doi: 10.3978/j.issn.2095-6959.2015.01.006
View this article at: http://dx.doi.org/10.3978/j.issn.2095-6959.2015.01.006

·AME 科研时间专栏·

专栏导读：AME Groups旗下出版了*Journal of Thoracic Disease*（《胸部疾病杂志》）、*Annals of Cardiothoracic Surgery*（《心胸外科年鉴》）、*Chinese Journal of Cancer Research*（《中国癌症研究》）和*Annals of Translational Medicine*（《转化医学年鉴》）等近20本英文医学学术期刊。2014年，AME Groups中文平台——“科研时间”的诞生，为广大从事临床和基础研究的科研工作者带来了福音，提供了更多科研交流和学习分享的机会。欢迎广大读者关注我们“AME科研时间专栏”，订阅我们的公众微信号(科研时间：amegroups)，给我们提出宝贵的建议和意见，以便于将这个专栏建设得更好，成为读者喜闻乐见的一个栏目。

AME详解STARD——如何规范写作诊断准确性试验报告?

胡志德

(济南军区总医院实验诊断科，济南 250031)

长期以来，由于对学术论文撰写没有统一的规范要求，很多学者在撰写论文的时候往往是根据个人习惯进行撰写，导致论文遗漏了部分重要的研究信息，削弱了论文的学术穿透力和应用价值，阻碍了科研成果的传播和应用。学术论文撰写规范一直困扰着从事临床研究的学者和期刊编辑们，规范论文写作的呼声也越来越高。在此背景下，国际上陆续出现了一些针对不同试验设计类型的论文报告规范，比如针对观察性研究的STROBE声明、针对随机对照试验的CONSORT声明等。在此大背景下，STARD报告规范应运而生。

STARD报告规范的全称是“Standards for Reporting of Diagnostic Accuracy”，该规范是M. Bossuyt P，Reitsma J等诊断试验设计领域的权威流行病学家于2003年制定的。该报告规范共25条，详细规定了诊断准确性试验的论文在各个章节中需要报告的内容。初步估计，到2004年，全世界已有400多家杂志在其稿约中明确要求：如果作者投递的论文属于诊断准确性试验论文，则在写作上必须遵循STARD报告规范。这400余家杂志中，包括了大名鼎鼎的*BMJ*，*JAMA*，*Lancet*等。时至今日，认可STARD报告规范的学术期刊数目还在不断增加。值得一提的是，据笔者观察，国内尚无学术杂志在其稿约中明确规定作者投送的诊断准确性论文需要遵循STARD报告规范，也鲜有作者在其论文中说明研究报告遵循了STARD规范。总体而言，国内杂志刊登的诊断准确性试验论文中不够规范、甚至违反规范的现象仍然俯拾皆是。由此可见，推广和普及STARD报告规范，提升诊断准确性试验论文的报告质量，已迫在眉睫。

在本文中，笔者拟结合长期从事诊断准确性试验系统评价和Meta分析的经历，对STARD报告规范中条目进行解析。

1 论文的标题 / 摘要 / 关键词 (条目 1)

STARD报告规范对论文条目的要求就一条，英文原文为“Identify the article as a study of diagnostic accuracy (recommend MeSH heading ‘sensitivity and specificity’)”。实际上就是要求作者在论文标题/摘要/关键词中说明研究属于诊断准确性试验。该条目较易于理解：论文的标题中应带有Diagnostic accuracy之类的字样；论文的摘要中应该介绍诊断敏感性、特异性等，同时建议作者将敏感性、特异性列为关键词，因为在考科兰推荐的诊断准确性试验检索策略中，sensitivity和

收稿日期 (Date of reception)：2014–12–22
通信作者 (Corresponding author)：胡志德，Email: hzdlj81@163.com

specificity是推荐的检索词。大多数诊断准确性论文都遵循了该条目。

2 前言（条目 2）

STARD报告规范对前言的撰写要求也只有一条，原文为State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups。该条目实际上也体现了科研论文前言的一般撰写原则，即陈述研究的背景、目前面临的困境以及本研究力图解决的问题。比如研究NT-proBNP在呼吸困难人群中诊断心力衰竭的价值，作者需要介绍的背景就应该包括：呼吸困难在临床上出现的概率有多少？人们对于呼吸困难的病因诊断有何困惑？研究在呼吸困难人群中诊断心力衰竭有何临床价值？作者为何提出NT-proBNP可以用于在呼吸困难人群中诊断心力衰竭？既往是否有类似研究，这些研究有何缺陷导致有必要开展新的研究区阐述该科学问题？如果作者是比较NT-proBNP和另一诊断手段(比如BNP)，也需要在前言中说明这种比较的必要性和意义。

3 研究对象（条目 3~6）

STARD报告规范有11个条目对材料和方法的撰写进行了规范，其中有4个条目与研究对象的募集有关。

3.1 Describe the study population: the inclusion and exclusion criteria, setting and locations where the data were collected

解析：从字面意思上理解，实际上就是要说明研究的纳入和排除标准，说明研究对象的来源。多数研究均能详细交代研究对象的来源(比如来自某个医院或某个区域)，但却鲜有研究阐述纳入和排除标准。比如研究NT-proBNP在呼吸困难人群中诊断心力衰竭的能力，国内大多数研究论文的撰写方法是：本研究纳入心力衰竭患者和非心力衰竭患者各100名，非心力衰竭患者中包括肺栓塞患者30名，急性冠状动脉综合征30名，健康对照30名，主动脉夹层10名。这种论文撰写方法本身是不规范的，因为缺少对纳入和排除标准的介绍。正确的方法应该是描述在研究时是如何招募到这200名研究对象的？有无统一的纳入和排除标准？纳入排除标准分别是什么？

3.2 Describe participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the (evaluated) index tests or the (golden) reference standard?

解析：该条目实际上与3.1中的规定有一些重叠。严格来讲，纳入和排除标准主要是一些基于病史、症状和体征的信息。比如在示例中，纳入标准就应该是：以呼吸困难为主诉的患者；排除标准就应该是：虽有呼吸困难，但是根据病史就能明确诊断或者排除心力衰竭的患者，比如创伤患者。这种通过统一的纳入排除标准确定研究对象的设计称为“单门设计(one gate design)。在经过单门设计募集到的人群中，目标疾病的发生率与临床实践中的发生率是接近的，因此属于真实世界(real world)的研究，结果有较强的说服力。有的研究是“双门设计”(two gate design)，没有统一的纳入排除标准，而是预先假设需要与目标疾病(心力衰竭)进行鉴别的疾病(比如肺栓塞、肺炎、主动脉夹层等)，然后收集相关的病例进行分析(如上述3.1中中文杂志的写作方法)。由于这些研究中各组研究对象的样本量是随意假设的，并没有反映目标疾病的真实发病率，因此并非真实世界的研究，结果说服力不强。除了单门和双门设计以外，还有一类比较特殊的，基于标本的研究，比如尿白细胞对尿路感染的诊断价值的回顾性研究，研究对象可能是在某一段时间内接受尿培养和尿常规检查的患者。这种设计虽然有纳入排除标准，但是这些标准并不是基于症状和体征的，研究人群中可能混杂一部分不需要鉴别诊断的患者，比如复诊患者，因此这类研究也是欠缺说服力的。

3.3 Describe participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected

解析：该条目实际上要求研究者报道是否连续或随机招募研究对象。考虑到连续招募对象更容易获得大样本，因此多数学者采用连续招募的方式募集研究对象，其目的在于保证研究对象的代表性，确保研究属于“真实世界”的研究。假定不是连续招募研究对象，则可能存在研究对象选择偏倚：一些病情更重，临床表现更典型的患者更容易进入研究，而病情较轻的患者则容易被剔除出研究，最终可能夸大待评价试验的诊断价值。

3.4 Describe data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?

解析：根据数据收集方式，诊断准确性试验可以分为前瞻性试验和回顾性试验。前者是指预先制定研究方案，并按照研究方案纳入研究对象，执行待评价试验和金标准；后者则是指回顾性地分析已有的病例，从中挑选出合格的患者，评价已经执行的待评价试验对目标试验的诊断价值。前瞻性研究不容易遗漏病例，能较好地控制病例选择偏倚，且可以进行盲法设计，因为具有较高的说服力。在论文中报道研究的数据收集时序，有助于读者判断研究结论的可靠性和应用价值。

值得一提的是，在修订的诊断准确性试验质量评价工具(QUADAS-2)中，对病例选择偏倚程度的判断就是基于三个问题：1)是否连续招募研究对象？2)是否纳入健康对照？3)是否有不恰当的纳入排除标准？这实际上也体现了诊断准确性试验在招募对象时的设计要领，即：设立统一的纳入排除标准、连续招募所有对象。

4 待评价试验 (条目 7~11)

STARD规范中共有5个条目规定了如何陈述待评价试验的相关信息，分别如下。

4.1 Describe the reference standard and its rationale

解析：该条目研究者报道金标准的相关信息。需要强调的一点是：在诊断准确性试验中，金标准是用于划分实验组和对照组的依据。言外之意，所有的对象，不论实验组还是对照组，都必须接受金标准检查。部分研究论文在报告金标准的执行情况时，未报道对照组是否也执行了金标准检查，这是一种不规范的报告行为。在NT-proBNP诊断心力衰竭的研究中，我们假定对照组中有部分病人没有执行金标准检查，比如肺栓塞病人，虽然经过了影像学检查确认患者患有肺栓塞，但是并不能确定患者是否同时合并心衰，在此背景下直接将患者划为对照组是不科学的。

4.2 Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard

解析：该条目研究研究人员详细介绍待评价试验和金标准的特征，以及何时执行了上述检查。比如评价NT-proBNP在呼吸困难人群中诊断心力衰竭的能力，研究者需要交代的是：何时采集血液进行了NT-proBNP的检测；NT-proBNP的检测是采用的是哪种方法；何时执行了金标准；金标准是如何执行的。对上述问题的阐述都应引用相关文献。

4.3 Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard

解析：该条目实际上也是要求患者详细介绍待评价试验和金标准的诊断诊断界值。有的金标准或待评价试验是主观检查，往往需要论文作者陈述是如何确定界值的。比如研究免疫组化法检测直肠组织中的骨钙素诊断直肠癌的价值，就需要对骨钙素阳性和阴性的标本进行定义。作者应该在论文中阐述自己定义阳性或阴性的依据。

4.4 Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard

解析：该条目强调的是：如果金标准或待评价试验是主观检查(比如各种精神类量表、影像学检查)，应该陈述是执行金标准或待评价试验的人数？是如何培训金标准或待评价试验的执行者？金标准或待评价试验的执行者临床经验如何？等方面的内容。

4.5 Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers

解析：该条目要求作者陈述盲法问题，即检测结果对负责诊断临床医生是否设盲；临床资料对负责执行待评价试验的医生设盲。对于一些客观的检查，比如实验室指标(比如NT-proBNP)，临床资料是否对待评价试验执行者设盲可能对研究结果可靠性的影响不大。但是对于一些主观的指标，比如精神科的量表，临床资料对待评价试验执行者设盲就显得尤为必要了。同理，在执行金标准时不应该参考待评价试验的结果，以便更加客观地评价待评价试验的诊断价值。

5 统计学分析 (条目 12~13)

STARD规范中共有两个条目规定了统计学分析需要报道的内容。

5.1 Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals)

解析：该条目要求作者介绍评价诊断性能的统计学方法。对于连续变量，一般采用受试者工作特征曲线法，以曲线下面积的大小来反映待评价试验的诊断价值；对于两分类变量，则直接构建四格表，计算对应的敏感性和特异性。需要说明的是，在交代统计学结果时，应明确敏感性、特异性、曲线下面积95%的可信区间。在比较两个待评价试验诊断性能时，应说明采用何种方法比较诊断性能。以连续变量为例，一般通过比较曲线下面积的大小来反映指标的鉴别能力，具体的统计学方法可以参考相关书籍。

5.2 Describe methods for calculating test reproducibility, if done

解析：该条目要求患者报道待评价试验和金标准的重复性或精密度。比如评价NT-proBNP对心力衰竭的诊断价值，金标准是两名临床医生综合临床表现和检查做出的判断。研究者在撰写论文时需要报道如何评价NT-proBNP的检测性能，特别是检测精密度。同时还需要报道如何评价金标准的重复性(即两名临床医师之间的一致性)。可以想象的是，如果两名临床医生诊断心力衰竭之间的一致性不好，说明金标准不够可靠，研究的说服力也将大打折扣。

6 研究对象的特征 (条目 14~16)

从条目14开始为作者需要在结果中阐述的内容，包括研究对象、检测结果和诊断性能等。条目14~16为报道研究对象特征的相关要求。

6.1 Report when study was done, including beginning and ending dates of recruitment

解析：该条目较容易理解，即报道试验开始的时间和日期。比如：从2008年1月开始募集研究对象，截止至2010年1月。大多数研究论文遵循了该条目的要求。

6.2 Report clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers)

解析：该条目要求研究者尽可能介绍研究对象的临床特征，一般通过列表的形式展示。这部分数据尤为重要，因为读者主要通过这部分数据了解研究对象的特征，并估计研究结论的适用范围。部分待评价试验与患者疾病分期密切相关，比如肿瘤标志物，因此研究者应对此方面的数据进行重点展示。

6.3 Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended)

解析：如前所述，在招募研究对象时，应采用统一的纳入排除标准，并连续招募研究对象。然而，在实际研究中，这种目标往往难以实现，比如有的研究对象可能不愿意签署知情同意书而无法进入研究，有的研究对象由于标本不足而无法进行实验室指标的检测等。在研究论文中，应详细阐述研究对象的募集流程，对于各种原因被剔除出研究的对象，应予以说明。STARD报告规范推荐流程图进行说明。

7 检测结果 (条目 17~20)

7.1 Report time interval from the index tests to the reference standard, and any treatment administered between

解析：该条目要求作者报道执行待评价试验和金标准的“时间差”，以及中间可能经历的治疗过程。一般而言，待评价试验和金标准最好在同一时间执行。如果不在同一时间执行，应该说明上述“时间差”对结果是否有影响。比如某些研究使用冻存血清进行研究，应说没冻存条件对检查结果是否有影响。

7.2 Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition

解析：简而言之，该条目研究报道实验组患者的疾病严重程度，比如肿瘤的分期、心衰的分级等。同时还需要报道对照组的病例组成。比如评价DCP对肝癌的诊断价值，研究者就需要在结果中陈述肝癌患者的分期、非肝癌患者的组成状况。

7.3 Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard

解析：该条目研究研究者以表格或图片的方

式陈述待评价试验在试验、对照组各个疾病中的分布情况。以DCP诊断肝癌为例，研究者需要陈述DCP在肝癌和各种非肝癌患者(比如肝炎、肝硬化)中的分布情况等。需要说明的是，部分待评价试验的结果可能难以确定，也应该在报告中说明。比如评价NT-proBNP对心力衰竭的诊断价值，可能会出现一些意想不到的情况，比如：部分患者的血液标本溶血严重，或者血浆不足，以至于仪器无法检测出结果；部分符合纳入条件的患者在采血前已经死亡等。这些情况均应在论文报告中说明。

7.4 Report any adverse events from performing the index tests or the reference standard

解析：大部分诊断准确性试验没有报道此条目的内容。该条目实际上是指报道在执行待评价试验或金标准过程中出现的不良事件。比如研究针吸细胞学检查对乳腺癌的诊断价值，研究者应该报道患者在接受针吸细胞学检查的过程中，出现的副反应(比如感染、出血等)，以便读者判断针吸细胞学检查的安全性。在诊断准确性试验中，人们关注的不仅仅是诊断准确性问题，同时也关注试验的安全性问题。

8 诊断性能 (条目 21~24)

8.1 Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals)

解析：本条目要求研究者报道诊断性能的相关指标及其95%可信区间，包括：敏感性、特异性、阳/阴性似然比、阳/阴性预测值、曲线下面积，诊断阈值等。一般建议以表格的形式展示。

8.2 Report how indeterminate results, missing responses and outliers of the index tests were handled

解析：本条目研究患者报道难以解释的中间结果，无法获取的结果最终是如何处理的。该条目与7.3中的条目实际上是一脉相承的，7.3中的条目强调的是作者应该报告是否有难以解释或者遗漏的中间结果，8.2中的条目则强调作者应报告这些数据是如何处理的。比如研究尿常规对尿路感染的诊断价值，尿路感染的金标准是细菌培养。但是部分研究对象在进行细菌培养时由于取材不当，导致尿液被污染了，无法执行金标准，也无从知晓患者是否患有尿路感染。这些研究对象就属于遗漏的结果，应在论文中说明这些数据是如何处理的(一般是剔除出研究)。

8.3 Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done

解析：该条目要求患者报道亚组分析的结果，为可选条目。比如评价NT-proBNP对心力衰竭的诊断价值，已知肾功能对NT-proBNP的结果有影响。研究者可以根据肾功能将研究对象分为肾功能受损组和无肾功能受损组，在两组中分别评价NT-proBNP对心力衰竭的诊断价值。

8.4 Report estimates of test reproducibility, if done

解析：该条目实际上与条目13是对应的。在条目13中介绍了待评价试验和金标准性能的分析方法，这里主要报道结果，包括精密度、线性、日间变异、批间变异等。对于主观的标准，应报道观察者之间的一致性。

9 讨论 (条目 25)

解析：诊断准确性试验队论文讨论部分的要求只有一条：Discuss the clinical applicability of the study findings。实际上就是要求作者讨论研究的临床价值。比如研究结果提示待评价试验队目标疾病的诊断价值如何？对疾病的诊断流程有何启示？对疾病的治疗模式有何启示等。当然，在具体的论文撰写过程中，为了提升论文的可读性和科学性，还应该同时讨论研究的创新性等问题。

作者：胡志德，AME学术沙龙委员、Section Editor（Systematic Review and Meta-analysis），工作于济南军区总医院实验诊断科，现为第二军医大学临床检验诊断学博士研究生。担任*Tumor Biology*，*Journal of Thoracic Disease*（*JTD*）等杂志审稿专家，以第一作者或通讯作者身份发表SCI论文13篇，并主持国家青年科学基金一项。

本文引用：胡志德. AME详解STARD——如何规范写作诊断准确性试验报告?[J]. 临床与病理杂志, 2015, 35(1): 14-18. doi: 10.3978/j.issn.2095-6959.2015.01.006

本文首先以中文发表于【科研时间】(doi: 10.3978/kysj.2014.1.369). 本文已获科研时间和作者同意将该文内容以中文在本刊发表。