包含了 C_5～C_{95} 区间); 用该方法检测,C_{50} −20%的样本检测结果一致,C_{50} +20%的样本不一定能得到一致结果; 需要用高于 C_{50} 更大百分率浓度的样本(如+30%)进行补充试验
	−20%	阴性结果≥90%(36/40)	
注：如 C_{50} 估计不准，−20%～+20%浓度范围也会变化,将导致浓度范围的一侧落在 C_5～C_{95} 区间之外。			

9 方法学比较

9.1 比较方法

9.1.1 为比较不同的检测方法,可用这些检测方法检测同一套样本,比较其检测结果的差异。在新的定性检测试剂或系统的性能验证中,用来比较的方法可以是另一种定性方法,如实验室目前正在使用的方法,也可以是诊断准确度标准,如巴氏试验、阴道镜检查、病史以及随访的合并结果。可据此分为“最高”和“较低”水平两种比较方法。

9.1.2 依据诊断准确度标准对候选方法进行评价为“最高”水平比较。候选方法的检测性能可以用诊断准确度来描述,即待评价方法的检测结果与诊断准确性评判标准的一致性程度,包括敏感性和特异性的评估、阳性和阴性结果的似然比以及接受者工作特征曲线(ROC)。

9.1.3 候选方法与另一种或几种方法之间的一致性进行评价为“较低”水平比较。非诊断准确度标准称为“比较方法”。使用比较方法评价某种候选方法时,不适合用敏感性和特异性来描述比较的结果。不能直接估计候选方法的校正信息,但是可以验证候选方法与比较方法的诊断等效性。

9.2 待测样本

患者样本、参考样本盘和室间质评样本均可用来研究方法的检测性能。由于室间质评样本存在基质效应,可能会带来两种方法比对结果不一致的错误结论。如在室间质评样本中分析物含量接近阳性或阴性阈值,或分析物分子与自然状态下的样本不同,抗体无法识其相应表位。

可从以下几个方面来考虑样本的要求：

a) 患者样本应该来自于与临床实践中预期状态相符的一组代表性群体,此群体应该按照年龄和性别进行合理组合；

b) 每份患者样本的采集量应足够大,以满足检测方法及比较方法的需要；

c) 在对一份样本进行多次采集时可能存在偏差,应将多次采集的同一样本进行混合或采取其他的方法尽量消除偏差；

d) 应尽量保证样本可用于常规检测,并且在检测前保持稳定；

e) 多数样本检测时应为新鲜样本，但某些样本（例如血迹）不是越新鲜越好；

f) 所有检测应尽可能在接近相同的时间完成，从而将检测时由于样本放置时间而造成的偏差控制到最小。

9.3 样本数量

为有效地评价方法的敏感性和特异性，作为最低要求，检测应持续到至少用比较方法获得 50 例阳性样本以确定此种检测方法的敏感性，并且至少用比较方法获得 50 例阴性样本以确定此种检测方法的特异性。当用 50 例阳性样本和 50 例阴性样本进行评价时，若某方法的敏感性和特异性均达到 90%，则其置信区间为 78%～97%（具有 19 个百分点的区间范围，Wilson 计分可信区间）。样本量增加，置信区间随之减小。假如可接受较宽的置信区间，则只需检测较少的样本。

评价者需要向统计学专家咨询被检样本数量，以满足评价者对于得到的统计学变量所提出的要求。此外，检测足够多的样本对于获得生物学变异也十分重要。

9.4 持续时间

比较检测的持续时间可为 10 d～20 d。样本检测分散为数天进行，可使评价者能获得一定数量的有代表性样本，并在常规的实验室使用情况下进行评价。如果可行，评价过程中所有检测样本都要妥善保存并且预留样本以备进一步的检测，必要时可以解决有争议的结果。进一步的检测可以是相同的比较方法、另一种比较方法或者使用临床诊断，它们能为解释检测方法结果的差异提供信息。

9.5 数据核查

所有数据应立即记录并核查，以早期发现分析系统及人为误差的来源。一旦发现一些结果是由可解释的误差引起的，应记录此误差状态，并且这些数据不能用于分析。如不能确定误差产生的原因，则保留原始结果。在比较待评价方法和比较方法检测结果是否存在差异之前，应剔除由技术原因造成的异常结果。如在进行数据比较分析之后发现了某个技术错误，则应对该错误进行校正并重新进行检测。

9.6 差异结果

在方法比较中，待评价方法和比较方法检测相同的样本，可能会产生有差异的结果，产生差异结果的原因可能是待评价方法的误差，也可能是比较方法并非 100%准确。如原因为后者，有差异的样本，可以用“金标准”、“参考方法”来检测确认。对于需将量化数值转化为定性结果的检测，应核查待评价方法和比较方法的检测结果，以确定有差异的结果是否在待评价方法或比较方法的临界点附近。也可分析量值以确定待评价方法和比较方法在检测样本之间的结果差异。此外，应仔细审阅有差异结果的样本相关患者的临床诊断或其他临床信息，找出对产生差异起主导作用的临床状态并加以深入研究。

如比较方法不是 100%准确，重新检测有差异结果的样本，可能不足以得出有统计学意义的灵敏度和特异性估计值，除非采用 100%准确的方法重新检测。这种情况下，要想获得敏感性和特异性的估计值，除不一致的结果的样本外，至少还需再检测一些结果一致的样本。需要重测的样本数量取决于：

——所估计的敏感性和特异性的理想精度；

——已知疾病的流行或状况和待验证方法；

——比较方法及临床诊断之间的关联。

以上均需要详细的统计学设计及缜密的数据分析。

9.7 参考样本盘

参考样本盘是指曾经被检测证实过的、或被成熟参考方法检测验证的，或临床诊断已证明的由多份样本组成的一套临床样本，其对于评价定性测定方法很有价值。参考样本盘其真实值已知，已用成熟的

方法检测验证或已明确其临床诊断。

参考样本盘应包括一些含有不同浓度的有价值的待测物的临床样本，还应尽量包括含一定浓度检测干扰物质的样本。这些可以干扰检测的物质因不同的定性检测有所不同。以下多种疾病状况或因素可能导致假阴性和假阳性结果：

——自身免疫病；

——螺旋体病；

——嗜异性抗体；

——多发性骨髓瘤等。

虽然参考样本盘的使用，对节省评价者的资源和时间以及提高工作效率有很大帮助，但参考样本盘并非总能得到。如评价者没有评价实验室日常检测的临床人群，应用参考样本盘的意义有限，因为实验室日常检测的临床人群代表某疾病或状态典型的流行趋势，没有对这类人群进行评价，就未将阳性和阴性预测值作为检测效果的指标考虑在内。参考样本盘和质评样本在常规临床检测中的应用将大大提高评价过程的可信度。

9.8 临床诊断

临床诊断可以作为“金标准”，主要评价检测结果与患者实际临床症状的符合程度。通常，应考虑以下几个关键点：

——用于方法比较研究的临床标本应包括典型的临床病例。研究对象应包括合理的年龄范围和性别的人群，使得患者标本具有代表性。

——为确立每个患者的临床病情信息，应有正确的临床病情评价标准。

——临床信息可以用来分析待评价检测方法和其比较方法两者检测结果的差异。

10 数据分析

10.1 数据分析分类

数据分析分为以下三类情况：

a) 当待评价方法与诊断准确度标准进行比较时，性能指标包括：敏感性和特异性、阳性结果和阴性结果的似然比、阳性预测值(PPV)和阴性预测值(NPV)。

b) 当待评价方法与用某种比较方法进行比较时，性能指标包括阳性百分比符合度(positive percent agreement，PPA)和阴性百分比符合度(negative percent agreement，NPA)，而非敏感性和特异性。PPA 和 NPA 评价的是待评价方法与比较方法的一致性，并非待评价方法的准确性。此外，由于个体临床诊断(由诊断准确度标准决定)未知，不能计算 PPV、NPV 以及阳性和阴性似然比。

c) 当某定性实验衍生于对定量或顺序尺度二分类时，宜使用 ROC 图表对其检测结果进行描述。

10.2 用诊断准确度标准进行比较评价

10.2.1 敏感性和特异性评价

常用“敏感性”和“特异性”评价临床定性测定试剂或系统的性能，如果样本来自诊断明确的患者时，通过表 4，这两个指标最容易计算。表的每个单元格代表相应样本的数量。敏感性、特异性、由所评价标本估计的疾病流行率、预测值计算见式(1)～式(5)。

表 4　待评价方法与诊断准确度标准相比较的 2×2 表

待评价方法	诊断准确度标准		总数
	阳性	阴性	
阳性	真阳性数(TP)	假阳性数(FP)	TP+FP
阴性	假阴性数(FN)	真阴性数(TN)	FN+TN
总数	TP + FN	FP + TN	N

注：TP、FP、FN 和 TN 可从表 5 中得到，其中 $N=n_{pos}+n_{neg}$。N 为样本总数，n_{pos} 为阳性样本总数，n_{neg} 为阴性样本总数。

$$SE=\frac{TP}{TP+FN}\times 100\% \qquad \cdots\cdots(1)$$

式中：

SE ——敏感性；

TP——真阳性数；

FN——假阴性数。

$$SP=\frac{TN}{FP+TN}\times 100\% \qquad \cdots\cdots(2)$$

式中：

SP ——特异性；

TN——真阴性数；

FP ——假阳性数。

$$prevalence=\frac{TP+FN}{N}\times 100\% \qquad \cdots\cdots(3)$$

式中：

prevalence ——由所评价标本估计的疾病流行率；

TP ——真阳性数；

FN ——假阴性数；

N ——样本总数。

$$PPV=\frac{TP}{TP+FP}\times 100\% \qquad \cdots\cdots(4)$$

式中：

PPV ——阳性预测值；

TP ——真阳性数；

FP ——假阳性数。

$$NPV=\frac{TN}{FN+TN}\times 100\% \qquad \cdots\cdots(5)$$

式中：

NPV ——阴性预测值；

TN ——真阴性数；

FN ——假阴性数。

如果真阳性率(敏感性)与假阳性率(1—特异性)等同，则该方法没有诊断价值。相反，如果敏感性和特异性均接近 100%，则该方法具有较高的诊断价值。

根据式(1)～式(5)计算得到的检测性能指标仅是对真实性能的估计值，因其仅针对研究群体的某一部分个体或样本。如果检测其他的个体或样本，或对同一部分样本在不同时间段进行检测，则检测性能的估计值可能在数值上存在差异。可利用置信区间和显著性水平对样本(个体)选择造成的统计不确定性进行量化，这种不确定性也会随着研究样本数的增加而减小。

检测性能的估计值同样会存在偏差(系统误差)。了解偏差的来源利于在研究过程中尽量避免或减少系统误差。简单地增加样本数量不能减小偏差，只有选择“正确的”研究个体、改变研究模式或数据分析程序才有可能减小或消除偏差。

10.2.2 对流行率和预测值的特殊说明

流行率是指某种疾病或状态在某一特定组群或群体的发生频率，通常用百分率来表示。流行率可评价某方法在评价者所研究群体中的检测性能，评估的流行率和流行率的期望值在该方法的常规应用中应该是近似的。

某种检测方法的预测值是将流行率和敏感性及特异性综合在一起来考虑的。阳性结果的预测值是指患病个体实际检测为阳性的比例，可由真阳性例数除以总阳性例数(真阳性和假阳性例数之和)计算得到。阴性检测结果的预测值是指未患病个体检测为阴性的比例，可由真阴性例数除以总阴性例数(真阴性和假阴性结果之和)计算得到。真阳性、真阴性、假阳性和假阴性结果的数量可用于研究流行率、待评价方法的敏感性和特异性。流行率的估计在判断待评价方法的阴性和阳性预测值方面至关重要。预测值仅适用于受试者的疾病流行率或流行谱与实际检测人群的疾病流行率相差不大的情况。

10.2.3 敏感性和特异性的置信区间

敏感性和特异性的可信限有多种计算方法，数据呈正态分布时，常用的较为简单的方法是正态近似二项式分布。当样本量较少或比率接近“0%”或“100%”的情况下，这种近似正态的假设通常不成立。

敏感性和特异性的精确可信限(clopper-pearson 方法)可通过二项分布、多种统计软件包或者从已经发表的表中获得。有能力计算精确可信限的用户可使用上述方法。除此以外，一种由 Wilson 创建的计分可信区间(score confidence interval)的直接计算方法可以应用于所有例子中。本标准建议使用计分可信区间，介绍如下。

敏感性的 95%计分可信区间表示见式(6)：

$$100 \times \frac{Q_{1,\mathrm{SE}} \mp Q_{2,\mathrm{SE}}}{Q_{3,\mathrm{SE}}} \quad \cdots\cdots(6)$$

其中，$Q_{1,\mathrm{SE}}$、$Q_{2,\mathrm{SE}}$和 $Q_{3,\mathrm{SE}}$的数据可通过式(7)、式(8)、式(9)计算：

$$Q_{1,\mathrm{SE}} = 2 \times \mathrm{TP} + 1.96^2 = 2 \times \mathrm{TP} + 3.84 \quad \cdots\cdots(7)$$

$$Q_{2,\mathrm{SE}} = 1.96\sqrt{1.96^2 + 4 \times \mathrm{TP} \times \mathrm{FN}/(\mathrm{TP} + \mathrm{FN})} \quad \cdots\cdots(8)$$

$$Q_{3,\mathrm{SE}} = 2 \times (\mathrm{TP} + \mathrm{FN} + 1.96^2) = 2 \times (\mathrm{TP} + \mathrm{FN}) + 7.68 \quad \cdots\cdots(9)$$

特异性的 95%计分可信区间表示为式(10)：

$$100 \times \frac{Q_{1,\mathrm{SP}} \mp Q_{2,\mathrm{SP}}}{Q_{3,\mathrm{SP}}} \quad \cdots\cdots(10)$$

其中，$Q_{1,SP}$、$Q_{2,SP}$和$Q_{3,SP}$的数据可从表4得到并通过式(11)、式(12)、式(13)计算：

$$Q_{1,SP}=2\times TN+1.96^2=2\times TN+3.84 \quad \cdots\cdots(11)$$

$$Q_{2,SP}=1.96\sqrt{1.96^2+4\times FP\times TN/(FP+TN)} \quad \cdots\cdots(12)$$

$$Q_{3,SP}=2\times(FP+TN+1.96^2)=2\times(FP+TN)+7.68 \quad \cdots\cdots(13)$$

式(7)～式(9)、式(11)～式(13)中，1.96由标准正态分布对应的95%可信区间得来。若置信水平改变，则1.96也应被相应的百分位点所替代。

10.2.4 两种方法敏感性和特异性的比较

如果定性检测中估计的敏感性和特异性被使用者接受，则不必再做其他的数据分析。然而，使用者可能想明确两种检测方法检测性能是否有统计学差异。待评价方法和比较方法都是相对于诊断准确度标准而独立的，式(1)～式(13)可以预测每种方法的敏感性、特异性等。只有被检标本是实验室检测的代表性标本时，估计的性能指标才能正确评价预期的实验室检测性能。具体如何选择具有代表性的研究标本，操作者可参照NCCLS GP10-A。

当某定性实验衍生于对定量或顺序尺度二分类时，可通过比较各自的ROC曲线进行评价。ROC曲线有助于区别临界值的选择所致的测定敏感性和特异性的差异还是诊断性能上真正的差异，具体如何比较ROC曲线，操作者可参照NCCLS GP10-A。

在方法比较时，如果特异性或敏感性不一致，仅比较敏感性或特异性不合适，因为临界值的改变虽然能提高敏感性或特异性，但常常是以降低特异性或敏感性为代价的。但如果一种方法的敏感性和特异性均优于比较用的方法，则同时比较敏感性和特异性有意义。如果一种方法的敏感性优于另一种方法，但其特异性较差，则难以判断哪种方法较好。

McNemar检验常用来判断两种方法的敏感性/特异性之间是否有统计学上的显著性差异。这种统计学检验没有假定在临床检测性能方面一种方法优于另外一种方法，而认为它们均有可能引起诊断错误。然而，此检验没有提供两种方法可能的差异程度的信息。采用敏感性之间的差异及特异性之间差异的可信区间可能更加有用。

因为待评价方法和比较方法检测的是相同的标本，可以认为数据是“配对的”。为计算正确的可信区间或者纠正McNemar检验统计量，数据需进行待评价方法、比较方法及诊断准确度标准的三方比较。表5列出了三方比较的情况。

表5 待评价方法、比较方法及临床诊断的三方比较

方法结果		总体样本数	真实诊断	
待评价方法	比较方法		阳性	阴性
阳性	阳性	$a_{pos}+a_{neg}$	a_{pos}	a_{neg}
阳性	阴性	$b_{pos}+b_{neg}$	b_{pos}	b_{neg}
阴性	阳性	$c_{pos}+c_{neg}$	c_{pos}	c_{neg}
阴性	阴性	$d_{pos}+d_{neg}$	d_{pos}	d_{neg}
总数		N	n_{pos}	n_{neg}
待评价方法真阳性数		$TP=a_{pos}+b_{pos}$		
待评价方法假阳性数		$FP=a_{neg}+b_{neg}$		
待评价方法假阴性数		$FN=c_{pos}+d_{pos}$		
待评价方法真阴性数		$TN=c_{neg}+d_{neg}$		

注：表中的数据如果按照表4的形式绘制可以得到两个表(待评价方法和比较方法与诊断准确度标准分别比较)，但是两个表4不能合并为一个表5。在绘制表5时，需要得到待评价方法、比较方法和诊断准确性评判标准对每一个个体或样本的三方检测结果。此三方结果的数据不能从表4中获得。

根据表5，待评价方法的估计敏感性见式(14)：

$$SE_{new}=\frac{a_{pos}+b_{pos}}{n_{pos}}\times 100\% \quad \cdots\cdots(14)$$

式中：

SE_{new}——待评价方法的估计敏感性。

比较方法的估计敏感性见式(15)：

$$SE_{old}=\frac{a_{pos}+c_{pos}}{n_{pos}}\times 100\% \quad \cdots\cdots(15)$$

式中：

SE_{old}——比较方法的估计敏感性。

估计的敏感性差异见式(16)：

$$SE_{new}-SE_{old}=\frac{b_{pos}-c_{pos}}{n_{pos}}\times 100\% \quad \cdots\cdots(16)$$

类似的，相应的估计特异性见式(17)：

$$SP_{new}=\frac{c_{neg}+d_{neg}}{n_{neg}}\times 100\% \quad \cdots\cdots(17)$$

式中：

SP_{new}——待评价方法的估计特异性。

比较方法的估计特异性见式(18)：

$$SP_{old}=\frac{b_{neg}+d_{neg}}{n_{neg}}\times 100\% \quad \cdots\cdots(18)$$

式中：

SP_{old}——比较方法的估计特异性。

估计特异性差异见式(19)：

$$SP_{new}-SP_{old}=\frac{c_{neg}-b_{neg}}{n_{neg}}\times 100\% \quad \cdots\cdots(19)$$

注意，如果待评价方法和比较方法分别使用了不同的相对独立的样本时，应采用比较两独立样本的公式计算其敏感性和特异性，而不宜采用三方比较的方法。

敏感性和特异性差异的近似可信区间可以根据配对资料之间差异的可信区间标准统计公式来计算。然而，按这种方法计算得出的差值，只是一个固定的值，可能并不可靠，特别是当两种检测方法的结果不一致，而且样本含量很小时，可靠性更差。因此，本标准推荐在所有情况下直接计算可信限的方法。

配对敏感性间差异的95%可信区间，$D=SE_{new}-SE_{old}$，可计算如下：

$$(D-\sqrt{Q_{5,dse}},D+\sqrt{Q_{6,dse}})$$

用式(6)、式(10)分别计算 SE_{new} 和 SE_{old} 的95%计分可信区间，然后应用式(20)、式(21)计算 $Q_{1,dse}$ 到 $Q_{6,dse}$。

$Q_{1,dse}=(a_{pos}+b_{pos})(c_{pos}+d_{pos})(a_{pos}+c_{pos})(b_{pos}+d_{pos})$（如果 $Q_{1,dse}=0$，那么 $Q_{4,dse}=0$，直接计算 $Q_{5,dse}$）

$Q_{2,dse}=a_{pos}d_{pos}-b_{pos}c_{pos}$（如果 $Q_{2,dse}>n_1/2$，那么 $Q_{3,dse}=Q_{2,dse}-n_1/2$；

如果 $0\leqslant Q_{2,dse}\leqslant n_1/2$，那么 $Q_{3,dse}=0$；

如果 $Q_{2,dse}<0$，那么 $Q_{3,dse}=Q_{2,dse}$）$Q_{4,dse}=Q_{3,dse}/\sqrt{Q_{1,dse}}$（如果 $Q_{1,dse}=0$，那么 $Q_{4,dse}=0$）

$$Q_{5,dse}=(SE_{new}-l_1)^2-2Q_{4,dse}(SE_{new}-l_1)(u_2-SE_{old})+(u_2-SE_{old})^2 \quad \cdots\cdots(20)$$

$$Q_{6,\mathrm{dse}}=(\mathrm{SE_{old}}-l_2)^2-2Q_{4,\mathrm{dse}}(\mathrm{SE_{old}}-l_2)(u_1-\mathrm{SE_{new}})+(u_1-\mathrm{SE_{new}})^2 \quad \cdots\cdots(21)$$

式中：

l_1——$\mathrm{SE_{new}}$ 95%可信区间的下限；

u_1——$\mathrm{SE_{new}}$ 95%可信区间的上限；

l_2——$\mathrm{SE_{old}}$ 95%可信区间的下限；

u_2——$\mathrm{SE_{old}}$ 95%可信区间的上限。

类似地，配对特异性间差异的 95%可信区间，$D=\mathrm{SP_{new}}-\mathrm{SP_{old}}$，可以计算如下：

$$(D-\sqrt{Q_{5,\mathrm{dse}}},D+\sqrt{Q_{6,\mathrm{dse}}})$$

用式(6)、式(10)分别计算 $\mathrm{SP_{new}}$ 和 $\mathrm{SP_{old}}$ 的 95%计分可信区间，然后应用式(22)、式(23)计算 $Q_{1,\mathrm{dsp}}$ 到 $Q_{6,\mathrm{dsp}}$。

$Q_{1,\mathrm{dse}}=(a_{\mathrm{pos}}+b_{\mathrm{pos}})(c_{\mathrm{pos}}+d_{\mathrm{pos}})(a_{\mathrm{pos}}+c_{\mathrm{pos}})(b_{\mathrm{pos}}+d_{\mathrm{pos}})$（如果 $Q_{1,\mathrm{dse}}=0$，那么 $Q_{4,\mathrm{dse}}=0$，直接计算 $Q_{5,\mathrm{dse}}$）

$$Q_{2,\mathrm{dse}}=a_{\mathrm{neg}}\,d_{\mathrm{neg}}-b_{\mathrm{neg}}\,c_{\mathrm{neg}}$$

（如果 $Q_{2,\mathrm{dse}}>n_2/2$，那么 $Q_{3,\mathrm{dse}}=Q_{2,\mathrm{dse}}-n_2/2$；

如果 $0\leqslant Q_{2,\mathrm{dse}}\leqslant n_2/2$，那么 $Q_{3,\mathrm{dse}}=0$；

如果 $Q_{2,\mathrm{dse}}<0$，那么 $Q_{3,\mathrm{dse}}=Q_{2,\mathrm{dse}}$）

$$Q_{4,\mathrm{dse}}=Q_{3,\mathrm{dse}}/\sqrt{Q_{1,\mathrm{dse}}}$$（如果 $Q_{1,\mathrm{dse}}=0$，那么 $Q_{4,\mathrm{dse}}=0$）

$$Q_{5,\mathrm{dse}}=(\mathrm{SP_{new}}-l_1)^2-2Q_{4,\mathrm{dse}}(\mathrm{SP_{new}}-l_1)(u_2-\mathrm{SP_{old}})+(u_2-\mathrm{SP_{old}})^2 \quad \cdots\cdots(22)$$

$$Q_{6,\mathrm{dse}}=(\mathrm{SP_{old}}-l_2)^2-2Q_{4,\mathrm{dse}}(\mathrm{SP_{old}}-l_2)(u_1-\mathrm{SP_{new}})+(u_1-\mathrm{SP_{new}})^2 \quad \cdots\cdots(23)$$

式中：

l_1——$\mathrm{SP_{new}}$ 95%可信区间的下限；

u_1——$\mathrm{SP_{new}}$ 95%可信区间的上限；

l_2——$\mathrm{SP_{old}}$ 95%可信区间的下限；

u_2——$\mathrm{SP_{old}}$ 95%可信区间的上限。

10.3 比较方法非诊断准确性评判标准

10.3.1 概述

当采用的比较方法非诊断准确度标准时不能直接估计敏感性和特异性。虽然也进行了类似的运算，但是估计值被称为 PPA 和 NPA 而非敏感性和特异性，以此来反映估计值并非来自于诊断准确度标准，而是待评价方法与比较方法的符合程度。此外，由于研究对象的诊断未知（由诊断准确度标准获得），不能计算诸如 PPV、NPV 和阳性、阴性似然比之类的量值。

如果假设比较方法的敏感性和特异性为已知，那么待评价方法的敏感性和特异性仍然可以用简单的公式进行估算。然而，这是建立在待评价方法和比较方法“有条件独立”的基础之上的，即如果待评价方法真阳性和真阴性的误差率与比较方法的相同，那么待评价方法和比较方法就是“有条件独立”。实际上，此假设方法不易证实，也不切实际。因此，估计待评价方法和比较方法的符合程度是有意义的。

10.3.2 待评价方法和比较方法的符合性

当诊断未知时，以 2×2 表形式表示待评价方法与比较方法的结果（见表 6）并计算两种方法符合的频率。注意表 5 中最后两栏结果的合计值在表 6 中给出。

表 6　真实诊断未知时的 2×2 表

待评价方法	比较方法		合计
	阳性	阴性	
阳性	a	b	$a+b$
阴性	c	d	$c+d$
合计	$a+c$	$b+d$	n

注 1：表 4 和表 6 的区别在于表 6 并没有显示诊断准确度标准的结果，而是比较方法的结果，这一结果可能不是真实的。

注 2：a、b、c、d 不再代表 TP、FP、FN 和 TN。表 4 的数据反映的是待评价方法与真实诊断的符合性，而表 6 则反映了待评价方法与比较方法的一致性。

总体符合率如式(24)所示：

$$OPA=\frac{a+d}{n}\times 100\% \quad \cdots\cdots(24)$$

式中：

OPA——待评价方法总体符合率。

仅计算总体符合率不足以描述待评价方法与比较方法的符合性。如果两个不同的 2 ×2 表中($b+c$)值相同，那么必然得到相同的总体符合率，然而实际上此时 b 和 c 各不相同。因而，需按式(25)、式(26)计算一对符合性量值，PPA 和 NPA，而总体符合率总是处于 PPA 和 NPA 之间。

$$PPA=\frac{a}{a+c}\times 100\% \quad \cdots\cdots(25)$$

式中：

PPA——阳性百分比符合度。

$$NPA=\frac{d}{b+d}\times 100\% \quad \cdots\cdots(26)$$

式中：

NPA——阴性百分比符合度。

计算 PPA 和 NPA 时应注意，通常待评价方法与比较方法的符合率在数值上同比较方法与待评价方法的符合率有区别[例如：比较方法与待评价方法的 PPA 为 $a/(a+b)$，而不是 $a/(a+c)$]。因而，评价者应清楚地说明所使用的计算。

符合率的一个主要缺陷是其并不是一个“正确性”的量度，事实上，两种检测方法可能高度符合，但是敏感性和特异性都很低。相反，两种检测方法不符合也并不意味着待评价方法是错误的而比较方法是正确的。

与待评价项目相关的疾病流行率(此情况下通常未知)严重影响符合率，因此对任何其他人群普遍应用符合率量值时必须注意。例如，待评价方法和比较方法在真实诊断为阴性时符合良好，而真实诊断为阳性时二者符合较差，当用于评价的标本的疾病流行率低时，两种方法的总体符合率将会偏高，当疾病流行率高时则会偏低。如果疾病流行率未知，那么无法对疾病流行率不同的人群使用符合率。

10.3.3　符合率的置信区间

符合率的 95%计分可信区间表示为：

$$100\times\frac{Q_1\mp Q_2}{Q_3} \quad \cdots\cdots(27)$$

其中，Q_1、Q_2 和 Q_3 的值可用式(28)、式(29)、式(30)得到：

$$Q_1 = 2 \times (a + d) + 1.96^2 = 2 \times (a + d) + 3.84 \quad \cdots\cdots (28)$$

$$Q_2 = 1.96\sqrt{1.96^2 + 4(a + d)(b + c)/n} \quad \cdots\cdots (29)$$

$$Q_3 = 2 \times (n + 1.96^2) = 2 \times n + 7.68 \quad \cdots\cdots (30)$$

式(28)～式(30)中，1.96 是从标准正态分布对应的 95%置信区间得来的。

PPA 的 95%计分可信区间计算方法与式(6)和式(10)类似，将“a”“b”“c”和“d”分别替换 TP、FP、FN 和 TN，得到式(31)和式(34)。

PPA 95%计分可信区间表示为：

$$100 \times \frac{Q_{1,\mathrm{ppa}} \mp Q_{2,\mathrm{ppa}}}{Q_{3,\mathrm{ppa}}} \quad \cdots\cdots (31)$$

式中：

$Q_{1,\mathrm{ppa}}$、$Q_{2,\mathrm{ppa}}$ 和 $Q_{3,\mathrm{ppa}}$ 的值可运用式(32)、式(33)、式(34)得到：

$$Q_{1,\mathrm{ppa}} = 2a + 1.96^2 = 2a + 3.84 \quad \cdots\cdots (32)$$

$$Q_{2,\mathrm{ppa}} = 1.96\sqrt{1.96^2 + 4ac/(a + c)} \quad \cdots\cdots (33)$$

$$Q_{3,\mathrm{ppa}} = 2(a + c + 1.96^2) = 2(a + c) + 7.68 \quad \cdots\cdots (34)$$

NPA 95%计分可信区间的计算方法为：

$$100 \times \frac{Q_{1,\mathrm{npa}} \mp Q_{2,\mathrm{npa}}}{Q_{3,\mathrm{npa}}} \quad \cdots\cdots (35)$$

$Q_{1,\mathrm{npa}}$、$Q_{2,\mathrm{npa}}$ 和 $Q_{3,\mathrm{npa}}$ 的值可运用式(36)、式(37)、式(38)得到：

$$Q_{1,\mathrm{npa}} = 2 \times d + 1.96^2 = 2 \times d + 3.84 \quad \cdots\cdots (36)$$

$$Q_{2,\mathrm{npa}} = 1.96\sqrt{1.96^2 + 4bd/(b + d)} \quad \cdots\cdots (37)$$

$$Q_{3,\mathrm{npa}} = 2 \times (b + d + 1.96^2) = 2 \times (b + d) + 7.68 \quad \cdots\cdots (38)$$

11 实例

11.1 用诊断准确度标准进行比较评价

真实诊断已知时，待评价方法和比较方法分别与诊断准确度标准进行比较。附录 B 列举了两种商品化的酶联免疫吸附试验(ELISA)试剂盒用于检测 154 例梅毒螺旋体感染状况已确定的患者血清。

待评价方法和比较方法的结果 2×2 表在附录 B 示例 1 和示例 2 中列出，并估计每种方法的敏感性和特异性。可使用 ROC 分析描述每种方法的性能指标。另外，当待评价方法的敏感性和特异性均大于比较方法的敏感性和特异性时，敏感性/特异性配对的联合统计比较也很有意义。为了比较两种方法的敏感性和特异性，需要在待评价方法、比较方法和诊断准确度标准之间进行三方比较，具体实例见附录 B 示例 3。

11.2 比较方法并非参考方法

如真实诊断未知，但仍需要进行待评价方法与另一方法的比较。附录 C 关于两种丙型肝炎病毒抗体(抗-HCV)检测方法结果评价就属于此种情况。例中，待评价方法为 ELISA 法，而比较方法是免疫印迹方法。

附 录 A
（资料性附录）
−20%～+20%浓度范围是否包含 C_5～C_{95}区间示例

如果某种试剂或系统检测血清中乙型肝炎表面抗原(HBsAg)的 C_{50}为 0.05 IU/mL，则用其对一份浓度为 0.05 IU/mL(通过定量检测确定)的样本多次检测，将获得 50%的阳性和 50%的阴性结果，对 C_{50}+20%浓度的样本(0.06 IU/mL)进行 40 次重复检测将产生 40 次阳性结果，对 C_{50}−20%浓度的样本(0.04 IU/mL)进行 40 次重复检测将产生 40 次阴性结果。如果+20%浓度样本 40 次检测得到36 次阳性结果 (90%)且−20%浓度样本 40 次检测得到 36 次阴性结果 (90%)，则可推断−20%～+20%浓度范围包含了 C_5～C_{95}区间。浓度范围 0.04 IU/mL～0.06 IU/mL 有 86%的可能包含了该方法 C_5～C_{95}阳性范围，换言之，用该方法检测，浓度≤0.04 IU/mL 或≥0.06 IU/mL(与 C_{50}相差 20%或更多)的样本，有 86%的可能将得到一致的检测结果。如果进行 60 次重复检测，则有 92.6%的可能，浓度范围 0.04 IU/mL～0.06 IU/mL 在或超出了该方法的 C_5～C_{95}阳性范围，换言之，用该方法检测，浓度≤0.04 IU/mL或≥0.06 IU/mL(与 C_{50}相差 20%或更多)的样本，有 92.6%的可能得到一致结果。

不同结果的推理过程相同。假设 C_{50}+20%(0.06 IU/mL)的样本 40 次检测得到 34 次 (85%)阳性结果，而 C_{50}−20%(0.04 IU/mL)的样本得到 40 次(100%)阴性结果。由于+20%样本浓度的阳性结果低于 90%，−20%样本浓度的阴性结果高于 90%，可认为浓度范围 0.04 IU/mL～0.06 IU/mL 只是部分在 C_5～C_{95}区间。浓度低于 0.04 IU/mL 的样本用该方法可得到一致结果，而浓度高于 0.06 IU/mL的样本用该方法不一定能得到一致结果。需要用另外一个样本浓度如 C_{50}+30%(0.065 IU/mL)进行补充试验。

附 录 B
（资料性附录）
用诊断准确度标准进行比较评价示例

B.1 示例 1

待评价方法与诊断准确度标准的 2×2 表见表 B.1。

表 B.1 待评价方法与诊断准确度标准的 2×2 表（诊断准确度标准：梅毒螺旋体感染）

待评价方法	阳性	阴性	合 计
阳性	86	3	89
阴性	6	59	65
合计	92	62	154

估计敏感性＝[TP/(TP＋FN]×100％＝[86/92]×100％＝ 93.5％
估计特异性＝[TN/(FP＋TN]×100％＝[59/62]×100％＝ 95.2％
评价样本的疾病流行率＝[(TP＋FN)/N]×100％＝(92/154)×100％＝ 59.7％
阳性预测值(PPV)＝[TP/(TP＋FP)]×100％＝[86/89]×100％＝ 96.6％
阴性预测值(NPV)＝[TN/(FN＋TN)]×100％＝[59/65]×100％＝ 90.8％
待评价方法的敏感性：

$$SE = (86/92) \times 100\% = 93.5\%$$

确切的 95％置信区间为[84.1％,98.2％]。
95％计分可信区间为[86.5％,97.0％],利用式(6)计算 95％计分可信区间：

$$Q_{1,SE} = 2 \times 86 + 3.84 = 175.84$$

$$Q_{2,SE} = 1.96\sqrt{3.84 + 4 \times 86 \times 6/92} = 10.047$$

$$Q_{3,SE} = 2 \times 92 + 7.68 = 191.68$$

下限：$100\%(Q_{1,SE} - Q_{2,SE})/Q_{3,SE} = 100\%(175.84 - 10.047)/191.68 = 86.5\%$
上限：$100\%(Q_{1,SE} + Q_{2,SE})/Q_{3,SE} = 100\%(175.84 + 10.047)/191.68 = 97.0\%$
待评价方法的特异性：

$$SP = (59/62) \times 100\% = 95.2\%$$

确切的 95％置信区间为[83.5％, 99.4％]。
95％计分可信区间为[85.7％, 99.3％],利用式(10)计算如下：

$$Q_{1,SP} = 2 \times 59 + 3.84 = 121.84$$

$$Q_{2,SP} = 1.96\sqrt{3.84 + 6 \times 59 \times 3/62} = 8.975$$

$$Q_{3,SP} = 2 \times 62 + 7.68 = 131.68$$

下限：$100\%(Q_{1,SP} - Q_{2,SP})/Q_{3,SP} = 100\%(121.84 - 8.975)/131.68 = 85.7\%$
上限：$100\%(Q_{1,SP} + Q_{2,SP})/Q_{3,SP} = 100\%(121.84 + 8.975)/131.68 = 99.3\%$

B.2 示例 2

比较方法与诊断准确度标准的 2×2 表见表 B.2。

表 B.2　比较方法与诊断准确度标准的 2×2 表(诊断准确度标准:梅毒螺旋体感染)

比较方法	阳性	阴性	合计
阳性	81	11	92
阴性	11	51	62
合计	92	62	154

估计敏感性=[TP/(TP+FN)]×100%=[81/92]×100%= 88.0%

估计特异性=[TN/(FP+TN)]×100%=[51/62]×100%= 82.3%

评价样本的疾病流行率=[(TP+FN)/N]×100%= (92/154)×100%= 59.7%

阳性预测值(PPV)=[TP/(TP+FP)]×100%=[81/92]×100%= 88.0%

阴性预测值(NPV)=[TN/(FN+TN)]×100%= [51/62]×100%= 82.3%

比较方法的敏感性:

$$SE = [81/92] \times 100\% = 88.0\%$$

95%计分可信区间为[80.7%,94.2%],利用式(6)计算如下:

$$Q_{1,SE} = 2 \times 81 + 3.84 = 165.84$$

$$Q_{2,SE} = 1.96\sqrt{3.84 + 4 \times 81 \times 11/92} = 12.790$$

$$Q_{3,SE} = 2 \times 92 + 7.68 = 191.68$$

下限:$100\%(Q_{1,SE} - Q_{2,SE})/Q_{3,SE} = 100\%(165.84 - 12.790)/191.68 = 79.8\%$

上限:$100\%(Q_{1,SE} + Q_{,SE2})/Q_{3,SE} = 100\%(165.84 + 12.790)/191.68 = 93.1\%$

比较方法的特异性:

$$SP = [51/62] \times 100\% = 82.3\%$$

95%计分可信区间为[71.0%,89.8%],利用式(10)计算如下:

$$Q_{1,sp} = 2 \times 51 + 3.84 = 105.84$$

$$Q_{2,sp} = 1.96\sqrt{3.84 + 4 \times 51 \times 11/62} = 12.401$$

$$Q_{3,sp} = 2 \times 62 + 7.68 = 131.68$$

$$100\%(Q_{1,sp} - Q_{2,sp})/Q_{3,sp} = 100\%(105.84 - 12.401)/131.68 = 71.0\%$$

$$100\%(Q_{1,sp} + Q_{2,sp})/Q_{3,sp} = 100\%(105.84 + 12.401)/131.68 = 89.8\%$$

B.3　示例 3

待评价的方法、比较方法和诊断准确度标准之间三方比较示例见表 B.3。

表 B.3　待评价的方法、比较方法和诊断准确度标准之间三方比较

方法结果		总标本数	诊断准确度标准	
待评价方法	比较方法		阳性	阴性
阳性	阳性	83	80	3
阳性	阴性	6	6	0
阴性	阳性	9	2	7
阴性	阴性	56	5	51
合计		154	93	61

敏感性比较：

$$D=\mathrm{SE}_{\mathrm{new}}-\mathrm{SE}_{\mathrm{old}}=92.5-88.2 \text{ 或者 } 100\%\times(6-2)/73=4.3\%$$

$$l_1=86.2\%\text{（根据 }\mathrm{SE}_{\mathrm{new}}\text{ 的置信区间得到）}$$

$$u_1=95.3\%$$

$$l_2=80.1\%\text{（根据 }\mathrm{SE}_{\mathrm{old}}\text{ 的置信区间得到）}$$

$$u_2=93.2\%$$

$$Q_{1,\mathrm{dse}}=(80+6)(2+5)(80+2)(6+5)=86\times7\times82\times11=543\ 004$$

$$Q_{2,\mathrm{dse}}=(80\times5)-(6\times2)=388$$

$$n_1/2=93/2=46.5<388=Q_{2,\mathrm{dse}}$$

$$Q_{3,\mathrm{dse}}=Q_{2,\mathrm{dse}}-n_1/2=388-46.5=341.5$$

$$Q_{4,\mathrm{dse}}=Q_3/\sqrt{Q_{1,\mathrm{dse}}}=341.5/\sqrt{543\ 004}=341.5/736.89=0.463\ 4$$

$$Q_{5,\mathrm{dse}}=(92.5-86.2)^2-2(0.463\ 4)(92.5-86.2)(93.2-88.2)+(93.2-88.2)^2=44.76$$

$$Q_{6,\mathrm{dse}}=(88.20-80.1)^2-2(0.463\ 4)(88.0-80.1)(95.3-92.5)+(95.3-92.5)^2=52.43$$

$$D-\sqrt{Q_{5,\mathrm{dse}}}=4.9-\sqrt{44.76}=-1.79$$

$$D+\sqrt{Q_{6,\mathrm{dse}}}=4.9+\sqrt{52.43}=11.54$$

$D=\mathrm{SE}_{\mathrm{new}}-\mathrm{SE}_{\mathrm{old}}$，其 95%置信区间为(−1.79%，11.54%)。

特异性比较：

$$D=\mathrm{SP}_{\mathrm{new}}-\mathrm{spec}_{\mathrm{old}}=95.1-83.6 \text{ 或 } 100\%\times(7-0)/61=11.5\%$$

$$l_1=86.5\%\text{（根据 }\mathrm{SP}_{\mathrm{new}}\text{ 的置信区间得到）}$$

$$u_1=98.3\%$$

$$l_2=77.6\%\text{（根据 }\mathrm{SP}_{\mathrm{old}}\text{ 的置信区间得到）}$$

$$u_2=85.6\%$$

$$Q_{1,\mathrm{dse}}=(3+0)(7+51)(3+7)(0+51)=(3)(58)(10)(51)=88\ 740$$

$$Q_{2,\mathrm{dse}}=(3\times51)-(0\times7)=155$$

$$n_1/2=61/2=31.5<155=Q_{2,\mathrm{dse}}$$

$$Q_{3,\mathrm{dse}}=Q_{2,\mathrm{dse}}-n_1/2=155-31.5=121.5$$

$$Q_{4,\mathrm{dse}}=Q_{3,\mathrm{dse}}/\sqrt{Q_{1,\mathrm{dse}}}=122/\sqrt{88\ 740}=121.5/297.89=0.409\ 5$$

$$Q_{5,\mathrm{dse}}=(95.1-86.5)^2-2(0.409\ 5)(95.1-86.5)(85.6-83.6)+(85.6-83.6)^2=61.83$$

$$Q_{6,\mathrm{dse}}=(83.6-77.6)^2-2(0.409\ 5)(83.6-77.6)(98.3-95.1)+(98.3-95.1)^2=30.51$$

$$D-\sqrt{Q_{5,\mathrm{dse}}}=12.2-\sqrt{61.83}=4.34$$

$$D+\sqrt{Q_{6,\mathrm{dse}}}=12.2+\sqrt{30.51}=17.72$$

$D=\mathrm{SP}_{\mathrm{new}}-\mathrm{SP}_{\mathrm{old}}$，其 95%置信区间为(4.34%，17.72%)。

由于敏感性差异的置信区间包括零，不能得出敏感性有统计学差异的结论。但是特异性的置信区间不包括零，因此表明特异性有统计学差异。

附 录 C
（资料性附录）
真实诊断未知比较评价示例

真实诊断未知比较评价示例 2×2 表见表 C.1

表 C.1 真实诊断未知比较评价示例 2×2 表（比较方法：免疫印迹）

待评价方法（ELISA 法）	阳性	阴性	合计
阳性	570	30	600
阴性	28	444	472
合计	598	474	1 072

待评价方法与比较方法的阳性符合率 $=[570/598]\times 100\%=95.3\%$

待评价方法与比较方法的阴性符合率 $=[444/474]\times 100\%=93.7\%$

总体符合率$=[(a+d)/n]\times 100\%=[1\ 014/1\ 072]\times 100\%=94.6\%$

用 10.3.3 的方法计算符合率的置信区间。

阳性符合率（ELISA 法/免疫印迹法）：

$$\mathrm{PPA}=[570/598]\times 100\%=95.3\%$$

95%计分可信区间（与免疫印迹法结果比较）为[93.3%，96.7%]，利用式（27）计算如下：

$$Q_{1,\mathrm{ppa}}=2\times 570+3.84=1\ 143.84$$

$$Q_{2,\mathrm{ppa}}=1.96\sqrt{3.84+4\times 570\times 28/598}=20.612$$

$$Q_{3,\mathrm{ppa}}=2\times 598+7.68=1\ 203.68$$

下限：$100\times(Q_{1,\mathrm{ppa}}-Q_{2,\mathrm{ppa}})/\ Q_{3,\mathrm{ppa}}=100\%\times(1\ 143.84-20.612)/1\ 203.68=93.3\%$

上限：$100\times(Q_{1,\mathrm{ppa}}+Q_{2,\mathrm{ppa}})/\ Q_{3,\mathrm{ppa}}=100\%\times(1\ 143.84+20.612)/1\ 203.68=96.7\%$

阴性符合率（ELISA 法/免疫印迹法）：

$$\mathrm{NPA}=[222/237]\times 100\%=93.7\%$$

95%计分可信区间（与免疫印迹法结果比较）为[91.1%，95.5%]，利用式（31）计算如下：

计算 95%的置信分数：

$$Q_{1,\mathrm{ppa}}=2\times 444+3.84=891.84$$

$$Q_{2,\mathrm{ppa}}=1.96\sqrt{3.84+4\times 30\times 444/474}=21.132$$

$$Q_{3,\mathrm{ppa}}=2\times 474+7.68=955.68$$

下限：$100\times(Q_{1,\mathrm{ppa}}-Q_{2,\mathrm{ppa}})/\ Q_{3,\mathrm{ppa}}=100\times(891.84-21.132)/955.68=91.1\%$

上限：$100\times(Q_{1,\mathrm{ppa}}+Q_{2,\mathrm{ppa}})/\ Q_{3,\mathrm{ppa}}=100\times(891.84+21.132)/955.68=95.5\%$

总体符合率$=[1\ 014/1\ 072]\times 100\%=94.6\%$

95%计分可信区间为[93.5%，95.4%]，利用式（35）计算如下：

$$Q_1=2\times 1\ 014+3.84=2\ 031.84$$

$$Q_2=1.96\sqrt{3.84+4\times 1\ 014\times 29/1\ 072}=20.847$$

$$Q_3=2\times 1\ 072+7.68=2\ 151.68$$

下限：$100\%\times(Q_1-Q_2)/Q_3=100\%\times(2\ 031.84-20.847)/2\ 151.68=93.5\%$

上限：$100\%\times(Q_1+Q_2)/Q_3=100\%\times(2\ 031.84+20.847)/2\ 151.68=95.4\%$

参 考 文 献

[1] GB/T 20468—2006 临床实验室定量测定室内质量控制指南

[2] EP12-A Vol.22 No.14 User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline.2002

[3] EP12-A2 Vol.28 No.3 User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline-Second Edition.2008

[4] Ross RS, Viazov S et al. Analytical performance characteristics and clinical utility of a novel assay for total hepatitis C virus core antigen quantification. J Clin Microbiol, 2010,48,1161-1168.

[5] Daniel HD, Abraham P et al. Evaluation of a rapid assay as an alternative to conventional enzyme immunoassays for detection of hepatitis C virus-specific antibodies. J Clin Microbiol,2005,43:1977-1978.

[6] Melo J, Nilsson C et al.Comparison of the performance of rapid HIV tests using samples collected for surveillance in Mozambique. J Med Virol. 2009 ,81:1991-1998.