Abstract

In 2003, the QUADAS tool for systematic reviews of diagnostic accuracy studies was developed. Experience, anecdotal reports, and feedback suggested areas for improvement; therefore, QUADAS–2 was developed. This tool comprises 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of risk of bias, and the first 3 domains are also assessed in terms of concerns regarding applicability. Signalling questions are included to help judge risk of bias.

The QUADAS–2 tool is applied in 4 phases: summarize the review question, tailor the tool and produce review-specific guidance, construct a flow diagram for the primary study, and judge bias and applicability. This tool will allow for more transparent rating of bias and applicability of primary diagnostic accuracy studies.

Systematic reviews of diagnostic accuracy studies are often characterized by markedly heterogeneous results originating from differences in the design and conduct of included studies. Careful assessment of the quality of included studies is therefore essential. Since its publication in 2003, the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool has been widely used (1, 2). More than 200 review abstracts in the Database of Abstracts of Reviews of Effects mention this tool, and it has been cited more than 500 times. The QUADAS tool is recommended for use in systematic reviews of diagnostic accuracy by the Agency for Healthcare Research and Quality, Cochrane Collaboration (3), and the U.K. National Institute for Health and Clinical Excellence.

The original QUADAS tool includes 14 items assessing risk of bias, sources of variation (applicability), and reporting quality; each item is rated “yes,” “no,” or “unclear” (1). Our experience, reports from users, and feedback from the Cochrane Collaboration suggested the potential for improvements. Users reported problems rating certain items (particularly those on patient spectrum, uninterpretable or intermediate test results, and withdrawals), possible overlap among items (for example, partial verification bias and withdrawals), and situations in which QUADAS is difficult to use (for example, topics for which the reference standard involves follow-up). Here we describe QUADAS–2, an improved, redesigned tool that is based on both experience using the original tool and new evidence about sources of bias and variation in diagnostic accuracy studies.

Methods

Development of QUADAS–2 was based on the 4-stage approach proposed by Moher and colleagues (4): define the scope, review the evidence base, hold a face-to-face consensus meeting, and refine the tool through piloting.

Define the Scope

We established a steering group of 9 experts in the area of diagnostic research, most of whom participated in developing the
Review the Evidence Base

We conducted 4 reviews to inform the development of QUADAS-2 (6). In the first review, we investigated how quality was assessed and incorporated in 54 diagnostic accuracy reviews published between 2007 and 2009. The second review used a Web-based questionnaire to gather structured feedback from 64 systematic reviewers who had used QUADAS. The third review was an update on sources of bias and variation in diagnostic accuracy studies that included 101 studies (7). The final review examined 8 studies that evaluated QUADAS. Full details will be published separately.

Evidence from these reviews informed decisions on topics to discuss at the face-to-face consensus meeting. We summarized reported problems with the original QUADAS tool and the evidence for each original item and possible new items relating to bias and applicability. We also produced a list of candidate items for assessment of studies comparing multiple index tests.

Hold a Face-to-Face Consensus Meeting

We held a 1-day meeting to develop a first draft of QUADAS-2 on 21 September 2010 in Birmingham, United Kingdom. The 24 attendees, known as the QUADAS-2 Group, were methodological experts and reviewers working on diagnostic accuracy reviews. We presented summaries of the evidence and split into smaller groups of 4 to 6 participants to discuss tool content (test protocol, verification procedure, interpretation, analysis, patient selection or study design, and comparative test items), applicability, and conceptual decisions. On the basis of the agreed outcomes of the meeting, steering group members produced the first draft of QUADAS-2.

Pilot and Refine

We used multiple rounds of piloting to refine successively amended versions of QUADAS-2. Online questionnaires were developed to gather structured feedback for each round; feedback in other forms, such as e-mail or verbal discussion, was also accepted. Participants in the piloting process included members of the QUADAS-2 Group; workshop participants at the October 2010 Cochrane Colloquium in Keystone, Colorado; systematic reviewers attending a National Institute for Health and Clinical Excellence technical meeting; and biomedical science students in Switzerland.

Pairs of reviewers piloted QUADAS-2 in 5 reviews on various topics. Interrater reliability varied considerably, with better agreement on applicability than on risk of bias (Appendix Table). An additional pair of experienced review authors piloted the tool on a review with multiple index tests. Feedback from these reviewers showed poor interrater reliability and problems applying the domain on comparative accuracy studies.

Appendix Table.
On the basis of these problems and the limited evidence base on the risk of bias and sources of variation in such studies, we decided that we cannot currently include criteria for assessing studies that compare multiple index tests within QUADAS-2. Feedback at all other stages of the process was positive, with all participants preferring QUADAS-2 to the original tool.

Role of the Funding Source

This article was funded by the Medical Research Council, National Institute for Health Research, Cancer Research UK, and the Netherlands Organization for Scientific Research (916.10.034). The sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the manuscript for publication.

QUADAS-2

The full QUADAS-2 tool is available from the QUADAS Web site (www.quadas.org) (Supplement). This tool is designed to assess the quality of primary diagnostic accuracy studies; it is not designed to replace the data extraction process of the review and should be applied in addition to extracting primary data (for example, study design and results) for use in the review. The QUADAS tool consists of 4 key domains that discuss patient selection, index test, reference standard, and flow of patients through the study and timing of the index tests and reference standard (flow and timing) (Table 1).

Table 1.
Risk of Bias and Applicability Judgments in QUADAS-2

The tool is completed in 4 phases: report the review question, develop review-specific guidance, review the published flow diagram for the primary study or construct a flow diagram if none is reported, and judge bias and applicability. Each domain is assessed in terms of the risk of bias, and the first 3 domains are also assessed in terms of concerns about applicability. Signaling questions are included to help judge the risk of bias; these questions flag aspects of study design related to the potential for bias and aim to help reviewers judge risk of bias.

Phase 1: Review Question

Review authors first report their systematic review question in terms of patients, index tests, and reference standard and target condition. Because the accuracy of a test may depend on where it will be used in the diagnostic pathway, review authors are asked to describe patients in terms of setting, intended use of the index test, patient presentation, and previous testing (8, 9).

Phase 2: Review-Specific Tailoring

The QUADAS-2 tool must be tailored to each review by adding or omitting signaling questions and developing review-specific guidance on how to assess each signaling question and use this information to judge the risk of bias (Figure 1). The first step is to consider whether any signaling question does not apply to the review or whether the core signaling questions do not adequately cover any specific issues for the review. For example, for a review of an objective index test, it may be appropriate to omit the signaling question about blinding the test interpreter to the results of the reference standard.

FIGURE 1.

Process for tailoring QUADAS-2 to your systematic review.

Review authors should avoid complicating the tool by adding too many signaling questions. Once tool content has been agreed upon, rating guidance specific to the review should be developed. At least 2 persons should independently pilot the tool. If
agreement is good, the tool can be used to rate all included studies; if agreement is poor, further refinement may be needed.

**Phase 3: Flow Diagram**

Next, review authors should review the published flow diagram for the primary study or draw one if none is reported or the published diagram is inadequate. The flow diagram will facilitate judgments of risk of bias and should provide information about the method of recruiting participants (for example, using a consecutive series of patients with specific symptoms suspected of having the target condition or of case patients and control participants), the order of test execution, and the number of patients undergoing the index test and the reference standard. A hand-drawn diagram is sufficient, as this step does not need to be reported as part of the QUADAS-2 assessment. **Figure 2** is a flow diagram of a primary study on using B-type natriuretic peptide levels to diagnose heart failure.

**FIGURE 2.**

Sample of a study flow diagram.

The diagram is based on a diagnostic cohort study on using B-type natriuretic peptide levels to diagnose heart failure. Based on data obtained from Smith H, Pickering RM, Struthers A, Simpson I, Mant D. Biochemical diagnosis of ventricular dysfunction in elderly patients in general practice: observational study. BMJ. 2000;320:906–8.

**Phase 4: Judgments on Bias and Applicability**

**Risk of Bias**

The first part of each domain concerns bias and comprises 3 sections: information used to support the judgment of risk of bias, signaling questions, and judgment of risk of bias. By recording the information used to reach the judgment (support for judgment), we aim to make the rating transparent and facilitate discussion among review authors independently completing assessments. The additional signaling questions are included to assist judgments. They are answered as “yes,” “no,” or “unclear” and are phrased such that “yes” indicates low risk of bias.

Risk of bias is judged as “low,” “high,” or “unclear.” If the answers to all signaling questions for a domain are “yes,” then risk of bias can be judged low. If any signaling question is answered “no,” potential for bias exists. Review authors must then use the guidelines developed in phase 2 to judge risk of bias. The “unclear” category should be used only when insufficient data are reported to permit a judgment.

**Applicability**

Applicability sections are structured in a way similar to that of the bias sections but do not include signaling questions. Review authors record the information on which the judgment of applicability is made and then rate their concern that the study does not match the review question. Concerns about applicability are rated as “low,” “high,” or “unclear.” Applicability judgments should refer to phase 1, where the review question was recorded. Again, the “unclear” category should be used only when insufficient data are reported.

The following sections briefly explain the signaling questions and risk of bias or concerns about applicability questions for each domain.

**Domain 1: Patient Selection**

**Risk of Bias: Could the Selection of Patients Have Introduced Bias?**

**Signaling question 1:** Was a consecutive or random sample of patients enrolled?
Applicability: Are There Concerns That the Included Patients and Setting Do Not Match the Review Question?

A study ideally should enroll a consecutive or random sample of eligible patients with suspected disease to prevent the potential for bias. Studies that make inappropriate exclusions (for example, not including “difficult-to-diagnose” patients) may result in overestimation of diagnostic accuracy. In a review on anti–cyclic citrullinated peptide antibodies for diagnosing rheumatoid arthritis (10), we found that some studies enrolled consecutive participants with confirmed diagnoses. In these studies, testing for anti–cyclic citrullinated peptide antibody showed greater sensitivity than in studies that included patients with suspected disease but an unconfirmed diagnosis (that is, difficult-to-diagnose patients). Studies enrolling participants with known disease and a control group without the condition may similarly exaggerate diagnostic accuracy (7, 11). Excluding patients with “red flags” for the target condition who may be easier to diagnose may lead to underestimation of diagnostic accuracy.

Concerns about applicability may exist if patients included in the study differ from those targeted by the review question in terms of severity of the target condition, demographic features, presence of differential diagnosis or comorbid conditions, setting of the study, and previous testing protocols. For example, larger tumors are more easily seen than smaller ones on imaging studies, and larger myocardial infarctions lead to higher levels of cardiac enzymes than small infarctions and are easier to detect, thereby increasing estimates of sensitivity (3).

Domain 2: Index Test

Risk of Bias: Could the Conduct or Interpretation of the Index Test Have Introduced Bias?

Signaling question 1: Were the index test results interpreted without knowledge of the results of the reference standard?

This item is similar to “blinding” in intervention studies. Knowledge of the reference standard may influence interpretation of index test results (7). The potential for bias is related to the subjectivity of interpreting index test and the order of testing. If the index test is always conducted and interpreted before the reference standard, this item can be rated “yes.”

Signaling question 2: If a threshold was used, was it prespecified?

Selecting the test threshold to optimize sensitivity and/or specificity may lead to overestimation of test performance. Test performance is likely to be poorer in an independent sample of patients in whom the same threshold is used (12).

Applicability: Are There Concerns That the Index Test, Its Conduct, or Its Interpretation Differ From the Review Question?

Variations in test technology, execution, or interpretation may affect estimates of the diagnostic accuracy of a test. If index test methods vary from those specified in the review question, concerns about applicability may exist. For example, a higher ultrasonography transducer frequency has been shown to improve sensitivity for the evaluation of patients with abdominal trauma (13).

Domain 3: Reference Standard

Risk of Bias: Could the Reference Standard, Its Conduct, or Its Interpretation Have Introduced Bias?

Signaling question 1: Is the reference standard likely to correctly classify the target condition?

Estimates of test accuracy are based on the assumptions that the reference standard is 100% sensitive and that specific disagreements between the reference standard and index test result from incorrect classification by the index test (14, 15).

Signaling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?
Applicability: Are There Concerns That the Target Condition as Defined by the Reference Standard Does Not Match the Question?

The reference standard may be free of bias, but the target condition that it defines may differ from the target condition specified in the review question. For example, when defining urinary tract infection, the reference standard is generally based on specimen culture; however, the threshold above which a result is considered positive may vary (16).

Domain 4: Flow and Timing

Risk of Bias: Could the Patient Flow Have Introduced Bias?

Signaling question 1: Was there an appropriate interval between the index test and reference standard?

Results of the index test and reference standard are ideally collected on the same patients at the same time. If a delay occurs or if treatment begins between the index test and the reference standard, recovery or deterioration of the condition may cause misclassification. The interval leading to a high risk of bias varies among conditions. A delay of a few days may not be problematic for patients with chronic conditions, but it could be problematic for patients with acute infectious diseases.

Conversely, a reference standard that involves follow-up may require a minimum follow-up period to assess whether the target condition is present. For example, to evaluate magnetic resonance imaging for early diagnosis of multiple sclerosis, a minimum follow-up period of approximately 10 years is required to be confident that all patients who will fulfill the diagnostic criteria for multiple sclerosis will have done so (17).

Signaling question 2: Did all patients receive the same reference standard?

Verification bias occurs when only a proportion of the study group receives confirmation of the diagnosis by the reference standard, or if some patients receive a different reference standard. If the results of the index test influence the decision on whether to perform the reference standard or which reference standard is used, estimated diagnostic accuracy may be biased (11, 18). For example, in a study evaluating the accuracy of d-dimer testing to diagnose pulmonary embolism, ventilation–perfusion scans (reference standard 1) were performed in participants with positive test results for this condition, and clinical follow-up was used to determine whether those with negative test results had pulmonary embolism (reference standard 2).

This method may result in misclassifying some false-negative results as true-negative because clinical follow-up may miss some patients who had pulmonary embolism but negative results on the index test. These patients would be classified as not having pulmonary embolism, and this misclassification would overestimate sensitivity and specificity.

Signaling question 3: Were all patients included in the analysis?

All participants recruited into the study should be included in the analysis (19). A potential for bias exists if the number of patients enrolled differs from the number of patients included in the 2×2 table of results, because patients lost to follow-up differ systematically from those who remain.

Incorporating QUADAS-2 Assessments in Diagnostic Accuracy Reviews

We emphasize that QUADAS-2 should not be used to generate a summary “quality score” because of the well-known problems associated with such scores (20, 21). If a study is judged as “low” on all domains relating to bias or applicability, then it is appropriate to have an overall judgment of “low risk of bias” or “low concern regarding applicability” for that study. If a study is judged “high” or “unclear” in 1 or more domains, then it may be judged “at risk of bias” or as having “concerns regarding applicability.”
At minimum, reviews should summarize the results of the QUADAS-2 assessment for all included studies. This could include summarizing the number of studies that had a low, a high, or an unclear risk of bias or concerns about applicability for each domain. Reviewers may choose to highlight particular signaling questions on which studies consistently rate poorly or well. Tabular (Table 2) and graphic (Figure 3) displays help to summarize QUADAS-2 assessments.

Table 2.
Suggested Tabular Presentation for QUADAS-2 Results

FIGURE 3.
Suggested graphical display for QUADAS-2 results.

Review authors may choose to restrict the primary analysis to include only studies at low risk of bias or with low concern about applicability for either all or specified domains. Restricting inclusion to the review on the basis of similar criteria may be appropriate, but it is often preferable to review all relevant evidence and then investigate possible reasons for heterogeneity (17, 22).

Subgroup or sensitivity analysis can be conducted by investigating how estimates of accuracy of the index test vary among studies rated high, low, or unclear on all or selected domains. Domains or signaling questions can be included as items in metaregression analyses to investigate the association of these questions with estimated accuracy.

The QUADAS Web site (www.quadas.org) contains the QUADAS-2 tool; information on training; a bank of additional signaling questions; more detailed guidance for each domain; examples of completed QUADAS-2 assessments; and downloadable resources, including an Access database for data extraction, an Excel spreadsheet to produce graphic displays of results, and templates for Word tables to summarize results.

Discussion

Careful assessment of the quality of included studies is essential for systematic reviews of diagnostic accuracy studies. We used a rigorous, evidence-based process to develop QUADAS-2 from the widely used QUADAS tool. The QUADAS-2 tool offers additional and improved features, including distinguishing between bias and applicability, identifying 4 key domains supported by signaling questions to aid judgment on risk of bias, rating risk of bias and concerns about applicability as “high” and “low,” and handling studies in which the reference standard consists of follow-up.

We believe that QUADAS-2 is a considerable improvement over the original tool. It would be desirable to extend QUADAS-2 to permit assessment of studies comparing multiple index tests, but we concluded that the evidence base for such criteria is currently insufficient and plan future work on this topic. We hope that QUADAS-2 will help to develop a robust evidence base for diagnostic tests and procedures, and invite further comment and feedback via the QUADAS Web site.

References

   CrossRef


Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality...
Appendix: The QUADAS–2 Group

University of Oxford: Doug Altman, Susan Mallett*; Memorial Sloan–Kettering Cancer Center: Colin Begg; University of Bristol: Rebecca Beynon, Jonathan A.C. Sterne*, Penny F. Whiting*; University of Amsterdam: Patrick M.M. Bossuyt*, Mariska M.G. Leeflang*, Jeroen Lijmer; American College of Physicians: John Cornell; University of Birmingham: Clare Davenport, Jonathan J. Deeks*, Khalid Khan; Bond University: Paul Glasziou; University of Sydney: Rita Horvath, Les Irwig, Petra Macaskill; University of Exeter: Chris Hyde; Maastricht University: Jos Kleijnen; University Medical Center Utrecht: Karel G.M. Moons, Johannes B. Reitsma*; Baseline Institute of Clinical Epidemiology and Biostatistics: Heike Raatz; University of Bern: Anne W.S. Rutjes*; National Institute for Health and Clinical Excellence: Beth Shaw, Toni Tan; Keele University: Danielle van der Windt; University of Florence: Gianni Virgili; Kleijnen Systematic Reviews: Marie E. Westwood*.

* = steering group members.

For members of the QUADAS–2 Group, see the 9.
RELATED TOPICS

Research and Reporting Methods

PUBMED ARTICLES

The quality of skilled birth attendants in Nepal: High aspirations and ground realities.

Impact of p53 status on TRAIL-mediated apoptotic and non-apoptotic signaling in cancer cells.

View More

Results provided by: PubMed
In 2003, the QUADAS tool for systematic reviews of diagnostic accuracy studies was developed. Experience, anecdotal reports, and feedback suggested areas for improvement; therefore, QUADAS-2 was developed. This tool comprises 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of risk of bias, and the first 3 domains are also assessed in terms of concerns regarding applicability. Signalling questions are included to help judge risk of bias. The QUADAS-2 tool is applied in 4 phases: summarize the review question, tailor the tool an Notes on use of Methodology checklist: the QUADAS-2 tool for studies of diagnostic test accuracy. Appendix F: Methodology checklist: the QUADAS-2 tool for studies of diagnostic test accuracy. The following checklists are taken directly from the 'QUADAS-2' publication (Whiting PF, Rutjes AWS, Westwood ME et al. and the QUADAS-2 group (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine 155: 529–36. The QUADAS website.