The STARD Statement for Reporting Diagnostic Accuracy Studies: Application to the History and Physical Examination

David L. Simel, MD MHS[1], Drummond Rennie, MD[2,3], and Patrick M. M. Bossuyt, PhD[4]

[1]Durham Veterans Affairs Medical Center and Duke University, Durham, NC, USA; [2]Institute for Health Policy Studies, University of California, San Francisco, CA, USA; [3]Journal of the American Medical Association, San Francisco, CA, USA; [4]Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands.

OBJECTIVE: The Standards for Reporting of Diagnostic Accuracy (STARD) statement provided guidelines for investigators conducting diagnostic accuracy studies. We reviewed each item in the statement for its applicability to clinical examination diagnostic accuracy research, viewing each discrete aspect of the history and physical examination as a diagnostic test.

SETTING: Nonsystematic review of the STARD statement.

INTERVENTIONS: Two former STARD Group participants and 1 editor of a journal series on clinical examination research reviewed each STARD item. Suggested interpretations and comments were shared to develop consensus.

MEASUREMENTS AND MAIN RESULTS: The STARD Statement applies generally well to clinical examination diagnostic accuracy studies. Three items are the most important for clinical examination diagnostic accuracy studies, and investigators should pay particular attention to their requirements: describe carefully the patient recruitment process, describe participant sampling and address if patients were from a consecutive series, and describe whether the clinicians were masked to the reference standard tests and whether the interpretation of the reference standard test was masked to the clinical examination components or overall clinical impression. The consideration of these and the other STARD items in clinical examination diagnostic research studies would improve the quality of investigations and strengthen conclusions reached by practicing clinicians.

CONCLUSIONS: The STARD statement provides a very useful framework for diagnostic accuracy studies. The group correctly anticipated that there would be nuances applicable to studies of the clinical examination. We offer guidance that should enhance their usefulness to investigators embarking on original studies of a patient's history and physical examination.

KEY WORDS: diagnostic accuracy; sensitivity; specificity.
J Gen Intern Med 23(6):768–74
DOI: 10.1007/s11606-008-0583-3

Received June 18, 2007
Revised December 10, 2007
Accepted March 5, 2008
Published online March 18, 2008

The Standards for Reporting of Diagnostic Accuracy (STARD) statement creates an opportunity to improve the reporting and science of diagnostic test evaluation, paralleling similar efforts for treatment trials.[1–3] STARD exists to help clinicians by guiding authors to provide essential information in an organized systematic format. When clinicians think of diagnostic tests, they often consider imaging or laboratory tests, overlooking that a patient's history and physical examination represents a diagnostic test. The STARD recommendations (Table 1) apply to all sorts of diagnostic test research while acknowledging differences between various types of studies. Studies of the history and physical examination differ from accuracy studies of imaging or laboratory tests. Some of the more obvious differences include the 2 populations of interest (both patients and examining clinicians), the use of unrecorded findings in clinicians' overall assessment of disease, and the clinical need to consider combinations of findings.

We reviewed and offer interpretations of the STARD recommendations that are particularly important for clinical examination diagnostic accuracy research. Focusing on only 3 of the items (item 4: describe carefully the patient recruitment process, item 5: describe participant sampling and address if patients were from a consecutive series, and item 11: describe whether the clinicians were masked to the reference standard tests and whether the interpretation of the reference standard test was masked to the clinical examination components or overall clinical impression) allows a brief screen of clinical examination articles most likely to be of higher quality.

METHODS FOR DEVELOPING STARD ADAPTATIONS

We appraised the STARD recommendations from the perspectives of a clinician, clinical researcher, and epidemiologist. The STARD recommendations are numbered 1 through 25: Items 1–2 cover the title and introduction, items 3–13 cover the methods, items 14–24 cover the results sections, and item 25 covers the discussion. We did not attempt to replicate the STARD literature review for each of these items. Instead, we suggest interpretations for several of the items based on our experience in the field of clinical examination research. We do not comment on the items where we had substantial agreement. The first author (Simel) participated in the development of similar recommendations for reporting clinical trials[4] and is the section editor for the Rational Clinical Examination Series published in the Journal of the American Medical Association. The second author (Rennie) is the deputy editor of the Journal

Table 1. Standard for Reporting Diagnostic Accuracy (STARD) recommendations*

Section and topic	Item	
Title/abstract/keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity")
Introduction	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups
Methods		Describe
Participants	3	The study population: the inclusion and exclusion criteria, setting, and locations where the data were collected
	4	Participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selection
	6	Data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	The reference standard and its rationale
	8	Technical specifications of material and methods involved including how and when measurements were taken, and cite references for index tests and reference standards
	9	Definition of and rationale for the units, cutoffs, and categories of the results of the index tests and the reference standard
	10	The number, training, and expertise of the persons executing and reading the index tests and the reference standard
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals)
	13	Methods for calculating test reproducibility, if done
Results		Report
Participants	14	When study was done, including beginning and ending dates of recruitment
	15	Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers)
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test results	17	Time interval from the index tests to the reference standard and any treatment administered between
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition
	19	A cross-tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard
	20	Any adverse events from performing the index tests or the reference standard
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence interval)
	22	How indeterminate results, missing responses, and outliers of the index tests were handled
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers, or centers, if done
	24	Estimates of test reproducibility, if done
Discussion	25	Discuss the clinical applicability of the study findings

Reprinted with permission from Clinical Chemistry[1]

of the American Medical Association, a participant in the STARD Initiative, and coeditor of the User's Guides to the Medical Literature.[5] The third author (Bossuyt) is a clinical epidemiologist who was the first author of the STARD statement.

ADAPTATIONS FOR THE TITLE, ABSTRACT, AND INTRODUCTION

Item 1. Identify the Article as a Study of Diagnostic Accuracy (Recommend MeSH Heading "Sensitivity and Specificity")

Using the term "diagnostic accuracy" in either the title or the abstract would be helpful for studies that explicitly target the history and physical examination. Most studies of diagnostic accuracy evaluate the sensitivity and specificity, expressed as likelihood ratios. Likelihood ratios are calculated from the sensitivity and specificity and represent the likelihood that a particular test result comes from a patient with disease as opposed to without disease. In the structured abstract, we also suggest including the terms *sensitivity*, *specificity*, or *likelihood ratios*. Whereas arguments can be made for other terms, a basic problem for clinicians is finding the information after publication.[6] Including "diagnostic accuracy" and the few terms we recommend would make the data much easier to retrieve.

Item 2. State the Research Questions or Study Aims, Such as Estimating Diagnostic Accuracy or Comparing Accuracy Between Tests or Across Participant Groups

Clinical examination studies generally involve diagnostic accuracy of symptoms and signs. The goal might be to identify the utility of individual findings, explicit combinations of findings, or the overall clinical assessment for the condition of interest (the "target condition"). Typically, the audience of such research is the clinician who wants to know the fewest number of most useful findings. This differs sharply from

studies of laboratory or radiologic tests where we may have only 1 test or where we seek to identify "the" best test among several. Other goals of clinical examination research include measuring observer variability, evaluating and understanding training effects, or quantifying the prevalence of disease.

ADAPTATIONS FOR REPORTING THE METHODS

Item 3. Describe the Study Population: The Inclusion and Exclusion Criteria, Setting, and Locations where Data were Collected

The details should be explicit enough so that readers can assess whether the studied patients were similar to their own. A narrow patient population does not necessarily mean the results lack generalizability. For example, a study of the joint findings in acute monoarticular gout done in a Veterans hospital with a 90% male population might generalize to other settings, unless there is some obvious reason why the findings of a woman's gouty joint would be different from that of a man's. In contradistinction, the diagnostic accuracy of symptoms and signs of adult streptococcal pharyngitis require validation in children.

Not only should the patients be described, but the clinician observers must also be characterized. Important variables to help the reader understand the clinician population include level of training (e.g., student, resident, fellow, attending physician), gender, and age.

Item 4. Describe Participant Recruitment; was Recruitment Based on Presenting Symptoms, Results from Previous Tests, or the Fact that the Participants had Received the Index Test or the Reference Standard? (Key Item)

The "index test" refers to the symptom, sign, or bedside maneuver that is being studied. The "reference standard" (also called the gold standard) is the finding that the investigators use to decide whether the patient has the condition in question. Ideally, investigators enroll study subjects in "real-time" as the patient enters the health care system but before they receive a reference standard test. By "real-time," we mean that the patients are recruited as part of clinical care as it is happening and not retrospectively. "Real-time" enrollment limits the bias from distorted study populations, reflecting the clinical reality of those who will ultimately use the results. Unfortunately, "real-time" enrollment for uncommon conditions takes time, potentially making a study less feasible.

When "real-time" enrollment is not feasible, sometimes investigators identify patients referred for the reference standard test. For example, an emergency physician might identify patients suspected of having meningitis through laboratory requests. In such a study, the submission of spinal fluid would prompt the investigator to go to the patient's bedside (or review the chart) and collect the requisite symptoms and signs. There is an inherent bias in such studies because the group of patients who undergo a lumbar puncture excludes those who are least likely to have meningitis. Conversely, this approach has less missing data because all the enrolled patients receive the reference standard test.

Item 5. Describe Participant Sampling was the Study Population a Consecutive Series of Participants Defined by the Selection Criteria in Items 3 and 4? If not, Specify how Participants were Further Selected. (Key Item)

Studies that enroll consecutive patients in "real-time" represent higher methodologic quality than those that prospectively enroll nonconsecutive patients. However, when are patients considered "consecutive" patients?

Clinical diagnosis usually involves the suspicion of a condition based on the presence of a symptom or risk factor. For example, we might be interested in knowing whether a patient with rhinorrhea and facial tenderness is more likely to have sinusitis than a patient with just rhinorrhea and no other symptoms; clinicians would almost never consider sinusitis in a patient with no symptoms. In this example, a research study evaluates all patients (consecutive) with rhinorrhea for the presence or absence of "facial tenderness." Once we have "used up" a symptom as an inclusion criterion, we can no longer evaluate its sensitivity or specificity.[7]

Sometimes, it is not practical to evaluate all eligible patients. When all potentially eligible patients are not included in the data tables, this results in incomplete inclusion and possible verification bias.[8] Verification bias occurs when the reference standard test was not obtained on all patients who were examined and enrolled. Complete data collection on patient accrual will help describe the potential magnitude of verification bias.

Item 6. Describe Data Collection: was Data Collection Planned before the Index Test and Reference Standard were Performed (Prospective Study) or After (Retrospective Study)

When "real-time" studies are not feasible, retrospective studies (especially for less common conditions) may be the investigator's best option. The investigator should describe the data collection protocol for patients included in the database, using the same careful inclusion and exclusion criteria they would have used when conducting a prospective study. Alternatively, investigators might resort to case–control studies.[9] Case–control studies cannot be used to estimate the prevalence of disease. In addition, the investigator's lack of control over the spectrum of disease may affect the results. However, this study design may be the only way to study rare disorders where the overwhelming majority of patients will not have the target condition.

Item 7. Describe the Reference Standard and its Rationale

A composite reference standard is often necessary. For example, when studying pigmented skin lesions for malignancy, not every lesion on every patient can undergo biopsy. Thus, most investigators would resort to a combined reference standard of either (1) biopsy results when performed or (2) lack of lesion change over time. When choosing to follow patients over time, the investigators should explicitly describe whether the patients followed differed from those who received the true reference standard (e.g., a biopsy).

When there are no laboratory, pathologic, or radiographic test that confirms the target condition, sometimes the results of the clinical examination itself become the reference standard. An example of such a disorder is Parkinson's disease.10 The pragmatic reference standard for Parkinson's disease might be the results of a confirmatory examination provided by an expert (or a panel of experts) and repeated later to confirm the findings.

Item 8. Describe Technical Specifications of Material and Methods Involved Including How and When Measurements were Taken and How the History and Physical Examination were Performed (and Cite Appropriate References)

The techniques of physical examination vary even on some of the most basic maneuvers. A description of exactly what was done are just as important for the clinical examination as they are for seemingly more objective laboratory tests. When questionnaires are used to systematically record the patient's risk factors and symptoms, the investigator should provide the questionnaire form in an appendix that describes how the form was administered and scored (e.g., written and self-administered, telephone, or direct interview). The methods section should note any study-specific training used to standardize the performance of physical examination items or use of questionnaires.

A second category of "how" the index tests were applied makes the clinical examination unique among diagnostic accuracy tests. Whereas the goal of a study might be to identify the useful symptoms and signs used in isolation, clinicians rely on many unmeasured components and the correct integration of the symptoms and signs into clinical reasoning. We recommend recording the clinicians' estimate of the overall probability of disease (on either a dichotomous, ordinal, or continuous scale) after they record their examination findings. The estimate of the probability of disease is a global measure of the clinical examination (the clinical "gestalt"). This single probability estimate combines the explicitly recorded findings with the unrecorded findings and general clinical observation.

Item 9. Describe Definition of and Rationale for the Units, Cutoffs, and Categories of the Results of the Index Tests and the Reference Standard

We recommend that investigators of the clinical examination establish the units a priori. Dichotomous results are always the easiest to manage, but sometimes they are not appropriate. Some index tests, for example, deep tendon reflexes, occur on ordinal scales (e.g., 0 to 4, where 0=no reflexes and 4= hyperactive reflexes with clonus). In other situations, clinicians may be uncertain about their clinical findings. When asked to evaluate whether or not they hear a third heart sound as an indicator of heart failure, a clinician may honestly state that they are "uncertain." They should not be forced to choose between present or absent, as this could introduce bias into the outcomes.

Item 10. Describe the Number, Training, and Expertise of the Persons Executing and Reading the Index Test and the Reference Standard Participants

Sometimes, investigators are interested in differences in measures of diagnostic accuracy as a function of performance. The levels of training can be a surrogate measure of expertise (e.g., years of practice) or the highest level of training received (e.g., specialty certification versus resident trainee).

For the reference standard tests, all clinicians reading the test should be able to give consistent results regardless of training level. When the interpretation of the reference standard test requires a clinician (e.g., radiographs or biopsy results), the investigators should determine the interobserver variability in the interpretation. Determining interobserver variability requires that at least 2 independent observers assess at least a sample of the reference standard tests.

Item 11. Describe Whether the Participants were Masked (Blinded) to the Reference Standard Outcome and Whether those Applying the Reference Standard were Masked to Clinical Examination Findings (Key Item)

Item 11 (independence), together with items 4 (real-time) and 5 (consecutive), represent the 3 most important items required for the highest quality clinical examination studies. The clinical examiner must record their findings blinded to the results of the reference standard test. Ideally, the reference standard should be interpreted without knowledge of the symptoms, signs, and overall clinical assessment.

Three aspects of independence require discussion. First, if consecutive patients cannot be enrolled, the examining clinician should preferably be blinded to the sampling framework and knowledge of the planned proportion of patients with the condition of interest (e.g., splenomegaly). Sometimes, it may be appropriate to blind the examining clinician to the research question. For example, if the study quantifies differences between the blood pressures measured by the nurse versus the physician, it might be best if neither was aware of the exact study question.

Second, the history (symptoms) and physical examination (signs) "inform" each other. The exact process for how the physician received the history should be recorded (e.g., inquired about symptoms herself, reviewed nurses notes, or perhaps reviewed a patient self-assessment questionnaire). When the goal is to reflect usual clinical care, the physician performing the physical examination should also obtain or be given the history.

Third, it is acceptable to allow the reference standard to be interpreted with an understanding that the patient is enrolled in a study. For example, informing a radiologist only that a chest radiograph is from a patient in a "pneumonia study" prevents expectation bias from knowing the individual symptoms, signs, or overall assessment. However, this implies that the radiologist knows that some patients in the study will have the target condition and that some will not.

Table 2. Tabular Display Derived from the 2×2 Matrix for Each Symptom and Sign Evaluated in Diagnostic Test Studies*

Results for each symptom and sign	*N*=sample size	*A*=True positive	*B*=False positive	*C*=False negative	*D*=True negative
Symptoms					
Symptom 1	100	20	15	10	55
Symptom 2	80	15	5	8	52
Signs					
Sign 1	100	18	5	12	65
Sign 2	90	24	15	6	45
Sign 3	100	20	0	25	55

**The data are the hypothetical results that show common issues in clinical examination research such as differences in the number of patients examined for each finding or a finding that has zero outcomes in 1 of the cells.*

Item 12. Describe Methods for Calculating or Comparing Measures of Diagnostic Accuracy and the Statistical Methods used to Quantify Uncertainty

Statistical hypothesis testing is not a major focus of clinical examination studies. Hypothesis testing is almost never done on sensitivity or specificity values. We recommend that authors do not subject likelihood ratios to statistical hypothesis testing but focus on the 95% confidence interval as a measure of the precision of (or confidence in) the estimated likelihood ratio. Although some general guidelines have been offered for what constitutes a "good" likelihood ratio,[5] the guidelines have no proven validity. A "good" likelihood ratio result is one that moves the pretest probability enough to affect clinical decision making.

Investigators should address whether they planned their sample size.[11,12] For some studies of diagnostic accuracy, investigators have no a priori sample size estimates because they attempt to enroll as many patients as they possibly can (often limited by time, expense, and patient availability). This is particularly true when investigators specify in advance a plan for multivariable modeling. For these analyses, investigators often resort to a "rule of thumb" in which for each independent symptom or sign they attempt to assemble a population that has at least 10 patients with the target disorder.[13] For example, if the investigators hoped to develop a 3-variable model, they would plan to enroll consecutive patients until they had at least 30 patients with the target disorder.

Item 13. Describe Methods for Calculating Test Reproducibility

Quantifying observer variability of the index test (measures include kappa statistic, weighted kappa, or intraclass coefficients) is sometimes the sole object of the research. For such studies, characterizing the clinicians becomes even more important as the studied sample should reflect a broad population of clinicians to insure generalizability.

Commonly held beliefs that index tests with high observer variability are useless do not always apply to symptoms and signs. Disagreement on a finding does not mean that the finding is worthless for all clinicians. Because much of the clinical assessment may be made from unmeasured characteristics (e.g., the clinicians intuition that a patient "looks" sick) or unmeasured combinations of findings, various symptoms and signs may be interpreted differently or weighted differently by a clinician. Variability in the interpretation of symptoms and signs represent a great opportunity to determine the components of the clinical examination that go into making an overall assessment. When multiple examiners contribute to the results, the data can be examined to deduce how more accurate physicians incorporate the individual findings into their assessment.

ADAPTATIONS FOR REPORTING THE RESULTS

Item 16. Report the Number of Participants Satisfying the Criteria for Inclusion that Did or Did Not Undergo the Index Tests and the Reference Standards; Describe Why Participations Failed to Receive Either Test (A Flow Diagram is Strongly Recommended)

A flow sheet provides an unambiguous display of the patient recruitment, enrollment, and application of the reference standard test that alerts the reader to verification bias. Typically, studies with verification bias overestimate sensitivity and underestimate specificity. Fortunately, advance planning may prevent compromised results. In the example of rhinorrhea and facial pain, most patients with rhinorrhea will have no facial pain, and the cost of evaluating all of them may be prohibitive. An investigator could plan to evaluate all rhinorrhea patients with facial pain and a prespecified random sample of patients without facial pain. This random selection facilitates later adjustment of the sensitivity and specificity by the sampling fraction (the percentage of patients who received the reference standard test), allowing the investigator to proceed with what might have been an impractical study. Both the raw and adjusted sensitivity, specificity, and likelihood ratios should be reported when adjusting for verification bias.[14]

Table 3. Data Display when the Results are Reported on an Ordinal Scale with 3 or More Levels*

Symptom or sign	Results of the reference standard		LR (95% CI)
	Condition present (*n*)	Condition absent (*n*)	
Level 3	25	2	54 (13–218)
Level 2	10	15	2.9 (1.4–5.9)
Level 1	11	180	0.26 (0.16–0.44)
Totals	46	197	

**The data (with hypothetical values) displayed in this fashion allow calculation of stratum-specific likelihood ratios.*

Table 4. Outcome Measures for Clinical Examination Research Calculated with Data from Table 2

	Sensitivity, specificity, and likelihood ratios			
	Sensitivity (*n*=number with condition)	Specificity (*n*=number without condition)	Likelihood ratio positive (95% confidence interval)	Likelihood ratio negative (95% confidence interval)
Symptoms				
Symptom 1	0.67 (*n*=30)	0.79 (*n*=70)	3.1 (1.9–5.2)	0.42 (0.25–0.71)
Symptom 2	0.65 (*n*=23)	0.91 (*n*=57)	7.4 (3.0–18)	0.38 (0.22–0.67)
Signs				
Sign 1	0.60 (*n*=30)	0.93 (*n*=70)	8.4 (3.4–20)	0.43 (0.28–0.67)
Sign 2	0.80 (*n*=30)	0.75 (*n*=60)	3.2 (2.0–5.1)	0.27 (0.13–0.55)
Sign 3	0.44 (*n*=45)	1.0 (*n*=55)	50 (3.1–803)	0.56 (0.43–0.72)

Item 19. Report a Cross-tabulation of the Results of the Index Tests (Including Indeterminate and Missing Results) the Results of the Reference Standard; for Continuous Results, Report the Distribution of the Test Results for the Results of the Reference Standard

Providing tabular raw data allows readers better insight into the outcomes compared to studies that report only the sensitivity and specificity values. When there is only 1 examination done on each patient, show the results in the standard 2×2 table (Table 2).

The sample size column represents the number of patients evaluated for each index symptom or sign. We recommend grouping symptoms and signs separately so that the clinician will be clear about terminology. For example, facial pain for sinusitis could represent the patient's report (a symptom), or the examiner could elicit pain by palpation (a sign).

When the index test has more than 2 levels of outcomes, the sensitivity and specificity lose relevance, so the tables require modification to show multilevel test results with the serial likelihood ratios (see Table 3). An examination for abdominal bruits for renal hypertension represents such a case where the bruit could be heard in systole and diastole (level 3), systole alone (level 2), or not at all (level 1).[15]

Item 21. Report Estimates of Diagnostic Accuracy and Measures of Statistical Uncertainty (e.g., 95% Confidence Intervals)

The proper index test depends on the research questions and study design. When the target audience of the research is clinicians, likelihood ratios of symptoms, signs, and the accuracy of the overall clinical suspicion relate directly to clinical decision making. The sensitivity and specificity (a fraction) should be given, together with the number of patients (Table 4). Calculating the likelihood ratios and their confidence intervals from the raw values (LR+=sensitivity/[1−specificity] and LR−=[1−sensitivity]/specificity) prevents rounding errors.

The likelihood ratios can be rounded off using 2 significant digits for LR values less than 1.0 (e.g., LR=0.23) and for LR between 1.0 and 10 (e.g., LR=5.7). LR values greater than 10 can be rounded to the nearest integer (e.g., LR=21). Occasionally, investigators find zero outcomes in 1 of the cells of the 2×2 table (Table 2, sign 3). The tabular raw data should show the zero cells, and the sensitivity or specificity would be reported as either 0% or 100% (depending on which cell is zero) with a one-sided 95% confidence interval (Table 4, sign 3). The likelihood ratios when any cell is zero require adding the value 0.5 to each cell for that particular index test.[16]

With multiple observations on each patient, the diagnostic accuracy estimates can be confusing (Table 5). The complexity occurs because (a) the patients should not be counted more than once and (b) the examiners may have evaluated differing numbers of patients. Adding up the results of all the patients examined creates a summary measure that violates the principle of counting each patient only once. With multiple examiners, the alternatives are (1) use the first examiner only (assuming that the examiner was selected randomly) or (2) evaluate the diagnostic accuracy for all the examiners. The first alternative is easy but gives broad confidence interval and loses information from the other examiners. The second approach uses summary measures to get the overall diagnostic accuracy results. Alternatives to these approaches are to report the range or median with interquartile range across all the examiners.

Table 5. Data Display When There are Multiple Examiners*

	Examiner (*n*=number examined)	Sensitivity (*n*=number with condition)	Specificity (*n*=number without condition)	Likelihood ratio positive (95% confidence interval)	Likelihood ratio negative (95% confidence interval)
Finding 1	Examiner 1 (*n*=100)	0.67 (*n*=30)	0.79 (*n*=70)	3.1 (1.9–5.2)	0.42 (0.25–0.71)
	Examiner 2 (*n*=80)	0.65 (*n*=23)	0.91 (*n*=57)	7.4 (3.0–18)	0.38 (0.22–0.67)
Summary finding 1	Range	0.65–0.67	0.79–0.91	3.1–7.4	0.38–0.42
	Summary measures	0.66 (95% CI 0.52–0.79)	0.84 (95% CI 0.74–0.93)	4.0 (2.4–6.6)	0.40 (0.27–0.59)

**We used the data from Table 2 for the first 2 symptoms, as if they were for the same symptom but by different examiners. Note that a summary estimate can be used to show the variability in performance between examiners. With only 2 examiners, it may be preferable to report the summary measure as the range. In this hypothetical example, the data are clinically and statistically homogenous, so we show a random-effects summary measure to demonstrate how the data could be displayed. Different summary measure calculations could lead to slightly different, although not clinically important, changes in the estimates (e.g., the Maentel–Haenzel chi-square for the LR+ is 4.4 with a 95% CI of 2.7 to 7.0).*

Item 22. Report how Indeterminate Results, Missing Responses, and Outliers of the Index Tests were Handled

Physician uncertainty about the presence of findings reflects the reality of clinical medicine. For example, the physician performing the ice test[17] for myasthenia gravis may be genuinely uncertain whether the result is "positive" or "negative." Uncertain results should never be excluded, nor should they be forced into a positive or negative category. When the investigators allow the examining clinicians to record uncertain results, the presentation of results parallels that of a multilevel index test.[18]

Sometimes, the LR for the uncertain level will indicate a truly indeterminate result, when the likelihood ratio confidence interval includes 1. However, some uncertain levels may have information, when the corresponding LR is less than 1 (decreasing the likelihood of disease) or greater than 1 (increasing the likelihood of disease). Once the investigators decide about the impact of uncertain results, they may justifiably choose to combine the data and create a more parsimonious display.

Item 23. Report Estimates of Variability of Diagnostic Accuracy Between Subgroups of Participants, Readers, or Centers, If Done

The goal of most studies is to identify a few findings accurate enough for diagnosis, but investigators should determine their independence.[19] Multivariable modeling with logistic regression offers the best opportunity for identifying the independently useful findings. Often, no individual findings are so useful that they dominant the clinical evaluation. Thus, the investigator should assess the impact of combinations of findings through the results of a logistic model. Once the independently useful variables are identified, some investigators construct a nomogram that shows an individual patient's probability of disease. By comparing the clinical gestalt to the results of a logistic model, clinicians can also decide whether their overall clinical judgment is better than the result of simple combinations of findings.

SUMMARY OF INTERPRETATIONS OF STARD FOR CLINICAL EXAMINATION RESEARCH

An understanding of the reporting requirements for diagnostic accuracy studies of the history, physical examination, and the physician's overall estimate of the likelihood of disease should lead to better study designs. Additionally, a more standardized approach to reporting should facilitate the work of evidence-based practitioners who use computerized approaches to identify relevant studies of the clinical examination. Real-time study designs that enroll consecutive patients and evaluate index tests independently of the reference standard will always provide the most methodologically sound results.

Funding: *The authors received no funding for the preparation of this manuscript.*

Conflict of Interest: *The authors have no financial conflict of interest with this manuscript. Drs. Simel and Rennie are the editors of the "Rational Clinical Examination Series" published in the Journal of the American Medical Association. Drs. Rennie and Bossuyt were members of the original STARD Steering Group.*

Corresponding Author: *David L. Simel, MD MHS; Durham Veterans Affairs Medical Center, 508 Fulton Street, Durham, NC 27705, USA (e-mail: david.simel@duke.edu).*

REFERENCES

1. **Bossuyt PM, Reitsma JB, Bruns DE, et al.,** for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Clinical Chem. 2003;49:1–6.
2. **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.
3. **Rennie D.** Improving reports of studies of diagnostic tests. The STARD initiative. J Am Med Assoc. 2003;289:89–90.
4. **Begg C, Cho M, Eastwood S, et al.** Improving the quality of reporting randomized controlled trials: the CONSORT statement. J Am Med Assoc. 1996;276:637–9.
5. **Guyatt G, Rennie D.** Users' Guides to the Medical Literature. Chicago, IL: AMA; 2002.
6. **Moons KG, Harrell FE.** Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol. 2003;10:670–2.
7. **Williams JW, Simel DL, Roberts L, Samsa G.** Clinical evaluation for sinusitis: the value of a good history and physical examination. Ann Intern Med. 1992;117:705–10.
8. **Begg CB.** Biases in the assessment of diagnostic tests. Stat Med. 1987;6:411–23.
9. **Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM.** Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51:1335–41.
10. **Rao G, Fisch L, Srinivasan S, et al.** Simel DL, Rennie D, eds. Does this Patient have Parkinson's Disease? J Am Med Assoc. 2003;289:347–53.
11. **Bachmann LM, Puhan MA, Riet G, Bossuyt PM.** Sample sizes of studies on diagnostic accuracy: literature survey. Br J Med. 2006;332:1127–9.
12. **Simel DL, Samsa GP, Matchar DB.** Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol. 1991;44(8):763–70.
13. **Harrell FE, Lee KL, Mark DB.** Multivariable prognostic models: issues in developing models, evaluating and measuring and reducing errors. Stat Med. 1996;15:361–87.
14. **Simel DL, Halvorsen RA, Feussner JR.** Quantitating bedside diagnosis: clinical evaluation of ascites. J Gen Intern Med. 1988;3:423–8.
15. **Turnbull JM.** Is listening for abdominal bruits useful in the evaluation of hypertension? J Am Med Assoc. 1995;274:1299–301.
16. **Hasselblad V, Hedges LV.** Meta-analysis of screening and diagnostic tests. Psychol Bull. 1995;117:167–78.
17. **Scherer K, Bedlack RS, Simel DL**. Simel DL, Rennie D, eds. Does This Patient Have Myasthenia Gravis? J Am Med Assoc. 2005;293:1906–14.
18. **Simel DL, DeLong ER, Feussner JR, Matchar DB.** Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Mak. 1987;7:107–14.
19. **Holleman DR, Simel DL.** Quantitative assessments from the clinical examination: how should clinicians integrate the numerous results? J Gen Intern Med. 1997;12:165–71.