Reporting and Concordance of Methodologic Criteria Between Abstracts and Articles in Diagnostic Test Studies

Carlos A. Estrada, MD, MS, Richard M. Bloch, PhD, Diana Antonacci, MD, L. Lorraine Basnight, MD, Sangnya R. Patel, MD, Sanjay C. Patel, MD, Wilhelmine Wiese, MD

OBJECTIVE: **To evaluate the quality and concordance of methodologic criteria in abstracts versus articles regarding the diagnosis of trichomoniasis.**

STUDY DESIGN: **Survey of published literature.**

DATA SOURCES: **Studies indexed in MEDLINE (1976–1998).**

STUDY SELECTION: **Studies that used culture as the gold or reference standard.**

DATA EXTRACTION: **Data from abstract and articles were independently abstracted using 4 methodologic criteria: (1) prospective evaluation of consecutive patients; (2) test results did not influence the decision to do gold standard; (3) independent and blind comparison with gold standard; and (4) broad spectrum of patients used. The total number of criteria met for each report was calculated to create a quality score (0–4).**

MEASUREMENTS AND MAIN RESULTS: **None of the 33 abstracts or full articles reported all 4 criteria. Three criteria were reported in none of the abstracts and in 18% of articles (95% confidence interval [95% CI] 8.6% to 34%). Two criteria were reported in 18% of abstracts (95% CI, 8.6% to 34%) and 42% of articles (95% CI, 27% to 59%). One criterion was reported in 42% of abstracts (95% CI, 27% to 59%) and 27% of articles (95% CI, 15% to 44%). No criteria were reported in 13 (39%) of 33 abstracts (95% CI, 25% to 56%) and 4 (12%) of 33 articles (95% CI, 4.8% to 27%). The agreement of the criteria between the abstract and the article was poor (κ −0.09; 95% CI, −0.18 to 0) to moderate (κ 0.53; 95% CI, 0.22 to 0.83).**

CONCLUSIONS: **Information on methods basic to study validity is often absent from both abstract and paper. The concordance of such criteria between the abstract and article needs to improve.**

KEY WORDS: **evidence-based medicine; periodicals; publishing; quality control; sensitivity and specificity; diagnosis.**
J GEN INTERN MED 2000;15:183–187.

When faced with clinical questions regarding which is the best treatment for a patient or which diagnostic test is worth performing, physicians are being encouraged to look for evidence rather than rely solely on clinical experience or "expert" opinion. Although some clinical questions have no supporting data, it is felt that the quality of care would be improved if physicians used the "best available evidence" in making decisions.[1,2] One of the major barriers is the time required to find, read, evaluate, and understand the evidence.[3]

The abstract is an important element in the communication from scientists to practitioners. High-quality abstracts reliably and succinctly reflect an article's methodology and content. The abstract provides enough information to allow readers to accurately assess if the article is relevant to their interests, has findings that could affect their practice, and uses methods that meet basic validity standards. The role of the abstract is as valuable to clinicians when it correctly directs them not to read an article as when it encourages reading a relevant article. Structured abstracts require authors to be explicit as to the design and findings described in the article and are used in journals to facilitate the communication of research findings.[4–6] In comparison with unstructured abstracts, some authors have shown that structured abstracts improve the amount and quality of the information provided.[7] In the case of clinical trials, however, no improvement in the quality of reporting was found when structured abstracts were used.[8] Reporting of methodologic characteristics in abstracts and full articles of clinical trials, review articles, and clinical practice guidelines could improve communication among practitioners.[4,9–12]

In the case of diagnostic test studies, fundamental methodologic deficiencies have been reported.[13–15] In view of the increasing amount of information in MEDLINE, clinicians use the title and the abstract to decide whether the full article is worth reading. One would hope that the abstract provides enough accurate information to facilitate appropriate decisions.[4] Our objective was to evaluate the quality and concordance of methodologic criteria in abstracts versus full articles on diagnostic test studies for trichomoniasis. We chose this specific example for several reasons. First, the materials had been collected in the process of preparing a meta-analysis of the wet mount and Pap smear for diagnosis of trichomoniasis.[16] Second, a gold standard diagnostic procedure existed in this area, which allowed for the comparison of other diagnostic procedures. Finally, the gold standard had existed for long enough that many diagnostic articles existed on the same topic.

Received from East Carolina University School of Medicine, Greenville, NC (CAE, RMB, DA, LLB, SRP, SCP, WW).

Presented in part at the regional meeting, Southern Society of General Internal Medicine, New Orleans, La, February 18–20, 1999, and at the 22nd national meeting, Society of General Internal Medicine, San Francisco, Calif, April 20–May 1, 1999.

Address correspondence and reprint requests to Dr. Estrada: East Carolina University School of Medicine, Pitt County Memorial Hospital, 600 Moye Blvd., Teaching Annex 389, Greenville, NC 27858 (e-mail: cestrada@brody.med.ecu.edu).

METHODS

We searched the MEDLINE database for articles published between April 1976 and February 1998 about diagnostic tests of vaginal trichomoniasis using Ovid 7.05 (Ovid Technologies Inc., New York, NY). The key words to identify diagnostic tests and trichomoniasis were then combined (Table 1). The search was limited to studies published in English and performed on humans. The search yielded a total of 374 articles, but we included only articles in which the test was compared with a gold standard (as reported in the full paper).[16] We required that the gold standard be the identification of the trichomonad by culture.[17,18] Each full paper was randomly assigned to 2 of 4 of the authors (CE, SCP, SRP, and WW) and independently reviewed for inclusion in the study ($\kappa = 0.87$; 95% confidence interval [95% CI] 0.81, 0.93). Disagreement was resolved by consensus among the 4 authors. Thus, 33 articles were included in this study and 341 articles were excluded: 37 were studies without a gold standard, 18 did not have an abstract, and 286 did not describe diagnostic tests. The 33 studies were published in 15 journals.

To assess the validity of the studies and minimize bias, we used primarily the criteria recommended by the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests.[19,20] Thus, we used the following methodologic criteria: (a) consecutive patients were recruited prospectively (to avoid selection and verification bias); (b) the test result did not influence the decision to perform the gold standard (to avoid verification bias); (c) the test and gold standard results were examined independently and without knowledge of the other result (to avoid measurement bias); and (d) the patient sample included an appropriate spectrum of patients in whom the diagnostic test would be used (to provide accurate estimates of sensitivity and specificity and to avoid spectrum bias). Other authors have proposed similar criteria.[14,15,21–24] Criteria were applied separately for abstracts and articles by different raters. Abstracts were randomly assigned and reviewed independently by 2 of 4 raters (CE, DA, LB, RB), and disagreement was resolved by consensus. Similarly, the articles were randomly assigned to 2 of 4 raters (CE, SCP, SRP, WW), with disagreement resolved by consensus. Each abstract and article were given 1 point for each criterion reported. An overall score of quality was obtained by adding these points (maximum score of 4). The κ interrater agreement values for prospective evaluation, influence decision, independent/blind comparison, and spectrum criteria were 0.76, 0.04, 0.65, and 1 for the abstracts, and 0.29, 0.09, 0.77, and 0.01 for the full paper, respectively. Values of κ reflect agreement that is slight (0 to 0.19), fair (0.2 to 0.39), moderate (0.4 to 0.59), substantial (0.6 to 0.79), and almost perfect (0.8 to 1).[25] Abstracts and articles also were examined for reports of sensitivity, specificity, likelihood ratios, or the numbers required to compute these values. We used the κ statistic to assess concordance of methodologic criteria between the abstract and the full article. We computed 95% CIs for proportions in small samples.[26] Data not normally distributed were described by using the median and the interquartile range (Q1–Q3). We compared medians with the Wilcoxon paired-rank test or the Wilcoxon-Mann-Whitney test as appropriate. We used a significance level of $P \leq .05$. We used Biblio-Link II and ProCite software (Research Information Systems, Carlsbad, Calif) to handle references and SPSS 8.0 software to perform statistical analyses (SPSS Inc., Chicago, Ill).

Table 1. Key Words to Identify Studies to Diagnose Trichomoniasis

Key Words to Identify Trichomoniasis	Key Words to Identify Diagnostic Test Studies
Trichomonas*	Sensitivity and specificity*
Trichomonas infections*	Diagnostic errors*
Trichomonas vaginalis	Diagnosis
Trichomonas vaginitis	False negative reactions*
Trichomon$ (textword)†	False positive reactions
	Diagnostic tests routine
	Multiphasic screening
	Likelihood functions
	Diagnosis-differential
	ROC curve
	Sensitivity (textword)
	Specificity (textword)

**Terms were exploded (captured all subheadings).*

†The dollar symbol ($) is a wild card; it retrieves any terms that begin with the letters preceding the sign.

RESULTS

The prospective evaluation of consecutive patients was reported in 5 (15%) of 33 abstracts (95% CI, 6.7% to 31%) and 11 (33%) of the articles (95% CI, 20% to 50%), as shown in Table 2. The agreement between the abstracts and the articles was moderate ($\kappa = 0.53$). The test result did not influence the decision to perform the gold standard in 17 (52%) of the abstracts (95% CI, 35% to 68%) and 26 (79%) of the articles (95% CI, 62% to 89%); the agreement between abstract and article was fair ($\kappa = 0.32$). Independent and blind comparison with the gold standard was reported in 2 (6.1%) of the abstracts (95% CI, 1.7% to 20%) and 4 (12%) of the articles (95% CI, 4.8% to 27%); the agreement between abstract and article was poor ($\kappa = -0.09$). A broad spectrum of patients was reported in 2 (6.1%) of the abstracts (95% CI, 1.7% to 20%) and 14 (42%) of the articles (95% CI, 27% to 59%); the agreement was slight ($\kappa = 0.16$).

None of the abstracts or full articles reported all 4 basic methodologic criteria, as shown in Table 3. Three criteria were reported in none of the abstracts and in 6 (18%) of the full articles (95% CI, 8.6% to 34%). Two criteria were reported in 6 (18%) of the abstracts (95% CI, 8.6% to 34%) and 14 (42%) of the full articles (95% CI,

Table 2. Methodologic Quality Criteria Concordance Between Abstracts and Articles to Diagnose Trichomoniasis (*N* = 33)

	Criteria Fulfilled and Reported, *n* (%) (95% confidence interval (95% CI))			Agreement Between Abstract and Article*
Methodologic Criteria	**Abstract**	**Full Article**	**None**	**κ (95% CI)**
Prospective evaluation of consecutive patients	5 (15) (6.7% to 31%)	11 (33) (20% to 50%)	22 (67) (50% to 80%)	0.53 (0.22 to 0.83)
Test result did not influence decision to do gold standard	17 (52) (35% to 68%)	26 (79) (62% to 89%)	6 (18) (8.6% to 34%)	0.32 (0.05 to 0.59)
Independent and blind comparison with gold standard	2 (6.1) (1.7% to 20%)	4 (12) (4.8% to 27%)	27 (82) (66% to 91%)	−0.09 (−0.18 to 0)
Broad spectrum of patients used	2 (6.1) (1.7% to 20%)	14 (42) (27% to 59%)	19 (58) (41% to 73%)	0.16 (−0.04 to 0.37)

**Values of κ reflect agreement that is slight (0 to 0.19), fair (0.2 to 0.39), moderate (0.4 to 0.59), substantial (0.6 to 0.79), and almost perfect (0.8 to 1). A negative κ reflects disagreement.*

27% to 59%). One criterion was reported in 14 (42%) of the abstracts (95% CI, 27% to 59%) and 9 (27%) of the full articles (95% CI, 15% to 44%). No criteria were reported in 13 (39%) of the abstracts (95% CI, 25% to 56%) and 4 (12%) of the full articles (95% CI, 4.8% to 27%). The quality score of the full article (median 2; Q1–Q3, 1 to 2) was higher than that of the abstract (median 1; Q1–Q3, 0 to 1) ($P < .001$). Five (15%) of the abstracts were structured (95% CI, 6.7% to 31%). The quality score of the abstract was higher among structured abstracts (median 1; Q1–Q3, 1 to 2) as compared with nonstructured abstracts (median 1; Q1–Q3, 0 to 1) ($P = .04$). The quality score of the full paper was no different among papers containing structured abstracts (median 2; Q1–Q3, 2 to 2) or nonstructured abstracts (median 2; Q1–Q3, 1 to 2) ($P = .4$).

The abstracts included sensitivity of the test in 23 (70%) of cases (95% CI, 53% to 83%), specificity in 10 (30%) (95% CI, 17% to 47%), likelihood ratio in 1 (3%) (95% CI, 0.5% to 15%), the sample of patients in 13 (39%) (95% CI, 25% to 56%), and the advantages of the tests in terms of lower costs or better outcomes in 7 (21%) (95% CI, 11% to 38%).

DISCUSSION

We examined diagnostic test studies for trichomoniasis to evaluate the extent to which abstracts in a well-studied area provided information to help a reader make decisions about reading the article. We found that information on methods basic to study validity were often absent from both abstract and paper. Also worrisome was that, overall, relatively few studies used a gold standard. Similar deficiencies in full articles have been found in a wide variety of diagnostic test studies published in 1984,[13] 1988,[14] and in 1995.[15] We also found poor to moderate concordance of the validity criteria between the abstract and the body of the article of diagnostic test studies for trichomoniasis. Other authors have noted discrepancies between data reported in the abstract and the

Table 3. Number of Methodologic Criteria Reported (*N* = 33)

	Abstracts			
Number of Criteria*	**Structured, *N* = 5** ***n* (%)**	**Non-Structured, *N* = 28** ***n* (%)**	**Total, *N* (%)** **(95% CI)**	**Full Article, *N* (%)** **(95% CI)**
4 criteria	0 (0)	0 (0)	0 (0)	0 (0)
3 criteria	0 (0)	0 (0)	0 (0)	6 (18) (8.6% to 34%)
2 criteria	2 (40)	4 (14)	6 (18) (8.6% to 34%)	14 (42) (27% to 59%)
1 criteria	3 (60)	11 (39)	14 (42) (27% to 59%)	9 (27) (15% to 44%)
none	0 (0)	13 (46)	13 (39) (25% to 56%)	4 (12) (4.8% to 27%)

**Prospective evaluation of consecutive patients; test result did not influence decision to do gold standard, independent and blind comparison with gold standard, and broad spectrum of patients used. Maximum number of criteria = 4.*

Table 4. Methodologic Quality Criteria of Diagnostic Test Studies

Valid gold or reference standard
Prospective evaluation of consecutive patients
Test result did not influence decision to do gold standard
Independent and blind comparison with gold standard
Broad spectrum of patients
Test measured independently of other clinical information
Reporting
Subjects (inclusion criteria, demographic information, comorbidities)
Test (how to perform, reproducibility, threshold, indeterminate results)
Prevalence, sensitivity, specificity, likelihood ratios (value and confidence intervals)
Subgroup analysis

body of the article. Among 6 large-circulation medical journals, Pitkin et al. found discrepancies in 18% to 68% of abstracts.[27] Discrepancies were of 2 types: data in the abstract and the body were different, and data were given in the abstract but not in the body. Our study is the first to report the quality of the abstract of a diagnostic test study and its concordance with the article.

Our study had certain limitations. First, the methodologic criteria were not always explicitly described in the abstract or full paper, resulting in less than ideal interrater agreements. We attempted to address this by discussing the full article among 4 authors, but we acknowledge that other reviewers may reach different decisions. Second, we chose a very specific disorder that has an accepted gold standard. Our findings may, therefore, not apply to abstracts about diagnostic tests for other disorders. However, we find this unlikely as the full articles on other diagnostic tests for other disorders have similar deficiencies.[13–15] The range of journals represented by the current sample was large enough to make it unlikely that different editorial judgments would be used for this subset. Third, the small sample size did not allow firm conclusions regarding the quality of papers that used structured abstracts. Finally, we used 4 methodologic criteria proposed by various groups; other criteria may result in different conclusions.

Basic methodologic criteria that determine diagnostic study validity[14,15,19,20,23,24,28] should be adhered to and clearly reported in the abstract as well as in the body of the paper of diagnostic test studies, in the same way as methodologic criteria are required for clinical trials.[29] The constraint on the number of words in the abstract could contribute to our finding of lower quality scores for the abstracts as compared with the full article. However, busy clinicians may expect that by reading an abstract they would be able to quickly identify an article whose validity or results make reading the article worthwhile, or not. We are concerned that decisions to read the articles, much less use the information within the articles, are being made without adequate information. We submit that a decision to read a valid study based on the abstract may improve efficiency and may provide information useful to practice. An abstract with insufficient information may lead the physician to refrain from reading a relevant and potentially useful article or to waste time reading a poor and potentially misleading article. Editors, reviewers, and authors should require that validity criteria be reported in the abstract as well as the article; a checklist is provided in Table 4.[13–15,19,20,23,28] The validity criteria should also be concordant between the abstract and the article. A checklist to improve the quality and concordance of the abstracts in general has been published.[30]

REFERENCES

1. Haynes RB, McKibbon KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL. How to keep up with the medical literature, I: why try to keep up and how to get started. Ann Intern Med. 1986;105: 149–53.
2. Shin JH, Haynes RB, Johnston ME. Effect of problem-based, self-directed undergraduate education on life-long learning. CMAJ. 1993;148:969–76.
3. Haynes RB. Some problems in applying evidence in clinical practice. Ann N Y Acad Sci. 1993;703:210–25.
4. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. Ann Intern Med. 1990;113:69–76.
5. Haynes RB. More informative abstracts: current status and evaluation. J Clin Epidemiol. 1993;46:595–7.
6. Harbourt AM, Knecht LS, Humphreys BL. Structured abstracts in MEDLINE, 1989–1991. Bull Med Libr Assoc. 1995;83:190–5.
7. Taddio A, Pain T, Fassos FF, Boon H, Ilersich AL, Einarson TR. Quality of nonstructured and structured abstracts of original research articles in the *British Medical Journal*, the *Canadian Medical Association Journal* and the *Journal of the American Medical Association*. CMAJ. 1994;150:1611–5.
8. Scherer RW, Crawley B. Reporting of randomized clinical trial descriptors and use of structured abstracts. JAMA. 1998;280:269–72.
9. The Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Checklist of information for inclusion in reports of clinical trials. Ann Intern Med. 1996;124:741–3.
10. Hayward RS, Wilson MC, Tunis SR, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. Ann Intern Med. 1993;118:731–7.
11. Rochon PA, Gurwitz JH, Cheung CM, Hayes JA, Chalmers TC. Evaluating the quality of articles published in journal supplements compared with the quality of those published in the parent journal. JAMA. 1994;272:108–13.
12. Pitkin RM, Branagan MA. Can the accuracy of abstracts be improved by providing specific instructions? A randomized controlled trial. JAMA. 1998;280:267–9.
13. Sheps SB, Schechter MT. The assessment of diagnostic tests: a survey of current medical research. JAMA. 1984;252:2418–22.
14. Arroll B, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. J Gen Intern Med. 1988;3:443–7.
15. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA. 1995;274:645–51.
16. Wiese WJ, Patel SR, Patel SC, Ohl C, Estrada CA. A meta-analysis of the wet mount and Papanicolaou smear for the diagnosis of vaginal trichomoniasis. J Gen Intern Med. 1999;14(suppl 2):79. Abstract.
17. Petrin D, Delgaty K, Bhatt R, Garber G. Clinical and microbiologi-

cal aspects of *Trichomonas vaginalis.* Clin Microbiol Rev. 1998; 11:300–17.

18. Sobel JD. Vaginitis. N Engl J Med. 1997;337:1896–903.
19. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended Methods. Updated June 6, 1996. Available from URL at: http://Som.Flinders.Edu.Au/Fusa/Cochrane/.
20. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994;120:667–76.
21. Kearon C, Julian JA, Newman TEGJS, for the McMaster Diagnostic Imaging Practice Guidelines Initiative. Noninvasive diagnosis of deep venous thrombosis. Ann Intern Med. 1998;128:663–77.
22. Wells PS, Lensing AW, Davidson BL, Prins MH, Hirsh J. Accuracy of ultrasound for the diagnosis of deep venous thrombosis in asymptomatic patients after orthopedic surgery: a meta-analysis. Ann Intern Med. 1995;122:47–53.
23. Jaeschke R, Guyatt GH, Sackett DL, For The Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA. 1994;271:703–7.
24. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. New York, NY: Churchill Livingstone; 1997.
25. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2nd ed. Boston, Mass: Little, Brown and Co; 1991.
26. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med. 1986;105:429–35.
27. Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. JAMA. 1999;281:1110–1.
28. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. J Gen Intern Med. 1989;4:288–95.
29. Junker CA. Adherence to published standards of reporting: a comparison of placebo-controlled trials published in English or German. JAMA. 1998;280:247–9.
30. Winker MA. The need for concrete improvement in abstract quality. JAMA. 1999;281:1129–30.

ANNOUNCEMENT

American Board of Internal Medicine

2000 ABIM Recertification Examinations in Internal Medicine, its Subspecialties, and Added Qualifications

The Board's Recertification Program consists of an at-home, open-book Self-Evaluation Process (SEP) and a proctored Final Examination which will be adminstered twice each year in May and November.

Final Examination Administration	May 2, 2000	November 8, 2000
Deadline for Completion of SEP Component	February 1, 2000	August 1, 2000
Deadline for Submission of FE Application	March 1, 2000	September 1, 2000

For more information and application forms, please contact:

Registration Section
American Board of Internal Medicine
510 Walnut Street, Suite 1700
Philadelphia, PA 19106-3699
Telephone: (800) 441-2246 or (215) 446-3500 Fax: (215) 446-3590
E-mail: request@abim.org Web Site: http://www.abim.org