# *The STARD Initiative and the Reporting of Studies of Diagnostic Accuracy*

With the current public attention on medical tests *(1)*, it is appropriate that we publish *(2)* this month, simultaneously with other journals,[1] the first result of an international project to improve the reporting of studies of the diagnostic accuracy of medical tests. The report *(2)* on Standards for Reporting Diagnostic Accuracy (STARD) reflects a multidisciplinary effort, 3 years in the making, that involved close collaboration of scores of clinicians, clinical investigators, clinical chemists, radiologists, clinical microbiologists, methodologists, and others. I review here some of the evidence that studies of diagnostic accuracy of medical tests need improving and I suggest that the STARD report *(2)* is a step in the right direction.

## What Is Diagnostic Accuracy?

"Diagnostic accuracy" refers to the ability of a test to identify a condition of interest. In studies of diagnostic accuracy, the results of one or more tests are compared with a reference ("gold") standard in a group of patients suspected of having the condition of interest. The term "accuracy" in this context thus refers to the amount of agreement between the studied test(s) and the reference standard.

Studies of diagnostic accuracy are distinct from studies of analytical characteristics of a test, such as analytical accuracy, and distinct from studies of nonanalytical factors, such as within-person biological variation. Both of these types of studies provide essential information about tests, but they are not the focus here. Similarly, studies of diagnostic accuracy must be distinguished from studies of outcomes related to testing (such as improvements in health) and from studies of cost-effectiveness of testing. Quantitative understanding of the diagnostic accuracy of a test is critical both in deciding on its potential to improve outcomes and in assessing the results of a test in an individual patient.

The term diagnostic accuracy is sometimes used in a broad sense of diagnosis (as above), but a good case can be made to use a more specific term when appropriate, e.g., prognostic accuracy for tests used in prognosis. There is need for improvement of these various types of studies of agreement between a test result and the relevant reference standard (gold standard), whether the reference standard is a current diagnosis or a state that is defined by subsequent events.

## The Need for Improvement

In 1995, Reid et al. *(3)* documented the poor quality of studies of diagnostic accuracy. They examined the adherence to seven elementary standards of clinical epidemiology and reporting of results in 112 studies of diagnostic tests published in four major medical journals. More than one-half the studies were of tests performed in clinical laboratories. (Approximately one-third were radiologic.) None of the studies showed evidence of adherence to all seven standards of study design and reporting. The studies rarely provided such key information as the reproducibility of the examined test and whether the final diagnosis was made with or without knowledge of the results of the test being studied. Another report from Feinstein's group *(4)* found similar problems in studies of genetic tests.

In 1999, Lijmer et al. *(5)* demonstrated that such methodologic flaws mattered. They showed that published studies with these flaws were associated with more optimistic estimates of diagnostic accuracy than were studies without such flaws. It thus became obvious that reports of studies must include information about all potential threats to their validity.

The inadequate reporting of studies of diagnostic accuracy has created an impasse for investigators attempting to do systematic reviews. Many published papers lack sufficient information even to calculate sensitivity-specificity pairs, which are needed to construct summary ROC curves *(6)*. At a 1996 workshop on systematic reviewing of studies of diagnostic tests, Jon Deeks, Douglas Altman, and others identified for attendees additional problems with the primary studies that systematic reviewers encountered. Among these was the difficulty of assessing the extent to which results of a published study could be legitimately applied in a different setting. During the workshop, items that an editor should request from authors of primary studies became apparent.

## Efforts to Improve Reporting of Studies in Clinical Chemistry

Concerned by the 1995 paper of Reid et al. *(3)*, I had made a list of the seven standards that they studied and used it as a checklist when working with authors of reports submitted to *Clinical Chemistry*. The first manuscript to which this checklist was applied contained six of the seven items of information on the list. The missing item was the imprecision of the test—and this was a manuscript from an investigator in the field of clinical chemistry, a specialty for which assessment of imprecision of assays is a part of one's being. The aphorism, "Mankind needs not so much to be educated as reminded", came to mind.

Other editors had become concerned with the quality of reporting of a different kind of clinical study, the randomized controlled trial (RCT) of drugs and other therapies. RCTs were (and are) highly influential, but the vulnerability to bias of a given study often could not be discerned from the published reports. In 1996, a group of medical

[1] Journals publishing the STARD document this month include *Annals of Internal Medicine*, *BMJ*, and *Radiology*. Editorials or commentaries have been written for several journals, including *JAMA* and *The Lancet*.

journals published a key guideline and checklist for reporting RCTs *(7)*. The guideline, called CONSORT (Consolidated Standards for Reporting Trials), appeared to offer a model for a similar effort for reporting studies of diagnostic tests.

In 1996, a small group of clinical chemists, with the help of Ed Huth, who was the Editor Emeritus of the *Annals of Internal Medicine* and a participant in the CONSORT effort, developed a list of items to include in a checklist for studies of diagnostic accuracy. This list was sent to clinical epidemiologists, statisticians, researchers, and editors. It was also presented for discussion at the 1997 Annual Meeting of the Council of Biology Editors (now the Council of Science Editors) and published in *Clinical Chemistry* for comment *(8)*. Comments from more than 50 individuals were incorporated in the checklist published in the Journal in 2000 *(9)* and reprinted in the Information for Authors since then. The checklist has been used in the review of most studies of diagnostic accuracy published in the Journal during 2001 and 2002.

The checklist has been useful, but it also has deficiencies. One of the most important is that, in our effort to make it succinct, it contains terminology that is foreign to many investigators. Moreover, it was designed for studies of the types of diagnostic tests (mostly quantitative assays) seen most commonly in the Journal; thus, its utility for other types of tests (e.g., radiologic) is unknown. In addition, with the limited budget of this single journal, no serious documentation of the effect of the checklist has been done. This, of course, violates a central tenet of evidence-based medicine (which prompted the development of the checklist) that interventions undergo assessment and refinement.

## The STARD Initiative

The STARD Initiative was undertaken with the aim "To make a statement with checklist for the reporting of the accuracy of a diagnostic test and to update this statement in the future when necessary". The stated objective of the checklist was to improve the quality of reporting of studies and thus to allow the reader to judge both the internal validity of the study and its applicability in other settings. A key component of the STARD document is a suggested flow diagram to follow patients enrolled in the study. Both the checklist and the flow diagram are intended to be applicable to studies of all types of diagnostic tests, including radiologic tests and elements of physical examination of the patient or items of the patient's medical history.

Authors of selected manuscripts submitted to *Clinical Chemistry* have used the new checklist during the last year. During this trial period, it has improved the reporting of studies, primarily by reminding authors to add information that often strengthened the authors' conclusions but had been omitted. The checklist has not played a role in editorial decisions about acceptability of papers for publication, nor is that its intended use. Similarly, the checklist and flow diagram cannot provide a cookbook approach to studies. Their role is to ensure clear and transparent reporting of studies, which will, as always, depend for their importance on the creativity and insights and effort of their authors.

As of this month, use of both the checklist and the flow diagram become part of the recommendations in the Information for Authors of *Clinical Chemistry*. Both are described in the STARD Statement, which is published in this issue *(2)*. I will not describe them further here. An accompanying document *(10)* provides elaboration and explanation of the items in the checklist along with examples.

An important part of the STARD plan is to evaluate its effect on the reporting of studies and to revise it in light of experience and new information and feedback. Comments and suggestions are welcomed; they can be sent directly to STARD@amc.uva.nl.

### References


1. Duenwald M. Putting cancer screening to the test. N Y Times; October 15, 2002.
2. STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 2003;49:1–6.
3. 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.
4. Bogardus ST, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodological standards. JAMA 1999;281:1919–26.
5. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of deign-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.
6. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993; 13:313–21.
7. Begg CB, Cho MK, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT Statement. JAMA 1996; 276:637–9.
8. Bruns DE. The clinical chemist. Clin Chem 1997;43:2211–2.
9. Bruns DE, Huth EJ, Magid E, Young DS. Toward a checklist for reporting of studies of diagnostic accuracy of medical tests. Clin Chem 2000;46:893–5.
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD Statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.



**David E. Bruns**

*Department of Pathology*
*Box 800214*
*University of Virginia Medical School*
*E-mail dbruns@virginia.edu*