

迈向完整、准确的诊断准确性研究报告： STARD计划

Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative

Patrick M. Bossuyt^{1*}, Johannes B. Reitsma¹, David E. Bruns^{2,3}, Constantine A. Gatsonis⁴,
Paul P. Glasziou⁵, Les M. Irwig⁶, Jeroen G. Lijmer¹, David Moher⁷, Drummond Rennie^{8,9},
Henrica C.W. de Vet¹⁰, for the STARD Group

作者单位

1. 荷兰阿姆斯特丹大学学术医学中心
临床流行病学与生物统计学系（阿姆斯特丹1100DE）
2. 美国夏洛茨维尔吉尼亚大学病理系
（维吉尼亚VA22903）
3. 美国华盛顿临床化学杂志（华盛顿
DC20037）
4. 美国普罗维登斯布朗大学统计学中
心（罗德岛RI02912）
5. 澳大利亚昆士兰大学全科实习中心
（昆士兰QLD4006）
6. 澳大利亚悉尼大学公共卫生与社区
医学系（悉尼NSW2006）
7. 加拿大安大略省渥太华Chalmers研究
组（渥太华K1N6M4）
8. 美国旧金山加利福尼亚大学卫生政
策研究学院（加利福尼亚CA94118）
9. 美国芝加哥美国医学会杂志（伊利
诺伊IL60610）
10. 荷兰阿姆斯特丹自由大学外部研究
学院（阿姆斯特丹1081BT）

*通讯作者: Department of Clinical
Epidemiology and Biostatistics, Academic
Medical Center, University of Amsterdam,
PO Box 22700, 1100 DE Amsterdam, The
Netherlands. Fax 31-20-6912683;
Email: Stard@amc.uva.nl.

摘要

背景 要理解诊断准确性研究的结果，读者必须了解其设计、实施、分析及结果。这只能通过作者完全公开研究资料才能实现。

目的 提高诊断准确性研究报告的准确性和完整性，以便读者评估研究的偏倚可能性及推广性。**方法** 诊断准确性研究报告标准（STARD）指导委员会通过文献检索，找出正确实施和报告的诊断性试验，并提取可能有用的条目形成一份泛表。再由研究人员、编辑及相关专业组织成员在为期两天的共识会议中对这份泛表进行缩减，以制定出一份诊断准确性研究的清单和通用流程图。

结果 检索已发表的诊断性研究指南，共找到33份清单，从中提取出75项可能有用的条目。根据当前可得的所有提示偏倚的证据，将之缩减至25项条目。制定出的典型流程图提供了招募患者的方法、执行“诊断措施”的顺序及接受试验诊断措施或/和金标准诊断措施的患者人数等信息。

结论 判断研究的价值有赖于完整、准确的报告。若医学期刊采用此清单和流程图，诊断准确性研究报告的质量将会有所提高，更好地服务于临床医生、研究人员、评价者、期刊及公众。

诊断措施（Diagnostic tests）千变万化，新检查方法不断涌现，已有检查技术日趋完善。设计和报告质量不佳的诊断研究所得出的夸大或偏倚结果会导致某些技术的过早推广，使临床医生作出错误的治疗决策。在某项诊断措施用于临床前进行严格评估，不仅能减少因错误估计其准确性而导致意外临床结果发生外，还能避免不必要的检查，降低医疗费用。因此，确定某项诊断措施诊断准确性的研究在这一评估程序中至关重要^[1-3]。

在诊断准确性研究中，将一个或多个待评估诊断措施所得结果与金标准诊断措施所得结果相比较，两组均在可能有目标病情的受试者中进行。在这里，“检查（test）”一词是指任何用于进一步

获取病人健康状况的方法，包括从病史和查体、实验室检查、影像学检查、功能检查和组织病理学检查中所获的信息。“目标病情”是指某种具体疾病或其它任何可能需进行临床干预（如尝试进一步诊断措施，或开始、调整或终止某项治疗）的显在病态表现。在该体系中，“金标准”被认为是当前已有的确定目标病情存在与否的最佳方法。金标准可以是单一方法，也可以是几种方法的组合。它不仅包括实验室检查、影像学检查、病理学检查，还包括对受试者专门的临床随访。“准确性”是指试验诊断措施（待测指标，index test）与金标准所得结果之间的吻合程度。诊断准确性可以用多种方式来表示，包括灵敏度和特异度、似然比、诊断优势比

及受试者工作特征（ROC）曲线下面积^[4-6]。

诊断准确性研究的内在和外在真实性存在着多种潜在威胁。一篇分析了1978至1993年间发表于四种主要医学期刊的诊断准确性研究的调查报告提示，这些研究的方法学质量充其量也就处于中等水平^[7]。但由于许多报告均缺少有关诊断研究的设计、实施、分析等关键要素的信息，该评价也因此

受到了局限^[7]。一些Meta分析的作者也证实了诊断研究的设计、实施等关键信息缺失的事实^[8,9]。同其它任何类型的研究一样，研究设计的缺陷可导致结果出现偏倚。有报告称，带有特定设计特征的诊断研究，与没有这类缺陷的研究相比，可能作出偏倚的、乐观的诊断准确性评估^[10]。

1999年于罗马召开的Cochrane学术年会上，

表 1 STARD诊断准确性研究报告清单
Table 1 STARD checklist for the reporting of studies of diagnostic accuracy

章节及主题	条目#	页数#
题目/摘要/关键词	1	确定文章为诊断准确性研究（推荐采用MESH主题词：“灵敏度和特异度”）
引言	2	陈述研究问题或目的，如评估诊断准确性或比较不同诊断措施间或不同受试人群间的准确性
方法	描述	
受试者	3	研究人群：纳入和排除标准，数据收集的背景和地点
	4	受试者募集：基于症状表现、既往检查结果或是受试者已接受过待测措施或金标准的事实？
	5	样本抽样：研究人群是否是根据条目3、4中的选择标准确定的样本序列？若“否”，具体说明是如何进一步选择样本的
	6	数据收集：数据收集是在指标检查和参考标准检查进行之前（前瞻性研究）还是之后（回顾性研究）？
检查方法	7	参考标准及其原理
	8	相关资料与方法的技术说明，包括何时和怎样进行测量的，和/或引用参考文献说明指标检查和参考标准检查
	9	指标检查和参考标准检查结果的单位、临界值和/或种类的定义及原理
	10	执行和读取指标检查和参考标准检查的人员的数量、培训及其专业技能
	11	读取指标检查和参考标准检查结果的人员是否不知道另一组检查的结果（盲法），描述其它任何可获得的临床信息
统计学方法	12	计算和比较诊断准确性的方法，以及不确定性的定量统计方法（如95%可信区间）
	13	如果已做，计算检查可重复性的方法
结果	报告	
受试者	14	研究何时完成，包括样本募集的起始日期
	15	研究人群的临床特征和人口统计学特征（如年龄、性别、出现的症状谱、合并疾病、现有治疗、样本募集中心）
	16	符合纳入标准，参与了或未参与指标检查和/或参考标准检查的人数；说明受试者为什么没有接受任一检查（强力推荐作出流程图）
检查结果	17	从指标检查至参考标准检查间的时间间隔，及在此期间是否予以了任何治疗
	18	在出现了目标情况的受试者中描述疾病严重程度（定义标准）的分布；在未出现目标情况的受试者中说明其它诊断
	19	指标检查结果（包括不确定的和丢失的结果）和参考标准检查结果的四格表；以及为获得进一步结果，检查结果的分布和参考标准检查结果的四格表
	20	在进行指标检查和参考标准检查的过程中是否出现了任何不良事件？
评价	21	诊断准确性的评估及统计学不确定性的衡量（如95%可信区间）
	22	指标检查的不确定结果、丢失结果以及异常值是如何处理的
	23	如果已做，评估诊断准确性在受试者亚组间、结果读取者间、各中心间的变异度
	24	如果已做，评估检查的可重复性
讨论	25	讨论研究结果的临床适用性

Cochrane诊断与筛查检查方法工作组讨论了诊断性试验方法学质量低、报告不标准的问题。工作组认为解决这些问题的第一步在于提高诊断性研究报告的质量。在CONSORT计划成功实施后^[11-13]，工作组致力于制定一份有关诊断准确性研究报告条目的清单。

制定诊断准确性报告标准（STARD）旨在提高诊断准确性研究报告的质量。完整、准确的报告便于读者发现研究中可能存在的偏倚（内在真实性），评价研究结果的普适性和推广性（外部效度）。

1 资料与方法

STARD指导委员会（全体委员及详细资料见附录）全面检索Medline, Embase, BIOSIS及Cochrane协作网方法学数据库（截至2000年7月），查找已发表的有关诊断性研究的实施和报告的文献。并筛查检出文献的参考文献，检索个人档案，并咨询了诊断性研究领域的专家。回顾分析了所有相关文献后，指导委员会提取出所有可能有用的条目，形成了一份泛表。

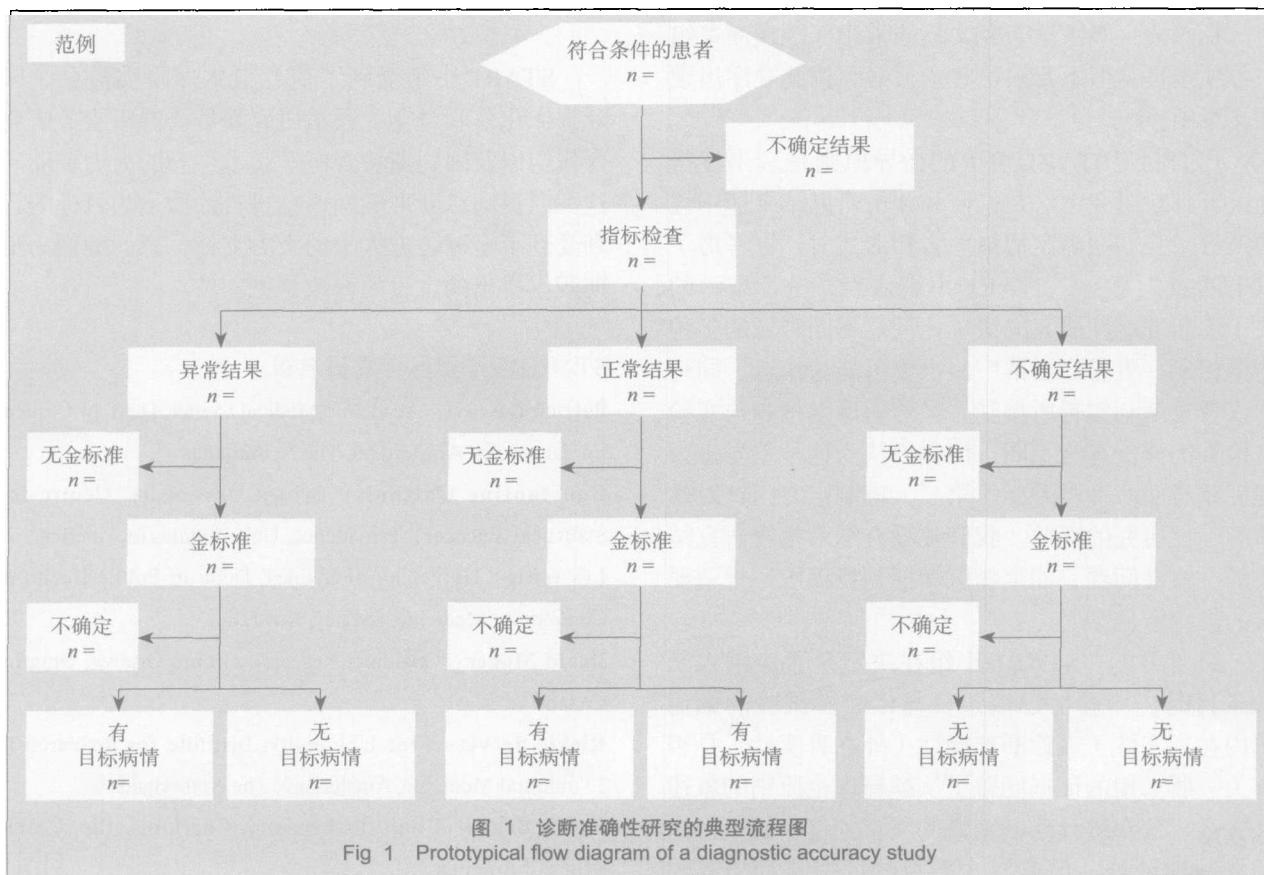
随后，STARD指导委员会邀请了来自研究人员、编辑、方法学家及专业组织等群体的专家，召开了一次为期两天的共识会议。该会议旨在是恰当缩减泛表，并讨论该清单的最佳格式和措辞。条目的保留与否尽可能循证决定。

此次会议由小组会议和全体会议共同构成。每个小组专攻列表中的一组相关条目。然后将各小组的建议在全体会议上进行讨论。指导委员会经过一夜的工作，根据各小组的建议和全体会议的进一步修改意见形成了STARD清单初稿。所有与会者于次日对初稿进行讨论并作进一步修改。STARD小组成员可通过电子邮件在下一轮讨论中提出进一步修改意见。

一些可能成为清单用户的人现场测试了会议版本的清单和流程图，从他们那里收集了一些评论意见。该版本之前曾放在CONSORT网站上，收集各方评论。STARD指导委员会讨论了所有评论意见，形成了最终的清单。

2 结果

检索出33个已发表的关于诊断性研究的临床指



南表单。基于这些表单以及指导委员会和STARD小组成员的建议，指导委员会制定出了一个含75项条目的列表。在2000年9月16、17日的会议上与会者达成共识，合并、删除部分条目，最终形成含25项条目的清单。清单的语法和格式的主要修改由参会人员完成。

会后，在评价的不同阶段时期，STARD小组收到了许多宝贵的评论和意见，据此确定了表1中的STARD清单。

流程图提供了患者募集方法（如基于有某种特定症状患者的连续序列，病例对照）、试验执行的次序、接受被评价的检查（指标试验）和参考检查的患者人数等信息（见图1）。

我们提供了一个反映诊断研究中最常用设计方案的典型流程图。关于其它设计方案的例图请参见STARD网站（www.consort-statement.org.htm）。

3 讨论

STARD计划的目的在于提高诊断研究报告的质量。清单中的条目和流程图有助于作者描述研究的设计与实施、试验的执行以及结果等方面的要点。

我们按一般医学研究文献的常用标题顺序来组织各项条目，但不要求这些条目必须按此顺序出现在文章中。

我们制定STARD清单的指导原则是选出能帮助读者判断研究中潜在偏倚和评价研究结果适用性的条目。我们在确定清单内容和形式时，也考虑了以下两点。第一，STARD小组认为，一个统一的关于诊断准确性研究的清单比针对不同领域的多种清单相比，更易广泛推广，也更可能为作者、同行评议者及期刊编辑所接受。虽然影像学检查与实验室检查的评价有所不同，但我们认为这种差异多为程度差异而非种类差异。第二，应制定专门针对诊断准确性研究的清单。我们并没有纳入报告研究结果的一般性问题，如生物医学期刊投稿统一规范要求的推荐意见^[14]。

只要可能，STARD小组在决定是否应纳入某一条目时，均会考虑该条目与诊断性试验的偏倚（内在真实性）或诊断准确性（外在真实性）是否有关。所采用的证据涵盖了从解释理论原则的叙述性文章、公布统计建模结果的文章到诊断研究得出的经验性证据。但有几个条目的证据是相当有限

的。

另有一篇背景文章阐释了每项条目的含义和原理，并简要概括了所用证据的类型和数量^[15]。这一背景文章将有助于STARD清单的应用、理解和推广。

STARD小组在诊断研究流程图的制定上投入了大量的精力。流程图能以透明的方式来交流有关研究设计及研究对象流向的重要信息^[16]。一份类似的流程图已成为CONSORT随机试验报告标准中的一个关键组成部分。鉴于诊断研究采用的设计方案各异，流程图对于诊断研究的意义更为重要。诊断准确性研究报告中的流程图可提示抽样和选择研究对象的过程（外在真实性）、研究对象的流向与试验时间及结果关系、未接受指标试验和/或参考标准检查的研究对象人数可能产生确认性偏倚，参考文献^[17-19]以及研究各阶段的研究对象人数，从而为比例的计算提供了正确的分母（内部一致性）。

STARD小组计划采用前-后效评价来衡量本声明对发表的诊断准确性研究报告的质量的影响^[13]。当有提示偏倚或变异度来源的新证据时，将会对STARD进行更新。我们欢迎各方意见，无论是关于内容的还是形式的，以此来进一步完善当前的版本。

STARD小组受到了荷兰卫生保险委员会、国际临床化学联合会、医学研究委员会健康服务研究合作组以及阿姆斯特丹医学学术中心的大力资助。这一以提高诊断准确性研究报告的质量为目标的计划受到了全球各方人士的大力支持，感谢他们为早期版本提出的宝贵的修改意见。

STARD指导委员会成员名单

Patrick Bossuyt: Academic Medical Center, Dept. of Clinical Epidemiology, Amsterdam, The Netherlands

Constantine Gatsonis: Brown University, Centre for Statistical Sciences, Providence, United States of America

Les Irwig: University of Sydney, Dept. of Public Health & Community Medicine, Sydney, Australia

David Moher: Chalmers Research Group, Ottawa, Ontario, Canada

Riekje de Vet: Free University, Institute for Research in Extramural Medicine, Amsterdam, The Netherlands

David Bruns: Clinical Chemistry, Charlottesville, United States of America

Paul Glasziou: Mayne Medical School, Dept. of Social & Preventive Medicine, Herston, Australia

Jeroen Lijmer: Academic Medical Center, Dept. of Clinical Epidemiology, Amsterdam, The Netherlands

Drummond Rennie: Journal of the American Medical Association, Chicago, United States of America

STARD小组成员名单

Doug Altman, Institute of Health Sciences, Centre for Statistics in Medicine (Oxford, United Kingdom);

Stuart Barton, British Medical Journal, BMA House (London, United Kingdom);

Colin Begg, Memorial Sloan-Kettering Cancer Center, Department Epidemiology & Biostatistics (New York, NY);

William Black, Dartmouth Hitchcock Medical Center, Department of Radiology (Lebanon, NH);

Harry Buijler, Academic Medical Center, Department of Vascular Medicine (Amsterdam, The Netherlands);

Gregory Campbell, US FDA, Center for Devices and Radiological Health (Rockville, MD);

Frank Davidoff, Annals of Internal Medicine (Philadelphia, PA);

Jon Deeks, Institute of Health Sciences, Centre for Statistics in Medicine (Old Road, United Kingdom);

Paul Dieppe, Department Social Medicine, University of Bristol (Bristol, United Kingdom);

Kenneth Fleming, John Radcliffe Hospital, (Oxford, United Kingdom);

Rijk van Ginkel, Academic Medical Center, Department of Clinical Epidemiology (Amsterdam, The Netherlands);

Afina Glas, Academic Medical Center, Department of Clinical Epidemiology (Amsterdam, The Netherlands);

Gordon Guyatt, McMaster University, Clinical Epidemiology and Biostatistics (Hamilton, Canada);

James Hanley, McGill University, Department Epidemiology & Biostatistics (Montreal, Canada);

Richard Horton, The Lancet, (London, United Kingdom);

Myriam Hunink, Erasmus Medical Center, Department of Epidemiology & Biostatistics (Rotterdam, The Netherlands);

Jos Kleijnen, NHS Centre for Reviews and Dissemination (York, United Kingdom);

Andre Knottnerus, Maastricht University, Netherlands School of Primary Care Research (Maastricht, The Netherlands);

Erik Magid, Amager Hospital, Department of Clinical Bio-

chemistry (Copenhagen, Denmark);

Barbara McNeil, Harvard Medical School, Department of Health Care Policy (Boston, MA);

Matthew McQueen, Hamilton Civic Hospitals, Department of Laboratory Medicine (Hamilton, Canada);

Andrew Onderdonk, Channing Laboratory (Boston, MA);

John Overbeke, Nederlands Tijdschrift voor Geneeskunde (Amsterdam, The Netherlands);

Christopher Price, St Bartholomew's - Royal London School of Medicine and Dentistry (London, United Kingdom);

Anthony Proto, Radiology Editorial Office (Richmond, VA);

Hans Reitsma, Academic Medical Center, Department of Clinical Epidemiology (Amsterdam, The Netherlands);

David Sackett, Trout Centre (Ontario, Canada);

Gerard Sanders, Academic Medical Center, Department of Clinical Chemistry (Amsterdam, The Netherlands);

Harold Sox, Annals of Internal Medicine (Philadelphia, PA);

Sharon Straus, Mt. Sinai Hospital (Toronto, Canada);

Stephan Walter, Mc-Master University, Clinical Epidemiology and Biostatistics (Hamilton, Canada).

参 考 文 献

- Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *Can Med Assoc J*, 1986; 134: 587-594.
- Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making*, 1991; 11: 88-94.
- Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol*, 1992; 27: 245-254.
- Griner PF, Mayewski RJ, Mushlin AI, *et al.* Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Ann Intern Med*, 1981; 94: 557-592.
- Sackett DL, Haynes RB, Guyatt GH, *et al.* The selection of diagnostic tests. In: Sackett D, editor. *Clinical epidemiology*, 2nd ed. Boston/Toronto/London: Little, Brown and Company; 1991: 47-57.
- Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*, 1978; 8: 283-298.
- Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA*, 1995; 274: 645-651.
- Nelemans PJ, Leiner T, de Vet HCW, *et al.* Peripheral arterial disease: Meta-analysis of the diagnostic performance of MR angiography. *Radiology*, 2000; 217: 105-114.
- Devries SO, Hunink MGM, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol*, 1996; 3: 361-369.
- Lijmer JG, Mol BW, Heisterkamp S, *et al.* Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 1999; 282: 1061-1066.
- Begg C, Cho M, Eastwood S, *et al.* Improving the quality of

- reporting of randomized controlled trials. The CONSORT statement. *JAMA*, 1996; 276: 637-639.
- 12 Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallelgroup randomized trials. *JAMA*, 2001; 285: 1987-1991.
 - 13 Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials. A comparative before-andafter evaluation. *JAMA*, 2001; 285: 1992-1995.
 - 14 International Committee of Medical Journal Editors. Uniform Requirements for manuscripts submitted to biomedical journals. *JAMA*, 1997; 277: 927-934. Also available at: ACP Online, <http://www.acponline.org>.
 - 15 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. The STARD Statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*, 2003; 49: 7-18.
 - 16 Egger M, Ju'ni P, Barlett C. Value of flow diagrams in reports of randomized controlled trials. *JAMA*, 2001; 285: 1996-1999.
 - 17 Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. *Med Decis Making*, 1987; 7: 139-148.
 - 18 Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. *Med Decis Making*, 1987; 7: 115-119.
 - 19 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*, 1987; 6: 411-423. 6 Bossuyt *et al.*: The STARD Initiative.

原载 *Clinical Chemistry*, 2003; 49(1): 1~6

刁 襄 译 秦 莉 姚 巡 审 校

收稿日期: 2006-05-11 修回日期: 2006-07-11

本文编辑: 蔡羽嘉