

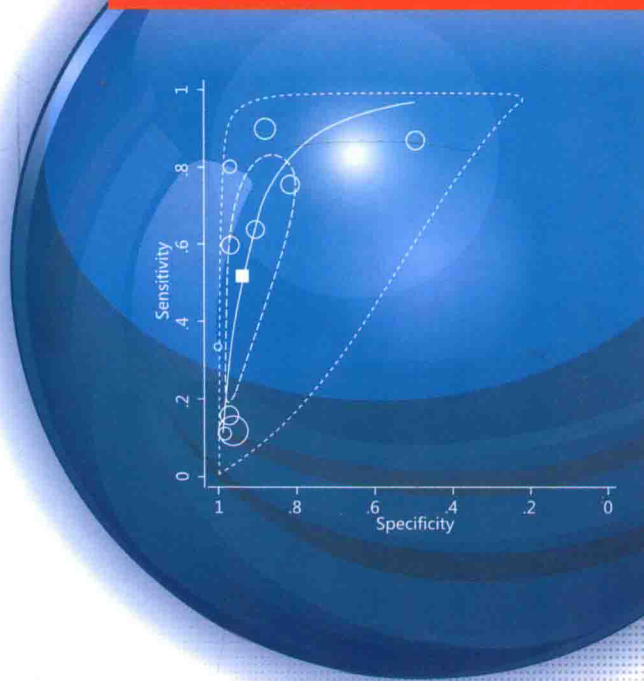
系统评价指导丛书

总主编 杨克虎

诊断试验

系统评价 / Meta分析 指导手册

主编 田金徽 陈杰峰



医药科技出版社

系统评价指导丛书

总主编 杨克虎

诊断试验系统评价/ Meta分析指导手册

主编 田金徽 陈杰峰

编者 (以姓氏笔画为序)

马 彬 田金徽 刘雅莉 杨克虎

李 伦 陈可欣 陈杰峰 陈耀龙

拜争刚 姚 亮 葛 龙

中国医药科技出版社

内 容 提 要

本书作为国内第一本诊断试验系统评价指导手册,有以下特色:第一,详尽阐述了诊断试验系统评价的起源、发展和特点以及制作方法和注意事项;第二,在讲解理论知识的同时,选取大量有代表性的实例深入剖析解读,以便于读者理解和灵活运用;第三,对诊断试验系统评价中软件的介绍,深入浅出,操作性强;第四,首次详细介绍了 GRAD 方法在诊断试验系统评价中的应用。

本书逻辑清晰,结构严密,章节安排得当,概念论述清晰,是诊断试验系统评价方法学感兴趣读者的首选。同时本书适合临床医学、药学及相关专业人员参考使用。

图书在版编目 (CIP) 数据

诊断试验系统评价: Meta 分析指导手册/田金徽, 陈杰峰主编. —北京: 中国医药科技出版社, 2015. 6

(系统评价指导丛书)

ISBN 978-7-5067-7568-7

I. ①诊… II. ①田… ②陈… III. ①流行病学-研究方法-手册 IV. ①R181.2-62

中国版本图书馆 CIP 数据核字 (2015) 第 110439 号

美术编辑 陈君杞

版式设计 郭小平

出版 中国医药科技出版社

地址 北京市海淀区文慧园北路甲 22 号

邮编 100082

电话 发行: 010-62227427 邮购: 010-62236938

网址 www.cmstp.com

规格 787×1092mm $\frac{1}{16}$

印张 16 $\frac{1}{4}$

字数 309 千字

版次 2015 年 6 月第 1 版

印次 2015 年 6 月第 1 次印刷

印刷 三河市百盛印装有限公司

经销 全国各地新华书店

书号 ISBN 978-7-5067-7568-7

定价 38.00 元

本社图书如存在印装质量问题请与本社联系调换

P 序 言

REFACE

系统评价发展至今已逾 20 年。在医学领域的诸多学科中，实属年轻，但从其本身的成长来看，其速度可谓空前：数量方面，仅 2014 年 MEDLINE 收录的 Meta 分析就达 10000 多篇，且每年以超过 20% 的速度递增；研究范围方面，从干预性系统评价迅速延伸到全面关注诊断、病因、预后、公共卫生和卫生政策等领域；研究类型方面，出现了系统评价再评价 (overview of review)、网状 Meta 分析 (Network Meta analysis)、快速系统评价 (rapid review)、概览系统评价 (scoping review)、证据地图 (evidence mapping) 等；研究方法方面，从基线风险、亚组分析，到检索策略和数据库的研究，再到统计学模型的选择，以及纳入研究的质量评价，无不更加系统和深入。*The Journal of Clinical Epidemiology* 最新发表的一篇文章甚至研究了在制作系统评价时，对比研究人员使用两个电脑显示器和单独使用一个显示器的问题。结果发现，同时使用两个显示器效率更高，尤其是在提取数据方面，每篇文章平均节省 23 分钟！

兰州大学循证医学中心引入、学习、制作和培训系统评价，迄今也近 10 年。这期间通过不断邀请国内外一流的方法学家前来讲学，持续派出中心师生赴国内外知名的学术机构进修，逐渐积累了较为丰富的系统评价制作经验。目前我们不仅在多个研究领域发表系统评价 100 余篇，同时也在国内较早开设了一系列系统评价培训班。在不断培训和被培训的过程中，在与国内外专家、学员、编辑和使用者的交流与互动中，把很多感悟、想法、经验和心得忠实记录，反复雕琢，并于 2010 年，推出了当时国内第一本《系统评价指导手册》。

系统评价及其方法学的持续发展和完善，以及国内研究者对系统评价越来越大的兴趣和需求，促使我们重新思考如何为读者奉献适应系统评价方法学发展的实用教材。我们打算推出系统评价指导丛书，旨在系统、全面和有针对性地指导如何撰写不同类型的系统评价。这次首先与读者见面的，是田金徽博士与陈杰峰博士主编的《诊断试验系统评价/Meta 分析指导手册》。

田金徽博士是我国较早学习、制作和教授诊断试验系统评价的研究人员之一。2007年,由他指导撰写的“有关硝酸盐还原法检测结核分枝杆菌耐药性的文献系统评价”在《临床检验杂志》发表,其后他本人及其研究团队在国内外期刊连续发表有关诊断试验系统评价及其方法学的论文已达27篇,并多次被邀请到全国有关培训班和学术会议讲授诊断试验系统评价方法。2011年~2012年间,田金徽博士专门赴英国 University of East Anglia 访学,师从国际系统评价资深专家宋福建教授学习间接比较和网状 Meta 分析方法。2014年7月,经过认真组织和筹备,由他主持在兰州大学循证医学中心成功举办了首届《诊断试验系统评价/Meta 分析强化培训班》,来自上海、广东、贵州、江西和甘肃等地40余位学员参加,取得了良好的效果。通过对国内外现有系统评价相关教材的深入研读和学习,结合在撰写系统评价方面的积累的经验和体会,田金徽博士组织中心骨干教师和研究生,联合台湾实证医学学会陈杰峰博士团队经过一年多的充分准备,共同完成了本书的编写工作。

作为国内第一本诊断试验系统评价指导手册,本书有以下特色:第一,详尽阐述了诊断试验系统评价的起源、发展和特点,以及制作方法和注意事项;第二,在讲解理论知识的同时,选取了大量有代表性的实例深入剖析解读,以便读者能够易于理解和灵活运用;第三,对诊断试验系统评价中软件的介绍,深入浅出,操作性强;第四,首次详细介绍了 GRADE 方法在诊断试验系统评价中的应用。本书逻辑清晰,结构严密,章节安排得当,概念论述清晰,在当前尚未有诊断试验系统评价方法学专著面世的情况下(Cochrane 诊断试验系统评价指导手册迄今尚未出版),本书应是对该主题感兴趣读者的首选。

下一步,我们将推出本丛书的其他分册,目前正在撰写的有《非随机试验系统评价》、《网状 Meta 分析》、《GRADE 系统介绍与实例解析》等。我们期待该丛书能够对广大系统评价的制作者和使用者有所帮助。

最后,特别感谢给予我们支持的国内外同行专家,是他们不断帮助我们开拓视野,完善思路,鞭策和鼓励我们向一流学习,努力站在循证医学研究的最前沿;更要感谢过去十年间在我们中心学习、培训过的来自全国各地的所有学员。正是他们宝贵的反馈、想法、建议和意见,催生了这套丛书的出版,也同时激励着我们不断学习、成长与进步。

杨克虎

2015年1月

C 目 录

CONTENTS

第一章 诊断试验简介	1
第一节 概述.....	1
第二节 诊断试验研究基本方法.....	2
一、诊断试验研究设计.....	2
二、诊断试验研究实施.....	3
三、诊断试验研究报告、质量评价与分级.....	6
第三节 诊断试验的评价指标及临床应用.....	9
第二章 诊断试验系统评价/ Meta 分析概述	14
第一节 系统评价/ Meta 分析简介.....	14
一、系统评价/ Meta 分析起源.....	14
二、系统评价/ Meta 分析定义.....	15
三、系统评价/ Meta 分析研究现状与挑战.....	17
第二节 诊断试验系统评价/ Meta 分析简介.....	24
一、诊断试验系统评价/ Meta 分析定义.....	24
二、诊断试验系统评价/ Meta 分析制作步骤.....	25
三、诊断试验系统评价/ Meta 分析研究现状.....	25
第三节 Cochrane 协作网与 Cochrane 诊断试验系统评价.....	29
一、Cochrane 协作网简介.....	29
二、Cochrane 诊断试验系统评价.....	30
第三章 诊断试验系统评价/ Meta 分析实施	36
第一节 选题.....	36
一、选题原则.....	36
二、题目构成.....	37
三、选题注意事项.....	38
第二节 诊断试验信息检索.....	39
一、信息检索基础.....	39

二、诊断试验信息资源检索	47
三、诊断试验系统评价/Meta 分析检索现状及存在的问题	83
第三节 背景与目的	88
一、背景撰写及注意事项	88
二、目的撰写及注意事项	89
第四节 纳入排除标准的确定	89
第五节 文献筛选及质量评价	91
一、文献筛选	91
二、质量评价	97
第六节 资料提取	100
一、资料提取主要内容	101
二、数据转换	101
第七节 常用统计软件及数据处理	102
一、常用统计指标	102
二、资料分析	104
三、常用统计软件	106
第八节 结果的呈现	145
一、文献检索结果	145
二、纳入研究基本特征	146
三、纳入研究质量评价	146
四、纳入研究结果及 Meta 分析结果	146
五、其他	148
第九节 讨论与结论撰写	148
一、讨论	148
二、结论	150
第十节 系统评价 / Meta 分析质量评价	151
一、AMSTAR 量表	151
二、PRISMA 声明	152
第四章 GRADE 在诊断试验系统评价 / Meta 分析中的应用	159
第一节 GRADE 概述	159
第二节 GRADE 分级原理	160
一、偏倚风险	162
二、不直接性	162
三、不一致性	163

四、不精确性·····	163
五、发表偏倚·····	163
六、升高证据质量的因素·····	164
七、挑战与未来的发展·····	164
第三节 GRADEpro 在证据分级中的运用·····	164
一、GRADEpro 简介·····	164
二、下载与安装·····	164
三、操作界面简介·····	165
四、数据输入·····	166
五、证据质量评估·····	171
六、结果呈现·····	174
七、结论解读·····	176
第四节 GDT 网站在证据分级中的运用·····	176
一、简介·····	176
二、操作流程·····	177
三、与 GARDEpro 的异同点·····	179
第五章 诊断试验系统评价/ Meta 分析案例剖析·····	180
第一节 关注无症状乳腺癌人群筛查·····	180
第二节 关注乳腺肿块诊断·····	183
第三节 关注乳腺癌诊断·····	186
第四节 预测乳腺癌新辅助化疗病理完全缓解·····	189
第五节 评估乳腺癌术前新辅助化疗后病理学完全缓解情况·····	191
第六节 检测乳腺癌新辅助化疗后残存病变·····	194
第七节 关注乳腺癌术前评估·····	197
第八节 预测早期乳腺癌无复发生存率·····	199
第九节 评估术中乳腺癌前哨淋巴结·····	201
第十节 评估乳腺癌腋窝淋巴结·····	204
第十一节 关注乳腺病变·····	207
第十二节 关注乳腺癌复发和转移·····	209
附录·····	214
附录一 Cochrane 中心一览表·····	214
附录二 Cochrane 系统评价工作组一览表·····	217
附录三 Cochrane 协作网研究领域/网络一览表·····	220
附录四 Cochrane 协作网方法学组一览表·····	221

附录五	中英文名词对照索引·····	222
附录六	诊断试验系统评价/Meta 分析主要术语·····	224
附录七	循证医学相关教材·····	227
附录八	系统评价/Meta 分析制作相关著作·····	229
附录九	系统评价/Meta 分析制作相关参考资源·····	231
附录十	SCI 收录诊断试验系统评价/Meta 分析的主要期刊列表·····	232

第一章 诊断试验简介

疾病的正确诊断是临床治疗的基础，只有正确诊断，才能保证有效干预。一般患者到医院就医，医生首先要判断患者是否患病、患何种疾病、疾病发展的程度等，在肯定和排除诊断中，需要根据诊断试验的准确性、可靠性、安全性、费用及患者的可接受性等方面科学地加以选择，并合理解释诊断结果，达到提高疾病的诊断水平和治疗效果之目的。

第一节 概 述

1. 诊断试验定义

诊断试验 (diagnosis test) 指临床上用于疾病诊断的各种试验，涉及临床采用的各种诊断手段和方法，它可为疾病正确诊断及其鉴别诊断提供重要依据，同时也可用于判断疾病的严重程度，估计疾病的临床过程、治疗效果及其预后，筛选无症状的患者和检测药物不良反应等。包括：病史和体检所获得的临床资料；各种实验室检查（如生化、血液学、免疫学、病理学检查等）；各种影像学检查（如 X 线、B 超、CT、PET/PET-CT、MRI 及放射性核素等）；其他特殊器械检查（如心电图、内镜等）；各种公认的诊断标准（如各种自身免疫性疾病的联合诊断标准等）。

2. 金标准定义

金标准 (gold standard) 又称为标准诊断试验 (standard diagnostic test)、参考标准 (reference test) 等，指当前医学界公认的诊断疾病最可靠的诊断方法，或者是一种广泛接受或认可的具有高敏感度和特异度的诊断方法。对大多数疾病而言，活体病理组织检查、手术探查、尸体解剖等均是具有普遍意义的金标准，也可由专家制定并得到的临床诊断标准和长期临床随访所获得的肯定诊断作为金标准。金标准的选择应结合临床具体情况，如肿瘤诊断应选用病理检查，胆石症以手术发现为金标准。如果金标准选择不当，就会造成对研究对象“有病”和“无病”划分上的错误。

3. 诊断试验分期

诊断试验的设计通常是一个横断面设计，包括诊断试验的准确性和可靠性等方面。

通常讨论的是诊断试验的准确性，从一个诊断指标最初在实验室被发现，到最后作为诊断试验进入临床应用，基本需要以下三个阶段：I 期是探索阶段，通常包括少量患者，比较确诊有病和健康者中诊断试验的表现。II 期是挑战阶段，将新的诊断试验与另一个已存在的诊断试验相比，在此阶段通常会增加研究对象，在更大范围的患者中评价其表现，包括不易诊断的患者和临床上需要鉴别诊断的对照，如有合并症或其他潜在混淆的状况等。前两期通常是回顾性病例对照研究。III 期是临床阶段，其目标是尽可能准确和无偏倚地获得诊断试验准确性和相对准确性的估计。III 期试验，估计诊断试验准确性的研究人群应尽可能地接近目标人群，通常是大样本，且是前瞻性研究，以避免回顾性研究常发生的偏倚。

(杨克虎)

第二节 诊断试验研究基本方法

一、诊断试验研究设计

1. 病例对照研究设计 (图 1-1)

即选择一组肯定患有某种疾病的患者 (病例组)，一组肯定不患有某种疾病的对象 (对照组，可为患其他疾病患者或健康人)，两组均进行诊断试验，根据结果评估诊断试验的准确性。病例对照研究设计选择研究对象时已明确患者是否患有某种疾病，容易发生选择性偏倚或部分核实偏倚。

2. 队列研究设计 (图 1-2)

即连续纳入所有怀疑患有某种疾病的患者，同步进行诊断试验和金标准检查，再盲法评估两者的结果。该研究设计保证纳入研究对象与临床实践具体情况相似，且每名可疑患者均进行金标准检查，能较好避免选择性偏倚或部分核实偏倚。

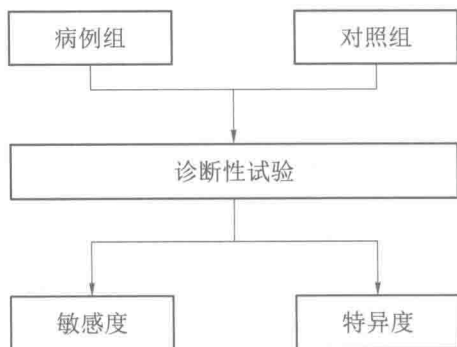


图 1-1 病例对照研究设计示意图

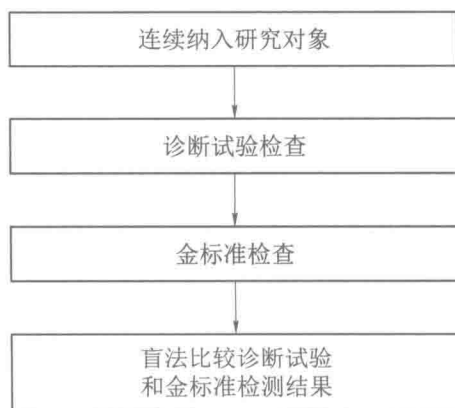


图 1-2 队列研究设计示意图

二、诊断试验研究实施

1. 确立金标准

金标准是指疾病公认的诊断标准,如果待评价的诊断试验未与“金标准”进行对比,就无法证明疾病诊断的准确性,但若所谓的“金标准”本身不准确,就存在参考试验偏倚,造成对研究对象“患病组”、“非患病组”划分上的错误,从而影响诊断试验评价的准确性。在实际工作中应根据临床具体情况选择合适的标准诊断方法作为“金标准”,如常用病理学检查作为肿瘤诊断的金标准,但若应用细针肝穿刺肿瘤细胞检查或病理学活检作为金标准诊断肝细胞癌常会造成遗漏,需要结合临床资料和影像学检查等。另外,金标准是相对的,任何一个金标准只是特定时期下医学发展的产物,它有相对稳定性,但不具有永恒性,即过去是金标准,现在则不一定是。因此,正确选择金标准是提高诊断试验研究与评价质量的关键。对有些疾病,如精神心理等疾病,往往缺乏金标准,即使恶性肿瘤诊断试验评价中约有 20% 的研究亦无明确的金标准。

2. 选择研究对象

纳入诊断试验的患者需要有合适的疾病谱,以保证样本具有代表性,即研究中所检查患者的疾病谱与诊断试验在临床应用时患者的疾病谱相同。但必须清楚两个问题:① 研究者想利用这个诊断试验解决什么问题,某一临床问题的定义决定了研究时患者的选择。若研究肝癌切除术后门静脉癌栓对预后的影响,这时研究对象应包括伴或不伴显微镜下癌栓和肉眼癌栓的患者;若研究目标是根治性切除肝癌患者,伴有肉眼癌栓的患者就不符合该条件,应挑选镜下癌栓的患者。② 患者的样本是否有代表性。研究对象包括两组人,一组是经过金标准确诊的患研究疾病的患者,这些患者需要一定合适的疾病谱,应包括早、中、晚期患者,或轻、中、重型患者,理想的情况各期比例与临床相一致。另一组是非病例组,也是由金标准证实未患研究的疾病,同时与病例组的其他特征相似。理想样本是那些临床需要鉴别诊断的患者。样本越有代表性,对研究疾病的判断就越准确。当然最科学的方法是从被测人群中随机抽样,这种做法在现实中难以实现,比较可行的方法是选取某一段时期内,那些因疑似某疾病在研究者医院中就诊的连续样本。但要注意研究场所,一般地说,在三级医院就诊的患者患严重疾病的比例高于在社区医院就诊者。如果一种诊断试验的测量是在三级医院进行的,则可能有较高的敏感度和较低的特异度,因为这里的疾病更重、合并症较多。相反,如果是在初级保健单位则相同的诊断试验可能有较低的敏感度和较高的特异度。

许多诊断试验设计采用健康者与诊断明确的患者比较来评价诊断试验的准确性。这种设计在研究早期或许有提示作用,因为没有纳入与该病混淆的其他疾病,亦即没有纳入检验结果呈“灰色带”的患者,从而高估该诊断试验的准确性。这种试验的研究对象不能代表试验应用的目标人群,从而产生“疾病谱偏倚”。

3. 样本量估算

进行诊断试验研究时需要一定的样本量,其意义是估计研究中误差与降低研究中的抽样误差。样本量过小,诊断试验的准确性指标就可能不稳定,影响对诊断试验结果的评价。诊断试验样本量通常根据被评价诊断试验的敏感度和特异度分别计算研究所需的患者数和非患者数,应用总体率的样本含量计算方法。样本大小估计与显著性水平 α 值、允许误差 δ 、试验敏感度和特异度有关。 α 值越小,所需样本量越大,一般取 $\alpha=0.05$ 。 δ 越大,样本量越小,一般 δ 取0.05或0.1。

(1) 当敏感度和特异度接近50%时,样本量估计公式为:

$$n = u_{\alpha}^2 \times p \times (1 - p) / \delta^2$$

如估计被评价诊断试验敏感度和特异度分别为70%和80%,评价该诊断试验所需样本量估算如下:

设 $\alpha=0.05$, $\delta=0.05$ 。即 $U_{\alpha}=1.96$

$$n_1 = 1.96^2 \times 0.70 \times (1 - 0.70) / 0.05^2 \approx 323$$

$$n_2 = 1.96^2 \times 0.80 \times (1 - 0.80) / 0.05^2 \approx 246$$

评价该诊断试验需要患病组人数为323名,未患病组人数为246名。在诊断试验评价的后期,由于研究对象入组时并不知道患病还是未患病,但研究者需要大致了解这个目标人群的患病率,还需要进一步测算需要多少样本量才能最后满足研究中研究对象至少有323名患病者和246名未患病者。

(2) 当敏感度或特异度 $\leq 20\%$ 或 $\geq 80\%$ 时,资料呈偏态分布,需对率进行转换,其公式为:

$$n = \left[57.3u_{\alpha} / \sin^{-1}(\delta / \sqrt{p(1-p)}) \right]^2$$

4. 同步独立、盲法比较测量结果

所谓“独立”指所有研究对象都要同时进行诊断试验和金标准的测定,不能根据诊断试验的结果有选择地进行金标准测定。原则上要求所有研究对象都经过“金标准”的评价以确定是否患有研究的疾病。但有些研究者只对部分的研究对象进行了“金标准”评价,这时需要了解这部分研究对象是如何产生的,若是随机产生的,那么在结果的分析阶段可以校正。若这些研究对象不是随机产生的,其结果会产生明显的偏倚。如果研究者根据诊断试验结果来决定该患者是否进一步去做金标准试验,就会产生“确定偏倚”。如金标准昂贵或是侵入性检查,研究者仅对诊断试验阳性者才进一步用金标准加以确诊,而阴性者则不再做进一步检查就简单地认定无病,造成假阴性资料的缺乏。这种偏倚在肿瘤诊断试验中非常普遍,如应用AFP检测诊断肝癌,AFP阴性者常会被认为无癌,但实际上AFP诊断肝癌的敏感度不到60%,这样就会出现很大一部分假阴性的患者,即结果会夸大诊断试验的敏感度,使其准确性受到影响。

所谓“盲法”指诊断试验和金标准方法结果的判断或解释相互不受影响。这里涉及

两个概念，一是金标准的判断是否盲法？意为金标准结果的判定与诊断试验的结果无关。当金标准的结果判定与诊断试验结果结合在一起时，将诊断试验检测结果与从参考试验方法得到的结果相结合，作为金标准的一部分，就会产生联合偏倚。由于诊断试验和金标准部分相同，这种偏倚导致对测量准确度的过度估计。另一概念是诊断试验判断是否盲法？意为诊断试验结果的判断受金标准结果的影响，研究者更愿意看到诊断试验是有效的，这样会高估诊断试验的准确性，这在量表性诊断工具中尤其要注意。

5. 诊断试验的可靠性分析

对诊断试验证据评价时，还要注意诊断试验的可靠性，又称重复性，是指诊断试验在完全相同条件下，进行重复试验获得相同结果的稳定程度。在实验条件相同情况下，研究中的观察测定仍可能存在测量变异。常见的原因有：① 观察者之间的变异，水平相似的 2 位放射科医师，在阅读相同的胸片时，诊断出肺门淋巴结的阳性率不一致；② 观察者自身的变异，同一位医师间隔 1 月后，重复阅读相同的胸片时结果也有差异；③ 测量仪器变异，血压计本身变异也会造成血压值的变化；④ 试剂变异，不同批次试剂存在着差异；⑤ 被测研究对象生物学变异，个体内和个体间会有不同等。诊断试验可靠性评价的设计与真实性评价不一样，评价指标主要用来评价测量变异的大小。

(1) 计量资料 用标准差及变异系数(CV)来表示。变异系数=标准差/均数×100%。变异系数和标准差越小，可靠性越好。

(2) 计数资料 用观察符合率与卡帕值(Kappa value)表示。观察符合率又称观察一致率，指 2 名观察者对同一事物的观察或同一观察者对同一事物 2 次观察结果一致的百分率。前者称观察者间观察符合率，后者称观察者内观察符合率。

如 2 名放射科医生分别对 100 名肺门淋巴结的胸片进行读片，结果见表 1-1。

表 1-1 两名放射科医生对 100 名肺门淋巴结胸片的读片结果

甲医生	乙医生		合计
	无肺门淋巴结	有肺门淋巴结	
无肺门淋巴结	60 (A)	10 (B)	70 (R ₁)
有肺门淋巴结	12 (C)	18 (D)	30 (R ₂)
合计	72 (C ₁)	28 (C ₂)	100 (N)

$$\text{观察的符合率} = (A+D)/N \times 100\% = (60+18)/100 \times 100\% = 78\%$$

Kappa 值是判断不同观察者间校正机遇一致率后观察一致率指标。其含义是实际符合率与最大可能符合率之比。计算过程如下：

$$\text{观察符合率}(P_0) = \frac{A+D}{N} \times 100\%$$

$$\text{机遇符合率}(P_c) = \frac{R_1 C_1 / N + R_2 C_2 / N}{N} \times 100\%$$

实际符合率=观察符合率-机遇符合率= $P_0 - P_c$

最大可能符合率=1-机遇符合率= $1 - P_c$

$$\text{Kappa} = \frac{\text{实际符合率}}{\text{最大可能符合率}} = \frac{P_0 - P_c}{1 - P_c}$$

根据这些公式，计算上例中的 Kappa 值：

$$\text{Kappa} = \frac{P_0 - P_c}{1 - P_c} = \frac{0.78 - 0.588}{1 - 0.588} = 0.466$$

Kappa 值充分考虑了机遇因素对结果一致性的影响，Kappa 值范围介于-1 到 1。Fleiss 提出三级划分：0.75~1.00 符合很好，0.40~0.74 符合一般，0.01~0.39 缺乏符合。

三、诊断试验研究报告、质量评价与分级

1. 诊断试验研究报告

Lijmer JC 等人在研究中发现许多研究设计缺陷都会导致过高估计诊断准确性研究结果。设计差的研究可能获得对诊断试验准确性过度夸大的结果，导致诊断试验在尚不成熟的情况下过早地用于临床，可能误导医生在个体患者的治疗中做出错误的决策。同时，一项设计实施良好的诊断研究，不一定有高质量的报告。很多杂志上发表的诊断研究由于未能提供足够的信息，用以判断该诊断试验的敏感度和特异度，因而无法对诊断试验的质量做出评价。并且具有方法学缺陷的研究往往会人为地夸大试验结果。为了规范和提高诊断性试验研究报告的质量，诊断准确性研究报告标准 (Standards for Reporting Diagnostic Accuracy, STARD) 指导委员会提取可能有用的条目形成一份清单。再由研究人员、编辑及相关专业组织成员在参加的共识会议中对这份清单进行讨论，最终制定出一份诊断准确性研究报告清单 (表 1-2) 和通用流程图 (图 1-3)。

表 1-2 STARD 诊断准确性研究报告清单

内容及主题	条目	描述
标题、摘要与关键词	1	能够判断是一篇诊断准确性研究 (建议采用医学主题词表中的敏感度和特异度)
前言	2	陈述研究问题或研究目的，如评估诊断试验的准确性或比较不同诊断试验的准确性，不同研究对象群体之间的准确性
方法		
研究对象	3	描述研究对象的纳入与排除标准，数据收集的机构和研究场所
	4	描述研究对象募集是基于存在某症状、各种检查结果，还是基于研究对象已经接受的被评价诊断试验或金标准
	5	描述研究对象的抽样是否根据上述条目 3 和条目 4 中纳入标准连续纳入研究对象，如果不是，需详细描述研究对象选择依据

续表

内容及主题	条目	描 述
诊断方法	6	描述数据收集的设计是在被评价诊断试验和金标准前（前瞻性研究）还是之后（回顾性研究）
	7	描述金标准及其使用的合理性
	8	描述被评价诊断试验和金标准的材料和方法的技术要点，包括何时、何种方法进行各种测量，以及被评价诊断试验和/或金标准的参考文献
	9	描述被评价诊断试验和金标准的定义、原理，所使用单位，以及采用的界值、结果分类方法
	10	描述实施和读取被评价诊断试验和金标准结果人员数量、是否经过培训及其技术专长
统计学方法	11	描述被评价诊断试验和金标准评判结果的人员是否设盲（即盲法实施），同时描述结果评价者可能获得的其他任何相关临床信息
	12	描述计算或比较被评价诊断试验准确性的各项指标的计算方法，描述结果的精确性（如 95%可信区间）
	13	如果进行了可重复性研究，描述可重复性计算的方法
结果		
研究对象	14	描述研究实施的时间，包括研究对象募集的起止时间
	15	报告研究对象的人口学和临床特征（如年龄、性别、症状呈现的情况、有无并发症、当前治疗、研究对象入组的场所）
	16	报告满足纳入标准的研究对象人数、接受和未接受被评价诊断试验或金标准的人数、描述研究对象未能接受被评价诊断试验或金标准的原因（建议使用流程图）
试验结果	17	报告研究对象接受被评价诊断试验和金标准之间的时间间隔，以及在此期间接受的任何干预措施
	18	报告具有目标状态的研究对象中疾病严重程度（给出明确定义）的分布；没有目标状态的研究对象报告其他疾病分布
	19	按照金标准分类分别报告被评价诊断试验结果（包括不明确的和缺失的结果），列出四格表，对于连续性结果变量，按照金标准分类分别报告连续变量分布
	20	报告被评价诊断试验和金标准中出现的任何不良事件
效应估计	21	报告被评价诊断试验准确性的效应值以及统计学的不确定性的指标（如 95%可信区间）
	22	报告被评价诊断试验无法解释结果、不确定性结果和中间结果的处理方法
	23	报告被评价诊断试验准确性和有效性的不同亚组、不同读取结果者和不同分中心间不同
	24	如果可能，报告诊断试验可重复性的估计
讨论		
	25	对研究结果的临床适用性进行讨论

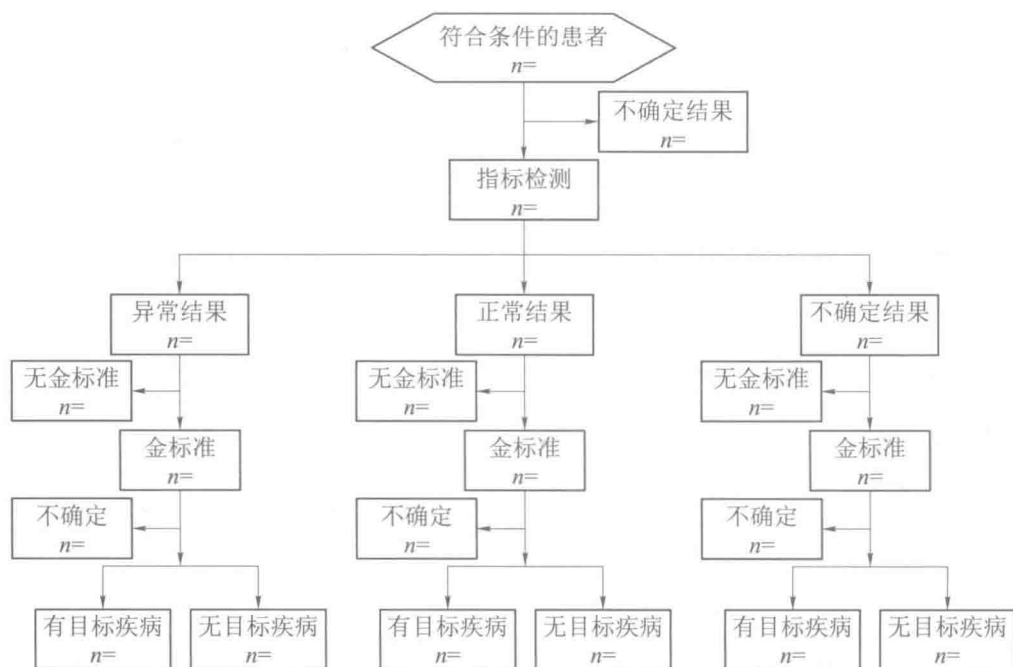


图 1-3 诊断准确性研究典型流程图

2. 诊断试验研究质量评价

详见第三章第五节。

3. 诊断试验研究证据分级

对诊断试验研究证据分级能够使证据使用者了解证据的适用范围，从而正确地使用证据。1998 年，由临床流行病学和循证医学专家 Bob Phillips、Chris Ball 和 David Sackett 等人共同循证制定了新的分级标准，2001 年 5 月正式发表于英国牛津循证医学中心的网络上，涉及治疗、预防、病因、危害、预后、诊断、经济学分析等 7 个方面，诊断方面见表 1-3。

表 1-3 牛津证据分级与推荐意见强度诊断部分

推荐级别	证据水平	条 目
I 级	I a	同质性好的诊断试验的系统评价或有试验基础可靠的临床指南
	I b	全部患者均同步做金标准和诊断试验检查且作独立的盲法比较
	I c	绝对的特异度高即阳性患者则可确诊；绝对的敏感度高即阴性患者则可排除
II 级	II a	同质性好但水平低于 I 级的诊断试验系统评价
	II b	1. 均同步作了金标准及诊断试验，也进行了独立盲法比较但研究对象局限且不连贯 2. 验证尚未确认的临床指南
III 级		研究对象并未全部作金标准检查，但作了适当指标的独立盲法比较
IV 级		没有独立利用金标准或未作盲法比较
V 级		专家意见