Published in final edited form as:

Urol Oncol. 2014 October ; 32(7): 1051–1060. doi:10.1016/j.urolonc.2013.10.003.

Assessing the quality of studies on the diagnostic accuracy of tumor markers

Peter J. Goebell, M.D.^a, Ashish M. Kamat, M.D.^b, Richard J. Sylvester, Ph.D.^c, Peter Black, M.D.^d, Michael Droller, M.D.^e, Guilherme Godoy, M.D.^f, M'Liss A. Hudson, M.D.^g, Kerstin Junker, Ph.D.^h, Wassim Kassouf, M.D.ⁱ, Margaret A. Knowles, Ph.D.^j, Wolfgang A. Schulz, Ph.D.^k, Roland Seiler, M.D.^l, and Bernd J. Schmitz-Dräger, M.D., Ph.D.^{m,n,*}

^aUrologische Klinik, Friedrich-Alexander-Universität, Erlangen, Germany ^bDepartment of Urology, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX ^cEORTC Headquarters, Brussels, Belgium ^dDepartment of Urology, Division of Surgery, University of British Columbia, Vancouver, Canada ^eDepartment of Urology, Mount Sinai Hospital, New York, NY ^fScott Department of Urology, Baylor College of Medicine, Houston, TX ^gOchsner Clinic Foundation, Tom and Gayle Benson Cancer Center, New Orleans, LA ^hUrologische Klinik und Poliklinik, Universität des Saarlandes, Saarland, Germany ⁱDepartment of Surgery (Urology), McGill University, Montreal, Quebec, Canada ^jSection of Experimental Oncology, Leeds Institute of Cancer and Pathology, St James's University Hospital, Leeds, UK ^kUrologische Klinik und Poliklinik, Heinrich-Heine-Universität, Düsseldorf, Germany ^lDepartment of Urology, University of Berne, Berne, Switzerland ^mSchön Klinik Nürnberg Fürth, Fürth, Germany ⁿUrologie 24, Nürnberg, Germany

Abstract

Objectives—With rapidly increasing numbers of publications, assessments of study quality, reporting quality, and classification of studies according to their level of evidence or developmental stage have become key issues in weighing the relevance of new information reported. Diagnostic marker studies are often criticized for yielding highly discrepant and even controversial results. Much of this discrepancy has been attributed to differences in study quality. So far, numerous tools for measuring study quality have been developed, but few of them have been used for systematic reviews and meta-analysis. This is owing to the fact that most tools are complicated and time consuming, suffer from poor reproducibility, and do not permit quantitative scoring.

Methods—The International Bladder Cancer Network (IBCN) has adopted this problem and has systematically identified the more commonly used tools developed since 2000.

Results—In this review, those tools addressing study quality (Quality Assessment of Studies of Diagnostic Accuracy and Newcastle-Ottawa Scale), reporting quality (Standards for Reporting of Diagnostic Accuracy), and developmental stage (IBCN phases) of studies on diagnostic markers in

*Corresponding author. Tel.: +49-911-971-4531; fax: +49-911-971-4532., bernd_sd@yahoo.de (B.J. Schmitz-Dräger).

This article reflects and summarizes discussions held at the 10th Meeting of the International Bladder Cancer Network (IBCN e.V.), Nijmegen, The Netherlands, 20.9.2012 to 22.9.2012.

bladder cancer are introduced and critically analyzed. Based upon this, the IBCN has launched an initiative to assess and validate existing tools with emphasis on diagnostic bladder cancer studies.

Conclusions—The development of simple and reproducible tools for quality assessment of diagnostic marker studies permitting quantitative scoring is suggested.

Keywords

Diagnostic accuracy; Study quality; IBCN classification; Oxford levels of evidence; QUADAS; NOS; STARD

Introduction

With rapidly increasing numbers of publications, assessments of study quality, reporting quality, and classification of studies according to their level of evidence (LoE) or developmental stage have become key issues in weighing the relevance of new information reported. Diagnostic marker studies are often criticized for yielding highly discrepant and even controversial results [1,2]. Thus, for an article on diagnostic accuracy of a molecular bladder cancer marker, it is often nearly impossible to judge the methodological rigor of that study and to conclude whether the published results can be translated to clinical practice.

The International Bladder Cancer Network (IBCN) has adopted this problem for the area of diagnostic and prognostic biomarker research, focusing on studies related to bladder cancer. Recently, the phases reporting and assessment optimization project has been proposed for developing a classification system to describe the developmental status of a given marker in analogy to the commonly accepted phases of clinical trials (phases I–IV) [3,4]. In addition, the IBCN has initiated an analysis of published tools that are used to assess study quality and reporting quality of biomarker studies, exploiting the resources of the IBCN.

Although the use of such tools for the assessment of diagnostic marker trials is recommended, these have generally not been implemented by users, e.g., readers or reviewers. Some of them have been used in systematic reviews and meta-analyses or in education research [5]; however, in many tools sufficient external validation remains pending. One important reason for underutilization of these tools in the urology community is that urology training programs, in general, do not incorporate education on trial design, management, and analysis for their residents; further difficulties of these instruments reside in their deficiency to define what may be considered sufficient or adequate quality. This is in part owing to the great variability in study settings and designs posing great challenges to a given tool with regard to its general applicability. As a consequence, application of most of the tools becomes rather complicated, further preventing their general use. These issues have fueled the development of numerous new instruments without finding a solution of existing problems.

In this context, it is the purpose of this review to introduce, classify, and analyze relevant available assessment tools designed to evaluate studies on the diagnostic accuracy of bladder cancer molecular markers. By this initiative, the use of assessment tools should be supported and, eventually, their practicability and applicability should be improved.

Current tools

A systematic review of medical databases by Dreier et al. [6] identified 17 tools designed to assess studies investigating the diagnostic accuracy of molecular markers. Only the instruments generated after 2000 and those more frequently cited in the literature were considered for this review. For this review, the tools were divided into 4 categories, based upon their objective:

- Study quality: e.g., Newcastle-Ottawa scale [7], Quality Assessment of Studies of Diagnostic Accuracy [QUA-DAS] [8], and the QUADAS-2 tool [9]
- Quality of reporting: e.g., Standards for Reporting of Diagnostic Accuracy [STARD] criteria [10,11]
- Study phases: e.g., IBCN criteria [3,4,12]
- Level of evidence: e.g., Oxford criteria 2001/2009 [13].

Study quality

Newcastle-Ottawa quality assessment scale—The NOS was designed to evaluate the quality of nonrandomized studies, discriminating between case-control trials and cohort studies [7]. Both scales include 3 categories with a total of 8 items (Table 1). When analyzing case-control trials, NOS addresses 3 areas including selection, comparability, and exposure, whereas in cohort studies it includes selection, comparability, and outcome.

This scale was originally developed for application in systematic reviews and meta-analyses. A study can be awarded a maximum of 1 star for each numbered item within the selection and exposure categories in case-control studies, or the selection and outcome categories in cohort studies. A maximum of 2 stars can be given for comparability, in either type of study, resulting in a maximum of 9 points. No cutoff for good or poor quality is provided. The questions are clear and apparently easy to answer; however, the options provided are difficult to apply to some study concepts. Furthermore, the NOS has been criticized for having a high interrater variability [14–17].

The discrimination between case-control studies and cohort trials, as well as its easy applicability, are important factors that explain why the NOS has been frequently used in the past, mainly for systematic reviews and meta-analyses [18,19].

Quality assessment of studies of diagnostic accuracy—The QUADAS instrument is presumably the most widely accepted tool for quality assessment. It is considered a retrospective instrument for evaluation of the methodological rigor of a study investigating the diagnostic accuracy of a given test. The QUADAS tool was developed through a Delphi procedure eventually reducing an initial list of 28 items down to 14 questions [8]. The items include patient spectrum, reference standard, disease progression bias, verification bias, review bias, clinical review bias, incorporation bias, test execution, study withdrawals, and indeterminate results (Table 2). The QUADAS tool is presented together with recommendations for scoring each of the items included. The QUADAS tool provides a matrix in which readers can examine the internal and external validity of a study.

Most items included in QUADAS relate to bias (items 3, 4, 5, 6, 7, 10, 11, 12, and 14); only 2 items relate to variability (items 1 and 2), whereas 3 relate to reporting (items 8, 9, and 13). The questions posed are focused and clear; their accompanying guidelines appear helpful. However, there is much room for subjective interpretation as several items may be answered differently by reviewers based upon their individual perception. Any item may be answered with “yes,” “no,” or “unclear”; however, no advice is provided on scoring, cutoff, and, as a result, on classifying a study as having good or poor quality.

The QUADAS tool has experienced fairly frequent use predominantly within systematic reviews and meta-analyses. For bladder cancer markers, Xia et al. [20] reported its use in a meta-analysis on the accuracy of survival in the diagnosis of bladder cancer.

Several reports have been published regarding the external validation of QUADAS [21,22]. Oliveira et al. [21] applied the QUADAS score alone and in combination with the STARD tool to assess a malaria test in a semiquantitative way. A combination of QUADAS criteria and STARD criteria was compared (see discussion of STARD later in the article) with the QUADAS criteria alone. Articles fulfilling at least 50% of QUADAS criteria were considered as having regular to good quality without providing a definition for this allocation. Of the 13 articles retrieved, 12 fulfilled at least 50% of QUADAS criteria; only 2 fulfilled the combined STARD/QUADAS criteria. The authors concluded that the STARD/QUADAS combination might have the potential to provide greater rigor when evaluating the quality of studies, given that it incorporates relevant information not contemplated in the QUADAS criteria alone.

Hollingworth et al. [22] used data from a systematic review of magnetic resonance spectroscopy in the characterization of suspected brain tumors to provide a preliminary evaluation of the interrater reliability of QUADAS. Nineteen publications were distributed randomly to primary and secondary reviewers for dual independent assessment. Most studies in this review were judged to have used an accurate reference standard. There was good correlation ($\rho = 0.78$) between reviewers in assessment of the overall number of quality criteria met. However, mean agreement for individual QUADAS questions was only fair ($\kappa = 0.22$) and ranged from no agreement ($\kappa < 0$) to moderate agreement ($\kappa = 0.58$). These findings suggest that different reviewers will reach different conclusions when using QUADAS. These findings are similar to those observed by Whiting et al. [23], reporting an adequate interrater reliability for individual items in the QUADAS checklist (range, 50%–100%; median = 90%).

Recently, the QUADAS-2 tool has been presented [9]. It basically follows the original QUADAS tool; however, items were now reduced down to 11 questions in 4 new domains (patient selection, index test(s), reference standard, and flow and timing) (Table 3). In contrast to the original scale, the QUADAS-2 tool provides advice on the rating of study quality. To date, experience concerning the use of this instrument is limited [24,25] and external validation is underway.

Quality of reporting

Although the general quality of the study and the reporting are difficult to separate from each other, the QUADAS tool has already been supplemented in 2003 by another tool, specifically addressing the issue of quality of reporting.

Standards for the reporting of diagnostic accuracy studies—The STARD tool was developed to improve the quality of reporting in diagnostic accuracy studies [10,11]. It comprises 25 items, mirroring the classical sections of an article including title, keywords, abstract (1 item), introduction (1 item), methods (11 items), results (11 items), and discussion (1 item). The reader may rate each item as either “present” or “absent” (Table 4).

Smidt et al. [26] reported on external validation by applying the STARD tool to 32 diagnostic accuracy studies, which was published in medical journals with an impact factor of at least 4 in 2000. All articles were independently reviewed by 2 experts at the beginning of the study and again almost 2 years later.

The overall interassessment agreement for all items of the STARD statement was 85% (Cohen $\kappa = 0.70$) varying from 63% to 100% for individual items. The interassessment reliability of the STARD checklist was satisfactory (intraclass correlation coefficients = 0.79 [95% CI: 0.62–0.89]). The authors concluded that although the overall reproducibility of the quality of reporting using the STARD statement was good, substantial differences were found for specific items. These disagreements were not likely caused by differences in the interpretation by the reviewers but rather by difficulties in assessing the reporting of these items because of lack of clarity within the articles.

In summary, despite some deficiencies concerning reproducibility, the STARD tool is a validated tool for the assessment of reporting quality. However, several issues have emerged with this tool. The underlying questions are not always easy to apply to a given article. Furthermore, no recommendations for scoring are provided that may allow classifying an article as having sufficiently good reporting or not.

Study phases

The definition of study phases addresses the need to identify the current status of development of a given procedure (treatment and diagnostic procedure). This should support an adequate and systematic development of new diagnostic or therapeutic concepts. Owing to a lack of recommendations for the development of diagnostic marker trials, the IBCN phases classification was developed in 2003 (and revised in 2007) in analogy to the 4 phases of clinical trials [3,4,12].

Phase I: Assay development and evaluation of clinical prevalence (feasibility studies)—This phase involves the identification of a target potentially suited for diagnostic use. Identification of the target may occur in many ways, classically by identifying the target in tumor cells. However, with the advent of molecular technology other ways or definitions of a variety of targets are conceivable. The key issue is whether a difference between tumor cells and normal urothelial cells can be demonstrated. It has to be noted that field effects are an integral part of the development of bladder cancer. This warrants inclusion of not only

“normal” adjacent tissue but also tissue and samples from healthy individuals as important controls in evaluating a markers definition.

Phase II: Evaluation studies for clinical utility—“This phase involves optimization of the assay technique (e.g., standardization and automatization) and/or interpretation of the assay results.” The ultimate goal of this phase is to develop hypotheses and to define standards that can be used to perform phase III studies.

Phase II trials are mostly single-institutional studies. However, adequately sized and representative samples of patients may be easier to achieve in a large collaborative network with sufficient numbers of specimens to define and select the most appropriate set of samples. In addition, identifying the sources of variability during this phase of biomarker development is required for designing a phase III study.

Based upon the results in such studies, adequate cutoff values will be defined for quantitative assays. It is essential that the outcome from phase II studies is translated into hypotheses that form the basis for phase III analyses.

Phase III: Confirmation studies—In phase III, hypotheses emerging from previous phase II studies are tested with sufficient power in a defined clinical setting using an independent, prospective, and controlled cohort of patients. The clinical utility of a given marker assay, its performance, and interpretation are established in this phase, the aim of which is the generation of (evidence based) information that may eventually be included into clinical guidelines.

Phase IV: Validation and technology transfer as application studies—The aims of phase IV studies are (1) to transfer the techniques and established methods of the assays and other aspects of the technology into clinical practice and (2) to evaluate the ability of investigators and clinicians at other institutions to apply these methods and interpret the results in a similar and comparable way.

The IBCN classification is easy to use; nevertheless, it has only been rarely applied in the past in systematic reviews [1].

Levels of evidence

A very important dimension in the assessment of articles is the consideration of the LoE that a given study provides. The Oxford Centre for Evidence-based Medicine Levels of Evidence May 2001/2009 classification has been designed to classify the relevance of scientific contributions based mainly upon the study design [13]. Initially aiming at the classification of clinical trials, the 2009 modification also included adaptations for prognostic and diagnostic marker studies as well as for economic and decision analyses. A 5-scale classification was developed, starting from expert opinion (LoE 5) and extending to validating cohort studies for diagnostic markers (LoE 1b) (Table 5). Levels 1 to 3 received subclassifications, with grade “a” representing systematic reviews or meta-analyses of respective trials and grade “b” representing results from a single study. It is of interest that absolute SpPins (case series reporting on highly specific tests, in which a positive result

confirms presence of a disorder) and absolute SnNouts (case series reporting on highly sensitive tests, in which a negative result excludes a disorder) were defined as LoE 1c.

Although the 2009 classification was simple and easy to use, early hierarchies that placed randomized trials categorically above observational studies were criticized for being simplistic. This criticism was met by introducing the 2011 classification providing more flexibility insofar that upgrading and downgrading of studies is possible (Table 6). Furthermore, different clinical settings, e.g., screening, diagnosis, prognosis, and therapy, are discriminated. Subclassifications “a to c” were eliminated, facilitating allocation to the different levels.

Somewhat surprisingly, randomized controlled trials are not listed as a separate level of evidence for diagnostic and prognostic studies, presumably owing to the fact that randomized controlled trials in this field are extremely rare.

Although the LoE classifications 2001/2009 have been well accepted by the scientific community, experience with the 2011 version is still limited.

Discussion

The problem of defining the quality of a given study is as old as scientific communication. An extensive literature search recently performed by Dreier et al. [6] yielded a total of 147 different tools developed to assess study quality. Although there has been a focus on therapeutic trials in the past, more recently instruments for assessment of diagnostic studies have also been developed. The large number of different assessment tools suggests that none of them is accepted as a “perfect” solution for the problem.

Doubtless, the challenge to develop a single tool that can be applied to all diagnostic studies is considerable. In contrast to clinical trials with similar designs comparing standard care vs. a new strategy using criteria defined by good clinical practice guidelines, diagnostic trials may differ with regard to a variety of parameters. Furthermore, the quality of studies is heterogeneous and numerous methodological shortcomings are apparent in the design of diagnostic accuracy studies (Table 7). Finally, definition of study quality is difficult as expectations are different and viewpoints may vary.

One of the most widely used tools to assess study quality is the QUADAS instrument. However, it has been designed—and thus far is exclusively used—for systematic reviews or meta-analyses [8]. One problem in using the QUADAS tool lies in the distinction between general study quality and reporting quality. Inevitably, the assessment of quality relates strongly to the reporting of results; a well-conducted study will score poorly in a quality assessment tool if the methods and results are not reported in sufficient detail. The intention of the STARD document was to complement quality assessment of diagnostic accuracy studies by providing a tool focusing on quality of reporting [10,11]. However, this requires the use of a second instrument and, in consequence, additional time.

Studies failing to report on aspects of quality may be considered as having inferior quality as faulty reporting generally reflects faulty methods. When using QUADAS, another important

factor to consider is the difference between bias and variability. Study bias will limit the validity of the study results, whereas variability may complicate the translation of study results into clinical practice.

It may be questioned whether a separate tool for assessment of reporting quality like STARD is necessary and reasonable, or if the study and the reporting quality are so closely linked that analysis along the STARD criteria is not likely to generate an added value with regard to study assessment [16]. In general, it would be considered preferable to assess overall quality using just a single tool.

A classification of the development phases concerning the status of a new test may well be necessary. The IBCN classification using a 4-phase scale (in analogy to the 4 phases of clinical trials) may constitute a first step in this direction [12]. Thus far, this classification is not generally accepted, despite its simplicity and similarity to clinical study phases. This may be partly due to a lack of precision in some of the definitions; however, the instrument is currently under review for further improvement.

The Oxford LoE 2001/2009 scale has been widely accepted for classifying the scientific effect of a new study [13]. This may be attributed to the facts that (1) it can be more or less universally applied to different study designs, (2) it is clearly structured, and (3) it can be easily used. Furthermore, the LoE classification is a rapid procedure and feasible even for inexperienced scientists. Although the revised 2011 version provides more flexibility, this feature makes the classification much more demanding as former LoE 1b studies may be downgraded to level 2, whereas convincing former level 3 studies might even be considered level 1 in the current system. However, this gain in flexibility may be traded in for a loss in reproducibility and discriminative power.

Neither QUADAS nor STARD can be used to provide a reproducible quantitative value or score for study or reporting quality. At best, the QUADAS instrument provides a qualitative assessment of study design, permitting the conclusion that weaknesses in certain parameters may alter some test findings more than other findings. However, there are several reasons for not incorporating a quality score into QUADAS. Scores are necessary if the investigator intends to use a quantitative indicator of quality to provide weight in a meta-analysis, or if a continuous variable in a meta-regression is required. As quality scores are very rarely used in these ways, the authors of QUADAS felt no need to introduce such a score. They stated that definitions on how to weigh and calculate quality scores might be in fact arbitrary, thus preventing development of an objective quality score [8,23]. In consequence, the application of scores without consideration of the individual items may dilute or entirely miss potential associations.

The authors of this review would challenge this line of reasoning, believing instead that it is necessary to add a semiquantitative estimation of the quality of a given study. In particular, we believe that journal editors and reviewers should be highly interested in tools permitting quantifying quality in a score, thereby permitting a more transparent review process. Furthermore, existing quality assessment tools still include arbitrary and debatable items notwithstanding the care invested in the development process. Finally, it should be the

intention of an assessment tool to permit estimates of study quality or reporting whether formal scoring is included.

Similar to the QUADAS instrument, the NOS has been developed for quality assessment in reviews and meta-analyses [7,27]. In contrast to QUADAS, it permits a semi-quantitative scoring although no cutoff for good/poor quality studies is provided. In a recent validation trial, the interrater reliability of the NOS varied from substantial for the length of follow-up to poor for both the selection of a nonexposed cohort and the fact that the outcome was not present at study outset [15,16]. Investigators reported no association between individual NOS items or overall NOS score and effect on estimates. Variable agreement for the NOS and the lack of evidence showing that it is able to discriminate studies with biased results underscores the need for more detailed guidance to apply this tool in systematic reviews [15,16].

In general, it may be hypothesized that reliability and reproducibility are better achieved by simpler instruments. Final conclusions cannot be drawn as systematic validation of intrarater and interrater reproducibility has rarely been reported for assessment tools in general. In particular, information concerning the use of the instruments by reviewers/investigators with limited experience is lacking.

Based upon this analysis, the authors feel that most instruments available are too complicated and time consuming for an application beyond systematic reviews and meta-analysis (e.g., in peer review or identifying the relevance of a study after reading the article). The reason underlying is the desire to generate comprehensive (ideal) assessment tools covering all possible aspects of quality/reporting quality. To improve the current situation we see a need for 2 measures. In the first step, a simple and robust assessment tool should be developed and validated. In the second step, journal editors and publishers must be encouraged to request reviewing based on such a tool. Acceptance by the reviewers can be obtained if the alternate review process will not require additional time.

As a starting point, the IBCN is planning to support assessment of marker studies through investigation of existing tools for analysis of studies on diagnostic accuracy, delineating limitations, proposing modifications, or, if considered necessary, developing new tools targeting the needs of potential users. Embedded in the phases reporting and assessment optimization initiative, the IBCN is preparing a validation trial for assessment tools focusing on studies on diagnostic accuracy. Instruments directed toward the assessment of study quality and reporting quality will be studied. In addition, further classification instruments for study phases and LoE will be included in this project.

References

1. Schmitz-Dräger, BJ.; Droller, M.; Lotan, Y.; Lokeshwar, VB.; van Rhijn, B.; Hemstreet, GP., et al. Molecular markers for bladder cancer screening, early diagnosis and surveillance. In: Soloway, M.; Khoury, S., editors. Bladder Cancer: 2nd International Consultation on Bladder Tumors. Vol. 21. Paris, France: ICUD Editions; 2012. p. 171-205.
2. van Rhijn BW, van der Poel HG, van der Kwast TH. Urine markers for bladder cancer surveillance: a systematic review. *Eur Urol.* 2005; 47:736–48. [PubMed: 15925067]

3. Goebell PJ, Groshen S, Schmitz-Dräger BJ, Sylvester R, Kogevinas M, Malats N, et al. The International Bladder Cancer Bank: proposal for a new study concept. *Urol Oncol.* 2004; 22:277–84. [PubMed: 15283883]
4. Lotan Y, Shariat SF, Schmitz-Dräger BJ, Sanchez-Carbayo M, Jankevicius F, Racioppi M, et al. Considerations on implementing diagnostic markers into clinical decision making in bladder cancer. *Urol Oncol.* 2010; 28:441–8. [PubMed: 20610281]
5. Cook DA, Levinson AJ, Garside S. Method and reporting quality in health professions education research: a systematic review. *Med Educ.* 2011; 45:227–38. [PubMed: 21299598]
6. Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. *GMS Health Technol Assess.* 2010; 6 [Doc07].
7. Wells, GA.; Shea, B.; O'Connell, D.; Peterson, J.; Welch, V.; Losos, M., et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available at: <http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp>
8. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol.* 2003; 3:25. [PubMed: 14606960]
9. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011; 155:529–36. [PubMed: 22007046]
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem.* 2003; 49:1–6. [PubMed: 12507953]
11. 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. [PubMed: 12507954]
12. Goebell PJ, Groshen SL, Schmitz-Dräger BJ. Guidelines for development of diagnostic markers in bladder cancer. *World J Urol.* 2008; 26:5–11. [PubMed: 18253738]
13. Oxford Centre for Evidence-based Medicine Levels of Evidence. Available at: <http://www.cebm.net>
14. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol.* 2010; 25:603–5. [PubMed: 20652370]
15. Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. *J Clin Epidemiol.* 2013 [pii: S0895-4356(13)00089-00089].
16. Hartling, L.; Hamm, M.; Milne, A.; Vandermeer, B.; Santaguida, PL.; Ansari, M., et al. Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments. Agency for Healthcare Research and Quality (US); Rockville (MD): 2012. [Report No.: 12-EHC039-EF]
17. Oremus M, Oremus C, Hall GB, McKinnon MC. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. *BMJ Open.* 2012; 2:e001368.
18. Tricco AC, Soobiah C, Antony J, Hemmelgarn B, Moher D, Hutton B, et al. Safety of serotonin (5-HT₃) receptor antagonists in patients undergoing surgery and chemotherapy: protocol for a systematic review and network meta-analysis. *Syst Rev.* 2013; 2:46. [PubMed: 23810058]
19. Tong H, Hu C, Yin X, Yu M, Yang J, Jin JA. Meta-analysis of the relationship between cigarette smoking and incidence of myelodys-plastic syndromes. *PLoS One.* 2013; 8:e67537. [PubMed: 23805315]
20. Xia Y, Liu YL, Yang KH, Chen W. The diagnostic value of urine-based survivin mRNA test using reverse transcription-polymerase chain reaction for bladder cancer: a systematic review. *Chin J Cancer.* 2010; 29:441–6. [PubMed: 20346223]
21. Oliveira MR, de Gomes AC, Toscano CM. QUADAS and STARD: evaluating the quality of diagnostic accuracy studies. *Rev Saude Publica.* 2011; 45:416–22. [PubMed: 21412577]

22. Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, et al. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. *Acad Radiol.* 2006; 13:803–10. [PubMed: 16777553]
23. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol.* 2006; 6:9. [PubMed: 16519814]
24. Blomberg BA, Moghbel MC, Saboury B, Stanley CA, Alavi A. The value of radiologic interventions and (18)F-DOPA PET in diagnosing and localizing focal congenital hyperinsulinism: systematic review and meta-analysis. *Mol Imaging Biol.* 2013; 15:97–105. [PubMed: 22752652]
25. Beynon R, Hawkins J, Laing R, Higgins N, Whiting P, Jameson C, et al. The diagnostic utility and cost-effectiveness of selective nerve root blocks in patients considered for lumbar decompression surgery: a systematic review and economic model. *Health Technol Assess.* 2013; 17(v–vi):1–88.
26. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC Med Res Methodol.* 2006; 6:12. [PubMed: 16539705]
27. Cook C, Cleland J, Huijbregts P. Creation and critique of studies of diagnostic accuracy: use of the STARD and QUADAS methodological quality assessment tools. *J Man Manip Ther.* 15:93–102. [PubMed: 19066649]

Table 1

NOS items for assessment of study quality of diagnostic studies [5]

Case-control studies	Cohort studies
A study can be awarded a maximum of 1 star for each numbered item within the Selection and Exposure categories. A maximum of 2 stars can be given for Comparability.	
<i>Selection</i>	<i>Selection</i>
1 Is the case definition adequate? <ul style="list-style-type: none"> a. yes, with independent validation b. yes, e.g., record linkage or based on self reports c. no description 	1 Representativeness of the exposed cohort <ul style="list-style-type: none"> a. truly representative of the average ____ (describe) in the community b. somewhat representative of the average ____ in the community c. selected group of users, e.g., nurses and volunteers d. no description of the derivation of the cohort
2 Representativeness of the cases <ul style="list-style-type: none"> a. consecutive or obviously representative series of cases b. potential for selection biases or not stated 	2 Selection of the nonexposed cohort <ul style="list-style-type: none"> a. drawn from the same community as the exposed cohort b. drawn from a different source c. no description of the derivation of the nonexposed cohort
3 Selection of controls <ul style="list-style-type: none"> a. community controls b. hospital controls c. no description 	3 Ascertainment of exposure <ul style="list-style-type: none"> a. secure record (e.g., surgical records) b. structured interview c. written self report d. no description
4 Definition of controls <ul style="list-style-type: none"> a. no history of disease (endpoint) b. no description of source 	4 Demonstration that outcome of interest was not present at start of study <ul style="list-style-type: none"> a. yes b. no
<i>Comparability</i>	<i>Comparability</i>
1 Comparability of cases and controls on the basis of the design or analysis <ul style="list-style-type: none"> a. study controls for ____ (Select the most important factor.) b. study controls for any additional factor (this criteria could be modified to indicate specific control for a second important factor.) 	1 Comparability of cohorts on the basis of the design or analysis <ul style="list-style-type: none"> a. study controls for ____ (select the most important factor.) b. study controls for any additional factor (this criteria could be modified to indicate specific controls for second important factor.)
<i>Exposure</i>	<i>Outcome</i>
1 Ascertainment of exposure <ul style="list-style-type: none"> a. secure record (e.g., surgical records) b. structured interview where blind to case/control status c. interview not blinded to case/control status d. written self report or medical record only e. no description 	1 Assessment of outcome <ul style="list-style-type: none"> a. independent blind assessment b. record linkage c. self report d. no description
2 Same method of ascertainment for cases and controls <ul style="list-style-type: none"> a. yes (select an adequate follow up period for outcome of interest) 	2 Was follow-up long enough for outcomes to occur <ul style="list-style-type: none"> a. yes (select an adequate follow up period for outcome of interest)

Case-control studies		Cohort studies	
	a. yes		b. no
	b. no		
3	Non-response rate	3	Adequacy of follow up of cohorts
	a. same rate for both groups		a. complete follow up—all subjects accounted for
	b. non respondents described		b. subjects lost to follow up unlikely to introduce bias— small number lost—> ____% (select an adequate % follow up, or description provided of those lost)
	c. rate different and no designation		c. follow up rate < ____% and no description of those lost
			d. no statement

Table 2

QUADAS tool for assessment of study quality of diagnostic studies [6]

Item	Yes	No	Unclear
1. Was the spectrum of patients representative of the patients who will receive the test in practice?	0	0	0
2. Were selection criteria clearly described?	0	0	0
3. Is the reference standard likely to correctly classify the target condition?	0	0	0
4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	0	0	0
5. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	0	0	0
6. Did patients receive the same reference standard regardless of the index test result?	0	0	0
7. Was the reference standard independent of the index test (i.e., the index test did not form part of the reference standard)?	0	0	0
8. Was the execution of the index test described in sufficient detail to permit replication of the test?	0	0	0
9. Was the execution of the reference standard described in sufficient detail to permit its replication?	0	0	0
10. Were the index test results interpreted without knowledge of the results of the reference standard?	0	0	0
11. Were the reference standard results interpreted without knowledge of the results of the index test?	0	0	0
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	0	0	0
13. Were uninterpretable/intermediate test results reported?	0	0	0
14. Were withdrawals from the study explained?	0	0	0

Table 3**QUADAS-2 tool for assessment of study quality of diagnostic studies [9]**

<i>Domain 1: Patient selection</i>	
A. Risk of bias	
Describe methods of patients selection:	
• Was a consecutive or random sample of patients enrolled?	Yes/no/unclear
• Was case-control design avoided?	Yes/no/unclear
• Did the study avoid inappropriate exclusions?	Yes/no/unclear
Could the selection of patients have introduced bias?	Risk: Low/high/unclear
B. Concerns regarding applicability	
Describe included patients (prior testing, presentation, intended use of index test and setting):	
Is there concern that the included patients do not match the review question?	Concern: Low/high/unclear
<i>Domain 2: Index test(s)</i>	
If more than one index test was used, please complete for each test.	
A. Risk of bias	
Describe the index test and how it was conducted and interpreted:	
• Were the index test results interpreted without knowledge of the results of the reference standard?	Yes/no/unclear
• If threshold was used, was it pre-specified?	Yes/no/unclear
Could the conduct or interpretation of the index test have introduced bias?	Risk: Low/high/unclear
B. Concerns regarding applicability	
Is there concern that the index test, its conduct, or interpretation differ from the review question?	Concern: Low/high/unclear
<i>Domain 3: Reference standard</i>	
A. Risk of bias	
Describe the reference standard and how it was conducted and interpreted:	
• Is the reference standard likely to correctly classify the target?	Yes/no/unclear
• Were the reference standard results interpreted without knowledge of the results of the index test?	Yes/no/unclear
Could the reference standard, its conduct, or its interpretation have introduced bias?	Risk: Low/high/unclear
B. Concerns regarding applicability	
Is there concern that the target condition as defined by the reference standard does not match the review question?	Concern: Low/high/unclear
<i>Domain 4: Flow and timing</i>	
A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2 × 2 table (refer to flow diagram):	
Describe the time interval and any interventions between index test(s) and reference standard:	
• Was there an appropriate interval between index tests(s) and reference standard?	Yes/no/unclear
• Did all patients receive a reference standard?	Yes/no/unclear
• Did patients receive the same reference standard?	Yes/no/unclear
• Were all patients included in the analysis?	Yes/no/unclear
Could the patient flow have introduced bias?	Risk: Low/high/unclear

Table 4

The STARD tool for assessment of reporting quality in diagnostic studies [8,9]

Section and topic	Item no.	On page no.
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	<i>The study population:</i> The inclusion and exclusion criteria, setting and locations where the data were collected.
	4	<i>Participant recruitment:</i> 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	<i>Participant sampling:</i> 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.
	6	<i>Data collection:</i> 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/or cite references for index tests and reference standard.
	9	Definition of and rationale for the units, cutoffs and/or 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 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 the recruitment.
	15	Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centers).
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or 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 diseases (define criteria) in those with the target condition; other diagnosis 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	Estimate of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).
	22	How indeterminate results, missing responses and outliers of the index tests were handled.

Section and topic	Item no.	On page no.
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.
	24	Estimate of the test reproducibility, if done.
Discussion	25	Discuss the clinical applicability of the study findings.

Table 5

Oxford Center of Evidence-Based Medicine (OCEBM) 2009 criteria for diagnostic and prognostic marker trials [13]

Level	Prognosis	Diagnosis
1a	Systematic review (SR) (with homogeneity) of inception cohort studies; CDR ^a validated in different populations	SR (with homogeneity) of level 1 diagnostic studies; CDR ^a with 1b studies from different clinical centers
1b	Individual inception cohort study with >80% follow-up; CDR ^a validated in a single population	Validating cohort study with good reference standards; or CDR ^a tested within one clinical centre
1c	All or none case series	Absolute SpPins and SnNouts ^b
2a	SR (with homogeneity) of either retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity) of level >2 diagnostic studies
2b	Retrospective cohort study or follow-up of untreated control patients in an RCT; derivation of CDR ^a or validated on split-sample ^c only	Exploratory cohort study with good reference standards; CDR ^a after derivation, or validated only on split-sample ^c or databases
2c	“Outcomes” research	–
3a	–	SR (with homogeneity) of 3b and better studies
3b	–	Non-consecutive study; or without consistently applied reference standards
4	Case series (and poor quality prognostic cohort studies)	Case-control study, poor or non-independent reference standard
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”	Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”

RCT = randomized controlled trial.

^a Clinical decision rule (these are algorithms or scoring systems that lead to a prognostic estimation or a diagnostic category).

^b An “Absolute SpPin” is a diagnostic finding whose specificity is so high that a positive result rules in the diagnosis. An “Absolute SnNout” is a diagnostic finding whose sensitivity is so high that a negative result rules out the diagnosis.

^c Split-sample validation is achieved by collecting all the information in a single tranche, then artificially dividing this into “derivation” and “validation” samples.

Table 6
Oxford Center of Evidence-Based Medicine (OCEBM) 2011 criteria for diagnostic and prognostic marker trials [13]

Question	Step 1 (level 1 ^a)	Step 2 (level 2 ^a)	Step 3 (level 3 ^a)	Step 4 (level 4 ^a)	Step 5 (level 5)
Is the diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards ^b	Case-control studies, or poor or non-independent reference standard ^b	Mechanism-based reasoning

^a Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; level may be graded up if there is a large or very large effect size.

^b As always, a systematic review is generally better than an individual study.

Table 7

Frequent methodological shortcomings and parameters varying between diagnostic accuracy trials

Study design
Sample size
Patient selection
Selection of adequate control population
Prevalence of target condition
Technique/standardization
Test experience
Insufficient operational definition of positive and negative test findings
Cutoff definition (e.g., post hoc definition)
Absence of a third category of indeterminate test findings
Use of an inappropriate gold standard or reference test
Lack of rater blinding

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript