

# Toward a Checklist for Reporting of Studies of Diagnostic Accuracy of Medical Tests

DAVID E. BRUNS,<sup>1\*</sup> EDWARD J. HUTH,<sup>2</sup> ERIK MAGID,<sup>3</sup> and DONALD S. YOUNG<sup>4</sup>

**Background:** “Diagnostic accuracy” refers to the ability of medical tests to provide accurate information about diagnosis, prognosis, risk of disease, and other clinical issues. Published reports on diagnostic accuracy of medical tests frequently fail to adhere to minimal clinical epidemiological standards, and such failures lead to overly optimistic assessments of evaluated tests. Our aim was to enumerate key items for inclusion in published reports on diagnostic accuracy, with a related aim of making the reports more useful for systematic reviews.

**Methods:** We examined published reports on shortcomings of studies of diagnostic accuracy. We prepared an initial draft of a checklist to address common errors and presented it at a meeting of editors. After incorporation of comments from editors, we published a revised version in *Clinical Chemistry* in 1997 for comment from readers. One of us (E.M.) additionally circulated copies of the draft to methodologists and others interested in Evidence-Based Medicine. We updated the checklist with input from these sources.

**Results:** The updated document lists items for inclusion in the title, abstract, methods, results, and discussion sections of published papers. Depending on the nature of the study, the total number of items for a single paper is ~40. We invite comments on this document, which is freely available at *Clinical Chemistry Online*, where it can be accessed readily from the Table of Contents for the July 2000 issue at [www.clinchem.org/content/vol46/issue7/](http://www.clinchem.org/content/vol46/issue7/). Comments (eLetters) can be posted there for general reading.

**Conclusions:** The suggested revisions incorporated in this report appear useful to ensure inclusion of additional information that can allow assessment of the validity of the conclusions and the applicability of the study in other settings. The list can be useful in formulating guidelines and a checklist, which will require testing by authors and study of their effect on published studies of diagnostic accuracy.

© 2000 American Association for Clinical Chemistry

Modern medical practice makes extensive use of laboratory tests, radiological imaging, and other technologies in decision-making regarding diagnosis, prognosis, monitoring, screening, and risk assessment. Despite remarkable technical testing methods and continuous improvement in analytical accuracy and precision of tests, the evaluation of the diagnostic accuracy of new tests often has received only modest attention. Thus, available tests may be analytically accurate but not provide reliable clinical information for diagnosis, risk stratification, and other clinical matters. Recent reports (1, 2) have documented that most published studies of diagnostic accuracy of clinical tests fail to meet (or fail to document adherence to) reasonable methodological standards. Moreover, the very recent study of Lijmer et al. (3) clearly demonstrated that failure to adhere to such standards produces overly optimistic estimates of diagnostic accuracy. This empirical evidence made clear the need for better clinical evaluations of medical tests (3).

In 1997, we proposed items to include in a checklist for reporting of studies of diagnostic accuracy (4). The proposal prompted valuable comments from numerous individuals, including statisticians, methodologists, editors, and researchers. Most comments have been positive, and all have been constructive. An eminent statistician correctly noted that most studies of diagnostic tests do not follow the model of a prospective clinical trial that appeared to guide the thinking in the proposed guidelines. He emphasized the need for “good information on exactly how patients were selected for inclusion in the study, and on clarity in reporting the results”. We have tried to keep our eyes on those needs, and hope that further work will

<sup>1</sup> Department of Pathology, University of Virginia School of Medicine, Charlottesville, VA 22908.

<sup>2</sup> *Annals of Internal Medicine*, 190 N. Independence Mall West, Philadelphia, PA 19106-1572.

<sup>3</sup> Department of Clinical Biochemistry, Amager Hospital, Italiensvej 1, DK2300 Copenhagen, Denmark.

<sup>4</sup> Department of Pathology and Laboratory Medicine, University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104-4283.

\*Author for correspondence. Fax 1-904-979-7599; e-mail [dbruns@virginia.edu](mailto:dbruns@virginia.edu).

produce a document that best reflects the prevalent study designs in this field.

In the updated checklist below, we have incorporated the vast majority of suggestions that we received. The present authors, however, did not agree even among themselves on every item that we proposed to each other. The wording below thus reflects compromises and not a unanimous opinion of the authors. One concern is the term “diagnostic accuracy”, which some feel will be confused with analytical accuracy. For better or worse, however, the term is well entrenched in the literature, and proposed alternatives seemed to have more problems than the present term.

“Diagnostic” is used here in the sense of a diagnostic system (5), with awareness that multiple factors (including medical history and physical examination) other than an evaluated test usually enter into a medical diagnosis or decision. In medicine, “diagnostic” tests find uses as aids for diagnosis, prognosis, monitoring, screening, and risk assessment, as mentioned above. Studies of “analytical” accuracy logically should precede studies of diagnostic accuracy; guidelines for such studies have been published [e.g., see Ref. (6)]. We do not address here how to report studies of nonanalytical sources of variation (including biological variation), which logically should be evaluated at an earlier stage of investigation, or studies of cost-effectiveness, which may require a distinct protocol.

#### **Checklist for Publications on Studies of Diagnostic Accuracy of Tests Used in Medical Case-Finding, Diagnosis, Prognosis, Risk Stratification, and Monitoring**

**Title:** Identify the study as an evaluation of a test in screening, risk assessment, diagnosis, prognosis, or monitoring. Specify the *disease* or *condition* and the *test(s)* studied.

**Abstract:** Use a *structured abstract*.

**Search terms:** Use Medline-compatible terms for *evaluated test* or *test(s)*, the *disease* or *outcome* (or both), the *criterion (gold) standard test*, and the type of *study design*. Include the terms *diagnostic accuracy*, *sensitivity* and *specificity*, and *diagnosis*.

**Introduction:** State the *research question* and why it came up. Cite a *systematic review* of the problem or provide a summary of how a *search for prior studies* was conducted. Summarize *design of study to address unresolved issues*. Indicate considerations regarding *fallibility of criterion (gold) standard* and effect on study design (e.g., use of double-reading of histologic sections). State the *hypothesis* and specific *objectives* of study.

#### **Study Protocol and Methods:**

1. *Study design*: prospective cohort, retrospective cohort, randomized clinical trial, etc.
2. *Patient-care setting* (e.g., ambulatory, general or referral practice, inpatient, volunteers) and summary of factors, especially other tests, that channeled patients to have the test under evaluation.

3. *Criteria for (a) inclusion and (b) exclusion* of subjects, especially regarding the results of any other tests and the criteria used in the interpretation of those tests. (c) *Consent procedures* and approvals of study.
4. Planned sample size and subgroup analyses; statistical power and resource considerations.
5. Methods to avoid *spectrum bias*<sup>5</sup> (e.g., consecutive series, statistically selected random sample, stratified random sample) and to define spectrum<sup>5</sup> of disease.
6. *Methods (and references) for (a) evaluated test(s) and (b) criterion (gold) standard test(s)*. References should include studies that validate the analytical performance of the tests. When no such studies have been published, provide key information. When an outcome is used as the criterion standard, indicate duration and methods of follow-up.
7. Indicate the *masking* (blinding) of those performing (a) evaluated test(s) and (b) criterion standard test(s) to avoid reviewer bias.<sup>6</sup> When multiple tests are to be evaluated, indicate whether the performance of each was without knowledge of results of the other(s).
8. Methods to avoid *verification bias*<sup>7</sup> (usually by application of criterion standard to all subjects) or to deal with its consequences.
9. Methods (and references) for *statistical analysis* including steps to deal with (a) potential for diagnostic accuracy to be overestimated when diagnostic rules are constructed by use of statistical modeling or by examination of more than one cutoff value for continuous variables, (b) repeated or serial measures, and (c) outliers.
10. *Cutoffs* used for quantitative tests and how they were determined; subjective criteria for qualitative tests.
11. For studies of prognostic tests, indicate whether the criterion standard or the evaluated tests influenced therapy (*treatment paradox*<sup>8</sup>).
12. Design features aimed at *ensuring comparability with other studies*.
13. Indicate that these guidelines were followed.

#### **Results:**

##### **A. Study Subjects:**

1. Inclusive *dates* of accrual of subjects.
2. *Sample size* achieved.
3. Numbers of subjects who were *excluded*, reasons for exclusions, and their timing.

<sup>5</sup> *Spectrum bias*: “Spectrum” refers to spectrum of disease. Bias may be introduced when, for example, only patients with advanced (presumably readily diagnosed) disease are included.

<sup>6</sup> *Reviewer bias* occurs when the performance or evaluation of a test is influenced by knowledge of the results of other tests.

<sup>7</sup> *Verification bias* may occur when the criterion standard test is applied to only a subset of the patients or subjects.

<sup>8</sup> *Treatment paradox* occurs when a positive test result successfully identifies patients at risk who then receive an effective therapy and thus have a favorable outcome despite the positive test result.

4. Number of *indeterminate test results* and their use (if any) in further data analysis.
5. *Demographics* of subjects and their clinical characteristics to include the presence or absence of disease and *spectrum* of disease (e.g., summary of signs/symptoms and disease stage).
6. *Deviations* from study protocol (e.g., loss to follow-up) and reasons.

#### B. Study Findings:

1. Data on *reproducibility* of evaluated test (e.g., analytical CV at relevant concentrations and during an appropriate temporal interval; measures of intraobserver or interobserver variability as appropriate for the way the tests were performed in the study).
2. Appropriate *tabulation* of key results (e.g.,  $2 \times 2$  contingency table).
3. *Measures of diagnostic accuracy* of test(s) and *confidence intervals* (e.g., areas under ROC curves, sensitivity/specificity pairs, likelihood ratios). When the diagnostic marker is not naturally dichotomous, posttest probabilities should be emphasized over methods relying on arbitrary cutoffs such as sensitivity and specificity. When a diagnostic rule was derived from statistical modeling of multiple potential predictors or from examination of multiple cutoffs for continuous variables, provide an unbiased validation of diagnostic accuracy on a data set not used to develop the diagnostic rule, or use a resampling technique that corrects for overfitting.
4. *Repository* where original data may be obtained (e.g., for use in systematic reviews).

#### Discussion:

1. Internal validity of study: Assess imprecision and bias, e.g., fallibility of criterion standard, losses to follow-up, verification bias, and reviewer bias.
2. Applicability to other settings, e.g., spectrum of dis-

ease, demographics of subjects. Positive and negative predictive values for relevant populations when appropriate.

3. Adequacy of sample size for potential uses of the report.
4. Interpretation of study results in context of other reports, e.g., potential utility of test as an addition to existing tests and clinical observation or as a replacement for other tests.
5. For studies of prognostic tests, discuss treatment paradox when appropriate.
6. Conclusions and implications for further research and clinical practice.

Supported in part by AACC, *Clinical Chemistry*, and the International Federation of Clinical Chemistry and Laboratory Medicine. This article is freely available at *Clinical Chemistry Online* ([www.clinchem.org](http://www.clinchem.org)), where it can be accessed readily from the Table of Contents for the July 2000 issue at [www.clinchem.org/content/vol46/issue7/](http://www.clinchem.org/content/vol46/issue7/). Comments (eLetters) can be posted there in the form of eLetters.

#### References

1. 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.
2. Bogardus ST, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodological standards. *JAMA* 1999;281:1919-26.
3. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
4. Bruns DE. The Clinical Chemist. *Clin Chem* 1997;43:2211-2.
5. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285-93.
6. Anonymous. Information for authors. *Clin Chem* 2000;46:1-5.