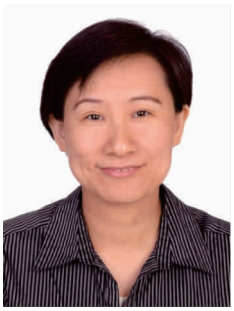
文章编号:1001-764X(2012)10-821-05

“方法学比较”法评价临床检验定量测量方法正确度时应注意的问题与对策*

陈宝荣(北京航天总医院检验科,北京 100076)



作者简介:陈宝荣,1967 年生,主任技师,硕士研究生导师。北京航天总医院检验科、医学参考实验室主任。主要从事临床化学检验与医学实验室管理。2005 年起致力于临床酶学参考体系建设及临床化学测量标准化工作。复现 JCTLM 列表参考方法 11 项,研制国家标准物质 10 种。建成中国第一个通过 ISO 17025 和 ISO 15195 认可的医学参考实验室。申报国家科技发明专利 6 项并成功转让 2 项。制定、参与制定行业标准 10 余项。参加或主持国际合作研究、科技部和北京市科研项目 10 余项。

摘要:探讨“方法学比较”法评价临床检验定量测量方法正确度时存在的问题,以帮助临床实验室正确评价这类方法的正确度。本文从实验方案的设计、实验准备、实验、数据处理、实验结果表述等五方面分析目前国内外临床实验室常用的“方法学比较“法在评价定量测量方法正确度时面临的问题,提出目前我国临床实验室采用该方法评价正确度时应注意的问题及解决的方法。严密的实验设计、恰当的实验样本、运行稳定的测量程序、训练有素的技术人员、适当的数据处理方法和评价标准是获得定量测量方法正确度评价正确结论的基础。行业标准的制定有助于临床实验室建立规范的评价定量测量方法正确度的“方法学比较”方法并获得可靠的评价结论。

关键词:参考方法;量值溯源;正确度;测量;方法学比较

中图分类号:R446　　**文献标志码**:A

“方法学比较”法是临床实验室进行定量测量方法性能评价的可靠方法。目前,国内一些实验室采用 CLSI EP9-A2[1] 文件描述的“方法学比较”法评价测量结果的正确度[2],也有部分实验室采用 CLSI EP15-A2[3] 文件提出的“方法学比较”法,即采用新方法和被替代的方法同时检测患者样本确认新方法的性能,以验证厂家声明的性能。笔者研究发现,不同的实验方案、不同的数据处理方法在处理相同的测量数据时可得出不一致的结论[4-5],应引起高度重视。现将“方法学比较“法评价定量检验结果正确度时应注意的问题及对策总结如下。

1　实验方案的设计

就像“好设计可以得到好作品”一样,一个经严密设计的“方法学比较”实验,可以使实验室获得可靠的定量测量方法正确度评价结果。临床实验室在设计“实验方案”时,应注意以下三个关键问题。

1.1　参考方法测量程序的性能验证　采用“方法学比较”法评价定量测量方法正确度性能时,比较方法应采用 JCTLM 列表的参考方法。这在 EP9-A2 和 HOKLAS Supplementary Criteria[6] 中有明确规定。参考方法[2]是一类经充分研究、用于获得与同类量测量标准没有关系的测量结果所用的方法,适用于评定由同类量的其他测量程序获得的被测量量值的测量正确度。一般用于校准或表征标准物质测量结果时采用。从理论上讲,参考方法的测量结果具有溯源性,但是不是只要运行的是参考方法,获得的测量结果就一定具有溯源性呢? 关于这一点,从事参考测量的技术人员都非常清楚,只有建立的参考方法测量程序的性能满足参考方法测量性能要求时,通过其获得的测量结果才是可溯源的。换句话讲,当采用“方法学比较”法进行正确度评价时,所用的参考方法测量程序应首先进行方法学性能验证,即溯源性确认。只有参考方法测量程序的测量性能达到参考方法性能要求时,才能用作“比较”方法。因此,在“实验方案”中应包含参考方法性能验证的内

* **基金项目**:国家质检公益性行业科研项目(20120066);北京市首都医学发展科研基金资助项目(2009-2100,2009-3182,2007-1045,2011-6032-02)。

容、要求及通过标准。

1.2 常规方法的校准 用“方法学比较”法评价正确度时另一实验主体是被评价的定量测量常规方法。众所周知,常规方法的正确度与校准密不可分,也就是说正确的校准是常规方法实现量值溯源的第一步。校准品的浓度、性质(溯源性、互通性、稳定性、均匀性)、校准方法(周期、模式)、校准品的正确使用是获得良好校准结果的重要因素。这些内容在“实验方案”中均应进行详细说明。应选择有溯源性定值、被测量浓度恰当、与测量样本有良好互通性、均匀且稳定的校准品,在确认满足测量方法性能要求的校准模式下校准或在有效的校准周期内进行“方法学比较”实验,以获得客观的评价结果。

1.3 评价样本及测量 采用“方法学比较”法评价正确度时另一重要因素是评价样本及测量。评价样本的类型、来源、浓度范围与其数量、测量条件包括实验人员选择、实验仪器、实验试剂及相关产品、测量程序及测量环境等是保证“方法学比较”法评价结果可靠的重要因素。在“实验设计”中应进行详细规定。

2 实验准备

采用“方法学比较”法评价定量测量方法正确度是方法学性能验证的重要组成部分,也是对待评价的定量测量方法是否可用于临床实验室进行患者样本测量的重要实验。ISO 15189[7]、欧盟器具令[8]、GB/T 21415-2008[9]等文件均规定实验室应使用正确可溯源的方法测量患者样本,表明用于患者样本测量的常规方法的正确度是临床实验室要重点关注的问题。为获得准确可靠的正确度评价结果,“实验准备”是常规方法正确度评价的重要步骤。在严密设计“方法学比较”实验方案的基础上,进行细致、适当的实验前准备是必不可少的环节。应重点关注以下五个方面。

2.1 人员准备 “正确度评价”是一项技术要求高、测量过程复杂的实验,对实验人员技术能力、责任心、统筹能力要求高。应选择经验丰富、具备良好技能与责任心的技术人员进行实验。实验前这些人员应熟知“实验方案”,并能按其要求正确执行。

2.2 仪器准备 应使用检定合格、性能稳定的仪器用于定量方法测量结果的正确度评价。实验前,应由有经验的工程师、实验室技术人员共同确认仪器的状态,对影响被测量分析性能的重要参数应按“实验方案”要求逐项验证、确认满足“实验方案”规定的性能指标后,方可进行“方法学比较”实验。

2.3 样本准备 样本准备也是临床实验室进行“方法学比较“的关键步骤之一。恰当数量的高质量的样本不仅可使临床实验室获得客观评价结果,还可明显降低实验成本。EP9-A2 文件规定应使用 40 份覆盖方法测量范围的不同浓度的患者样本评价测量结果的正确度,并规定了各浓度水平样本的例数。EP15-A2 文件推荐使用浓度水平覆盖方法可报告范围的 20 份新鲜患者样本评价测量结果的正确度。王治国等在《临床检验方法确认与性能验证》[10]一书中提出应使用至少 10 个常规分析样本进行新检验方法的正确度估计。Westgard[11]提出应使用 40 份患者样本评价常规方法的正确度,样本应覆盖整个方法的可报告范围及代表方法在常规应用预期疾病的浓度范围。到底用多少份样本评价方法的正确度性能合适?各“标准”没有给出一致的说法,这难免让实验室的技术人员有无所适从的感觉。样本数量过少可能导致评价结果可信度降低,样本数量大会增加评价结果的可信度,但成本将明显增加,究竟选择多少份样本进行定量测量方法的正确度评价恰当呢?笔者研究发现,使用 20 份以上均匀覆盖定量测量方法测量范围的新鲜/冰冻(-80 ℃)患者单人份样本评价测量结果正确度与使用 40 份甚至更多份数样本采用相同的统计学方法评价时得到的结果基本一致。如果评价样本浓度范围及分布恰当,20 份样本已能满足方法正确度评价要求,没有必要再增加评价样本数量。

样本质量对方法的正确度评价也有一定影响。实验室应依据实验方案要求在正式实验前准备符合要求的样本,即应尽量选择外观正常、无明显溶血、乳糜、黄疸等有可能影响被测量准确测量的样本。如有可能,还应确认这些样本中不含有可能干扰被测量准确测量的药物或其他成分。评价样本的有效样本量应同时满足参考方法和常规方法正确度评价实验需要。值得注意的是实际工作中参考方法和常规方法的线性范围有时相差悬殊,此时,应依据线性范围窄的方法选择评价样本,浓度应均匀覆盖方法的线性范围。还应注意不宜使用经各种预处理的样本如反复冻融、稀释、浓缩、有效成分添加等。

2.4 试剂、耗材的准备 试剂与耗材涉及参考方法与常规方法两部分。实验前实验室应以列表方式准备实验所需试剂及相关耗材。关键试剂特别是一些不稳定试剂的性能应按“实验方案”要求在实验前进行验证并确认满足方法性能要求。实验所需的全

部试剂、耗材在“实验方案”中均应根据被测量常规实验情况设计一定备用量。

3　实验中应注意的问题

样本稳定性、测量周期与批次是目前“方法学比较“法评价正确度时应特别关注的问题。关于样本稳定性,通常情况下,被测量不同,样本的稳定性不同。有时,同一被测量在不同样本类型、不同保存状态中稳定性可能也不同,实验室应了解待评价被测量的这些特性,并应根据这些特性确定样本有效的测量时间,如对于大多数临床化学检验项目,单一样本测量时间室温状态下应不超过 2 h。Westgard 等建议在不同天、不同批测量评价样本,实验周期至少 5 d,每天测量 2 ~ 5 个患者样本。EP15-A2 文件规定应在 3 ~ 4 d 内,每天在 4 h 内完成 5 ~ 7 份样本的测量。EP9-A2 文件没有规定测量时间要求。笔者研究发现,不同的测量项目实验室应根据其测量方法的运行稳定性和实验室的测量能力决定单一样本允许测量时间。上述不同学者给出了不同的实验周期,有专家认为这可以增加实验结果的可信度。笔者以为,正确度性能评价应基于运行稳定的测量程序进行,如果实验室运行的测量程序本身还存在这样或那样导致测量结果变异大的情况,实验室应优先解决测量程序精密度问题,没有必要通过延长测量周期、增加测量批次来增加评价结果的可信度。多个实验已证实对于一个运行稳定的测量程序,多日、多批次测量与单日单/多批次测量可获得非常一致的结果,是否可以据此理解为只要测量程序稳定,单日单/多批次测量结果也可获得满意的方法正确度评价结果?这有待于更多实验研究验证。

4　实验数据处理

“方法学比较”实验不同于一般的定量测量实验,通常由具有良好技能、熟知实验程序的技术人员在适宜的实验环境中使用性能可靠的仪器、试剂及相关产品、运行稳定的操作程序进行实验。从理论上讲,这类实验获得的数据应足以代表实验室运行程序的测量水平,不应再剔除实验数据。但由于“方法学比较”实验也是一种科学试验,在统计学上不可能是完全的理想状态,实验室应根据实验性质在“实验方案”中约定数据剔除标准,应尽量与其他文件/标准等一致。EP9-A2 文件给出了离群值的判断标准,不少学者使用这一标准评价实验数据有效性,基本满足“方法学比较”法实验数据有效性判断要求。

在采用“方法学比较”法评价定量测量方法正确度时,通过实验获得的有效数据的统计学处理方法是实验室获得“正确度评价”正确结论的关键环节之一。在医学实验中,大多数方法均基于配对 t 检验进行分析,有些方法使用原始数据作统计分析;有些方法则使用数据的相对或其他形式作统计分析如 Bland-Altman 图形法[12];一些标准如 EP9-A2 等,除对原始数据进行相应的统计分析外,还增加医学决定水平处允许偏移的评价及限定,形成新的评价标准。总体上看各方法实验数据的统计学处理原则似乎并没有改变。

虽然各方法实验数据总的统计学处理原则一致,但既往研究[13-17]发现不同统计学方法处理相同实验数据群常导致不同的结论。在临床实验室评价二种定量方法一致性时,由于受统计软件限制常会选用配对 t 检验、线性回归、EP9-A2 文件方法、Bland-Altman 图形法等方法进行分析。对于配对 t 检验:结论由 t 检验假设决定(总均数与 0 比较)。根据公式:$t=\frac{\bar{d}-0}{s/\sqrt{n}}$,若 $t>1.96$,则 $P<0.05$,示两法均数不等;若 $t<1.96$,则 $P>0.05$,示两法均数相等。1987 年 Howell[18]曾使用配对 t 检验作为衡量一致性的方法。根据文献[19]配对 t 检验对系统误差敏感,对随机误差不敏感。BeAdard 等认为[20]配对 t 检验的本质是对“差异”的检验,而非对“一致”的检验。当采用此法评价时常存在二个误区:一是两法均数相等是否等价于两法测量结果一致?二是 t 或 P 值受那些因素影响?这常导致分析结果与实际情况不符。如文献[4]中 B 法与 IFCC 法的比较结果,用配对 t 检验分析时与其他统计方法得出的结论不一致。关于 EP9-A2 文件方法:要求评价常规方法正确度时应与参考方法比较,包括直线回归方程和回归方程的 95% 偏移预测区间与设定偏移值比较二部分。对直线回归方程评价斜率和截距与“0”的差别是否有显著性,在配对 t 检验中,若 $P<0.05$,直线方程成立。反之亦然。此时会遇到与 t 检验类似的问题:数据越分散,越不可能拒绝斜率为 1、截距为 0 的假设。栾荣生等将相关系数用于评价定量数据两种测量结果的一致性[21]。通常相关系数仅能反映两变量线性关系的密切程度,不能检出恒定或成比例的偏移,如文献[4]中所述的 4 种常规方法与 IFCC 法相关系数均 >0.99,各常规方法结果与 IFCC 法结果高度相关,但 4 种常规方法中的

D 法测量结果明显低于 IFCC 法,已远超出临床和统计学可接受范围。因此线性回归方法对系统偏差不敏感,不能正确反映两法一致性。同样对相关系数 r 的检验即使 $P<0.05$,也无法得出“两法一致性“的正确结论。基于此 CLSI 建立了 EP9-A2 文件的方法,即在直线回归方程比较基础上增加回归方程的 95% 偏移预测区间与设定偏移值比较内容,也即各常规方法与 IFCC 参考方法是否等价还要依据回归方程的 95% 偏移预测区间与设定偏移值比较后判断。根据 EP9-A2 文件,若设定偏移值 > 预测偏移区间的上限,示预测偏移小于可接受偏移,常规方法与 IFCC 参考方法等价(文献[4]中的 B、C 常规方法)。若设定偏移值 < 预测偏移区间的下限,示预测偏移超过可接受偏移,常规方法与 IFCC 参考方法不等价(文献[4]中的 D 常规方法)。若设定偏移值包含预测偏移区间,根据 EP9-A2 文件认为常规方法与 IFCC 参考方法等价(文献[4]中的 A 常规方法)。值得注意的是,当设定偏移值包含预测偏移区间时,根据评价方法测量结果仅部分符合评价标准,不表示两法完全相等,文献[4]中的 A 常规方法平均偏移为 -5.92%,已超出设定偏移值,按实际情况应判定为不等价,这与 EP9-A2 文件方法结论不符。另由于该法引入直线回归方程的 95% 偏移预测区间与设定偏移值比较条件可检出恒定偏移,但对随机误差检出仍不敏感(如文献[4]中的 C 常规方法)。事实上文献[4]中的 A 法、C 常规方法与 IFCC 法直线回归方程斜率相等、截距不等,因此两法血清测量结果并不相同,但根据 EP9-A2 文件方法判断结论相同。可见该法评价正确度是否存在标准较松的情况,尚需进一步证实。

20 世纪 80 代 Bland 等建立的 Bland-Altman 图形法是评价两种定量测量法一致性较好的简便、直观的统计学方法,通过对每一对测量值、测量差值、变异均值的观察,计算依据偏移均值、标准差、95% 置信区间评价恒定偏移的基本图形,观察图形变化趋势、测试数据是否有≥95% 的点落在“均值 ±2s”内。可同时观察系统误差与随机误差。对解决配对 t 检验、线性回归分析和 EP9-A2 文件方法评价“两法一致性”的不足有明显的帮助,其核心思想是描述被评价方法结果与两法均值的变异性(相对偏移)与两法均值的变量关系,以“被评价方法结果与两法均值的变异性均值 ±2s”,观察被评价测量方法结果是否落在上述界值范围内,若超过 95% 的观察点落在此范围,则两法一致。本法由于被评价方法结果是与两法均值做比较,不符合 ISO 18153[22]、ISO 15189 文件对测量结果正确度评价要求。

在充分理解 Bland-Altman 图形法设计思想的基础上,笔者曾对 Bland-Altman 图形法进行改进,建立以“IFCC 参考方法做比较方法、各酶常规方法为被评价方法、被评价方法与 IFCC 参考方法结果的变异性均值 ±2s”为界值的改良 Bland-Altman 图形法。由于比较方法为 IFCC 参考方法,因此,在加入被评价方法的平均偏移值小于 5% 的条件后,将改良 Bland-Altman 图形法用于 AMY 常规方法的正确性评价,可清晰识别由系统误差或随机误差引起的测量结果不正确,效果优于其他统计学方法。

EP15-A2 文件是实验室用于验证检验方法的测量性能与厂家的声明是否一致的文件,其中包括正确度验证内容,文件给出二种方法:“分析标准物质”方法和“方法学比较”方法。在采用“方法学比较”方法评价正确度时,一方面该文件没有明确规定比较方法必须是参考方法;另一方面在数据处理时,该文件需要通过配对 t 检验计算方法的平均差值(偏移)、差值的标准差、置信区间,将观测的偏移与厂家的声明进行比较。笔者以为,该文件的“比较方法”和所谓的“偏移”的评价标准有待明确。不适用于评价正确度。2010 年 5 月发布的“HOKLAS Supplementary Criteria No. 38”文件中明确规定在“方法学比较”中,如果比较的方法不是参考方法,则比较的差异不能用“偏移”表示。

5 结果表述

在采用“方法学比较”法评价定量测量方法正确度时,按不同标准/文件的统计学方法(或相应软件)处理数据时可能给出不同的结论,如等效、等价、正确、一致等。事实上这些术语表达的含义并不完全相同。笔者以为,用于正确度评价的比较方法通常为基准方法或其他公认的参考方法(JCTLM 列表参考方法),同时被评价方法与比较方法又具有测量一致性,可以理解为被评价方法与参比方法等效、等价、一致,其测量结果正确度性能应使用“正确“一词表述,可能更为恰当。关于如何正确表述“检验结果正确度”,行业内目前尚未形成一致观点。有些学者用“正确”表述,有些学者用“等效、一致、等价”等表述,这急需规范。如有可能,行业制定相关标准可能有助于此方法评价正确度结果表达的尽早规范。

6　小结

“方法学比较”法是一种评价临床检验定量测量方法正确度的可靠方法，其可信度依赖于严密的实验设计、恰当的实验样本、运行稳定的测量程序、训练有素的技术人员、适宜的数据处理方法和评价标准。相关行业标准的制定将有助于临床实验室建立规范的评价定量测量方法正确度的“方法学比较”方法并获得可靠的评价结论。

7　参考文献

[1] CLSI EP9-A2. Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline-Second Edition[S]. Clinical and Laboratory Standards instituition, 2002.

[2] JCGM 200. International vocabulary of metrology-basic and general concepts and associated terms(VIM)[S]. 2008.

[3] CLSI EP15-A2. User verification of performance for precision and trueness; Approved Guideline-Second Edition[S]. Clinical and Laboratory Standards instituition, 2006.

[4] 陈宝荣，孙慧颖，邵燕，等. 四种血清 α-淀粉酶厂家系统测量结果的正确性评价[J]. 中华检验医学杂志，2012，35(4)：1-7.

[5] 陈宝荣，孙慧颖，邵燕，等. 用 IFCC 参考方法评价 11 个常规系统测量血清 α-淀粉酶结果的正确性[J]. 临床检验杂志，2011，29(4)：309-313.

[6] HOKLAS Supplementary Criteria No. 38：“Medical Testing“Test Category-Performance Verification of Automated Analysers[S].

[7] ISO 15189. Medical laboratories-Particular requirements for quality and competence[S]. International Standards Organization, 2003.

[8] Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices[S]. 1999.

[9] GB/T 21415-2008. 体外诊断医疗器械生物样品中量的测量-校准品和控制物质赋值的计量学溯源性[S]. 2008.

[10] 王治国. 临床检验方法确认与性能验证[M]. 北京：人民卫生出版社，2009，177-180.

[11] Westgard JO. A method evaluation decision chart (MEDx chart) for judging method performance[J]. Clin Lab Sci, 1995, 8(5): 277-283.

[12] Bland JM, Altman DG. Statistical method for assessing agreement between two methods of clinical measurement[J]. Lancet, 1986, 1(8476): 307-310.

[13] Bland JM, Altman DG. Measuring agreement in method comparison studies[J]. Statistical Methods in Medical Research, 1999, 8(2): 135-160.

[14] Dewitte K, Fierens C, Stöckl D, *et al*. Application of the Bland-Altman plot for interpretation of method-comparison studies: a critical investigation of its practice[J]. Clin Chem, 2002, 48(5): 799-801.

[15] Hanneman SK. Design, analysis, and interpretation of method-comparison studies[J]. AACN Adv Crit Care, 2008, 19(2): 223-234.

[16] 陈卉. Bland-Altman 分析在临床测量方法一致性评价中的应用[J]. 中国卫生统计，2007，3(24)：308-310.

[17] Kost GJ, Tran NK, Abad VJ, *et al*. Evaluation of point-of-care glucose testing accuracy using locally-smoothed median abslute difference curves[J]. Clin Chim Acta, 2008, 389(1-2): 31-39.

[18] Howell DC. Statistical methods for psychology[M]. 2nd ed. Boston: Duxbury Press, 1987.

[19] 余松林. 医学统计学[M]. 北京：人民卫生出版社，2002：48-57.

[20] BéAdard M, Martin NJ, Krueger P, *et al*. Assessing reproducibility of data obtained with instruments based on continuous measurements[J]. Exp Aging Res, 2000, 26(4): 353-365.

[21] 栾荣生. 流行病学研究原理与方法[M]. 第 4 版. 成都：四川大学出版社，2002，114-121.

[22] ISO 18153. *In vitro* diagnostic medical devices measurement of quantities in biological samples metrological taceability of values for catalytic concentration of enzymes assigned to calibrators and control materials[S]. International Standards Oganization, 2003.

（**收稿日期**：2012-08-31）

（**本文编辑**：王海燕，陈维忠）