Risk of bias assessment of test comparisons was uncommon in comparative accuracy systematic reviews

Yang, Bada ; Vali, Yasaman ; Sharifabadi, Anahita Dehmoobad Sharifabadi; Harris, Isobel ; Beese, Sophie; Davenport, Clare; Hyde, Christopher; Takwoingi, Yemisi; Whiting, Penny; Langendam, Miranda; Leeflang, Mariska M G

DOI:
10.1016/j.jclinepi.2020.08.007

ELSEVIER

Journal of Clinical Epidemiology ■ (2020) ■

Journal of Clinical Epidemiology

REVIEW

Risk of bias assessment of test comparisons was uncommon in comparative accuracy systematic reviews: an overview of reviews

Bada Yang[a,*], Yasaman Vali[a], Anahita Dehmoobad Sharifabadi[b], Isobel Marion Harris[c], Sophie Beese[c], Clare Davenport[c,d], Christopher Hyde[e], Yemisi Takwoingi[c,d], Penny Whiting[f], Miranda W. Langendam[a], Mariska M.G. Leeflang[a]

[a]*Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105AZ, Amsterdam, The Netherlands*
[b]*Department of Radiology, Faculty of Medicine, University of Ottawa, Roger Guindon Hall, 451 Smyth Rd #2044, Ottawa, Ontario K1H 8M5, Canada*
[c]*Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK*
[d]*NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK*
[e]*Exeter Test Group, Institute of Health Research, College of Medicine and Health, University of Exeter, Exeter, UK*
[f]*Population Health Sciences, Bristol Medical School, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK*

Accepted 10 August 2020; Published online xxxx

Abstract

Objectives: Comparative diagnostic test accuracy systematic reviews (DTA reviews) assess the accuracy of two or more tests and compare their diagnostic performance. We investigated how comparative DTA reviews assessed the risk of bias (RoB) in primary studies that compared multiple index tests.

Study Design and Setting: This is an overview of comparative DTA reviews indexed in MEDLINE from January 1st to December 31st, 2017. Two assessors independently identified DTA reviews including at least two index tests and containing at least one statement in which the accuracy of the index tests was compared. Two assessors independently extracted data on the methods used to assess RoB in studies that directly compared the accuracy of multiple index tests.

Results: We included 238 comparative DTA reviews. Only two reviews (0.8%, 95% confidence interval 0.1 to 3.0%) conducted RoB assessment of test comparisons undertaken in primary studies; neither used an RoB tool specifically designed to assess bias in test comparisons.

Conclusion: Assessment of RoB in test comparisons undertaken in primary studies was uncommon in comparative DTA reviews, possibly due to lack of existing guidance on and awareness of potential sources of bias. Based on our findings, guidance on how to assess and incorporate RoB in comparative DTA reviews is needed.

Keywords: Diagnostic accuracy; Bias; Test comparison; Meta-analysis; Systematic review

1. Introduction

Clinicians and health care providers need to select the most accurate diagnostic tests available to reduce the consequences of false positive and/or false negative test results. The best available evidence regarding the accuracy of a test is summarized in a diagnostic test accuracy systematic review (DTA review), which aims to assess the accuracy of a test of interest (index test) against a reference standard. DTA reviews usually assess the accuracy of a single index test, but increasingly also compare the accuracy of two or more index tests. These ''comparative DTA'' reviews enable decision makers to choose the most accurate test, especially when evidence regarding the effectiveness of a test on patient outcomes is unavailable [1].

DTA reviews, like any other type of systematic review, should include assessment of the risk of bias (RoB) in included studies [2,3]. In a DTA review assessing the

Conflict of interests: All authors declare that there are no conflicts of interest.

Funding: Amsterdam UMC (the Netherlands) and National Institute for Health Research (the United Kingdom) provided funding for this study. The funding organizations had no role in the design, collection, analysis, and interpretation of the data or the decision to approve publication of the finished manuscript.

* Corresponding author. Department of Epidemiology and Data Science, Amsterdam UMC, Room J1b-210, Location AMC Meibergdreef 9, 1105AZ Amsterdam, The Netherlands. Tel: +31(0)20 5666948.

E-mail address: b.d.yang@amsterdamumc.nl (B. Yang).

https://doi.org/10.1016/j.jclinepi.2020.08.007

What is new?

Key findings

Risk of bias assessment was rarely undertaken in systematic reviews that compare the accuracy of two or more diagnostic tests. In reviews that performed risk of bias assessment, this was performed with tools that were not designed for diagnostic accuracy studies.

What this adds to what was known?

Risk of bias assessment is a key feature in systematic reviews. While risk of bias assessment for individual test accuracy was almost always performed, the assessment for comparative accuracy was usually overlooked.

What is the implication and what should change now?

Comparative diagnostic test accuracy systematic reviews that do not undertake risk of bias assessment should be interpreted with caution. Guidance and tools that address potential biases in test comparisons should be developed.

accuracy of a single test, the ideal study design is one in which a consecutive series or random sample of patients undergo both the index test and the reference standard. These studies are assessed for RoB and concerns regarding applicability using the QUADAS-2 tool [4]. However, when comparing the accuracy of two or more index tests, reviews should ideally include studies that directly compare the index tests in a single study. QUADAS-2 was not developed to assess the RoB of the comparison between tests in a primary study. Therefore, additional sources of bias need to be considered (Table 1). For example, both index tests should ideally be evaluated in the same patients, so that any differences between the tests cannot be attributed to differences in factors that may influence test accuracy (e.g., patient spectrum). Alternatively, in a study where patients are randomized to undergo one index test or another, the randomization should be adequate to produce index test groups that are comparable with regard to patient spectrum. Reviews that fail to consider these and other potential sources of bias may draw conclusions based on flawed comparisons, which have the potential to mislead users of evidence. Such an approach in DTA reviews would be analogous to systematic reviews of interventions that do not assess the internal validity of the comparison (e.g., allocation concealment) in randomized controlled trials (RCTs).

We are not aware of the RoB tools for test comparisons in primary studies, although Cochrane recommends that when using QUADAS-2 to assess comparative studies included in a comparative DTA review, review authors should consider adding a ‘‘comparative domain’’ with relevant signaling questions [5,6]. At least two Cochrane DTA review protocols have implemented this recommendation, by adding signaling questions to QUADAS-2 such as ‘‘Were the same participant selection criteria used for those allocated to each test?’’ and ‘‘Was each index test result interpreted without knowledge of the results of other index tests or testing strategies?’’ [7,8]. In addition, Wade et al describe an effort to modify QUADAS-2 to assess quality of studies comparing the accuracy of colposcopy technologies [9]. They added 10 signaling questions to QUADAS-2 specifically for test comparisons, such as ‘‘Were the index and the comparator tests independent?’’ and ‘‘Was there an appropriate interval between the index and the comparator tests?’’. These efforts indicate that RoB of test comparisons is an issue pertinent to comparative DTA reviews, and that there is a need to identify RoB tools currently in existence.

Therefore, the objective of this overview of reviews is to assess (1) whether recent comparative DTA reviews assessed the RoB of test comparisons undertaken in included primary studies, and if so, (2) which methods have been used for RoB assessment.

2. Methods

2.1. Study design

This is a methodological overview of comparative DTA reviews. The protocol of this overview was preregistered on PROSPERO (CRD42018099111).

2.2. Terms and definitions

We define a ‘‘comparative DTA review’’ as a DTA review that compared the accuracy of two or more index tests. Because not all DTA reviews with multiple index tests will have the intention of comparing tests, we also required that the review should contain at least one statement, anywhere in the review, indicating a comparison between the accuracy of the index tests. A comparison between an index test and the reference standard is not a test comparison because a reference standard is needed to assess the accuracy of an index test.

We refer to a primary DTA study that compared two or more index tests as a ‘‘comparative DTA study’’. Examples of comparative DTA studies include the ‘‘paired’’ design, in which a series of patients undergo both index tests and the reference standard, and the ‘‘randomized’’ design, in which patients are randomly allocated to receive either index test A or B, and subsequently receive the same reference standard (Figure 1). We refer to a primary DTA study that only includes one index test as a ‘‘noncomparative DTA study’’.

Table 1. Differences between single test accuracy and comparative accuracy studies

Characteristic	Single test accuracy study	Comparative accuracy study
Clinical question being answered	How accurately can a single index test classify individuals who have or do not have the target condition?	How does the accuracy of index test A compare with that of index test B?
Ideal study design	A study in which a consecutive series or random sample of patients all undergo a single index test and the reference standard	1). A study in which each participant undergoes all index tests and the reference standard (paired or within-subject design) or 2). A study in which participants are randomly allocated to an index test and all participants get the same reference standard
Examples of risk of bias[a]	The patient spectrum of those who have the target condition only includes advanced disease	The patient spectrum differs between those who get index test A and those who get index test B
	The index test was interpreted with knowledge of the reference standard result	Index test A was interpreted with knowledge of the results of index test B
	The reference standard does not correctly classify the target condition	Results of index test A and B are verified against a different reference standard for each test
	There is an inappropriate time interval between the index test and the reference standard	There is an inappropriate time interval between index test A and B

[a] The examples of risk of bias given for comparative accuracy studies are additional to those given for single test accuracy studies.

2.3. Data sources and searches

We searched MEDLINE (Ovid interface) from January 1st 2015 to February 15th, 2018, to identify recently published DTA reviews. However, *post hoc* we restricted the eligibility to reviews published in the most recent complete year at the time of search (from 1st January to 31st December 2017) because we identified substantially more records than we had anticipated (n = 11,702). We regarded this to be a safe decision, as earlier reviews are less likely to contain examples of RoB assessment of test comparisons, compared with more recent reviews. The search strategy contained terms relating to systematic reviews and meta-analyses in combination with terms relating to test accuracy (Appendix A). No specific keywords filtering comparative DTA reviews were applied. We did not restrict on language or type of index test.

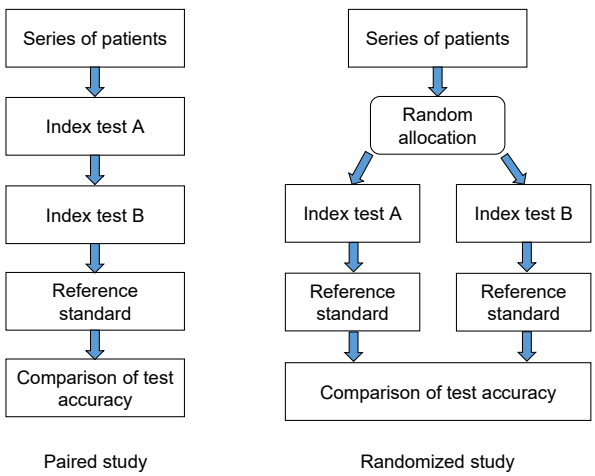

Fig. 1. Examples of comparative DTA study designs. Abbreviations: DTA, diagnostic test accuracy.

2.4. Review selection

We included comparative DTA reviews consistent with our definition: systematic reviews (1) evaluating the accuracy of two or more index tests, verified by a reference standard, and (2) containing at least one sentence in which the review authors made a comparison between index tests. We considered literature reviews to be systematic reviews if the review reported a search strategy and if the review explicitly reported eligibility criteria for study selection. We did not consider comparisons restricted to multiple thresholds of a single test. We included comparative DTA reviews regardless of whether comparative DTA studies were included in the review, as it is possible that reviews detailed methods for RoB assessment in these studies, but did not find any studies.

We excluded reviews of animal studies, review protocols, reviews of predictive accuracy, and reviews for which the full-text articles could not be retrieved. We also excluded reviews without a clinical target condition (e.g., reviews assessing the accuracy of database search filters) and reviews evaluating only one index test that made a comparison with a test not evaluated in the same review.

Each article was assessed for eligibility by two authors independently (B.Y. assessed all articles, and independent assessment was conducted by A.S., I.M.H., or S.B.). The assessment of eligibility was performed in two steps. First, the titles and abstracts of each article were reviewed. Potentially eligible articles were then further assessed based on their full-text. Disagreements were resolved by discussion, or by taking into account the opinion of a third assessor.

2.5. Data extraction

While extracting data for our objectives, we also extracted data to describe the characteristics of our cohort of DTA reviews. As reviews often described more than one test comparison, we identified and extracted data on the first reported comparison in the article.

We extracted data using a piloted form that included the following items (the complete list of data extraction items is available in Appendix B).

- Objective of the review.
- Characteristics of the first comparison in the review (we recorded the number of index tests, type of studies in the comparison, type of index test and the target condition). We also assessed whether authors explicitly specified the ''role of the index test'', namely replacement, triage or add-on, in the introduction [10], following STARD 2015 reporting guidelines [11].
- Inclusion of comparative DTA studies (we recorded whether comparative DTA studies were mentioned as inclusion criteria and whether these were separately reported from noncomparative DTA studies).
- Whether or not RoB assessment of individual test accuracy was performed (if yes, we recorded which tool was used).
- Whether or not RoB assessment of test comparisons in comparative DTA studies was performed (if yes, we recorded which tool was used, which items were present in the tool, and whether existing tools were modified for test comparisons).
- If RoB assessment for comparisons was not performed, whether authors discussed potential RoB issues in test comparisons as a limitation of the evidence in the discussion.

When reviews failed to identify any studies for inclusion but reported clear methods for RoB assessment, we nevertheless extracted data as if the methods had been applied. For instance, if a review reported that it would perform RoB assessment in the methods but did not include any comparative DTA studies in the review, we recorded ''the review performed RoB assessment''.

Data from each review were extracted by two authors independently (B.Y. extracted data for all reviews, and independent extractions were conducted by Y.V., A.S. or M.W.L.). Disagreements were resolved by discussion.

2.6. Data synthesis and analysis

We used descriptive statistics to summarize categorical variables as frequencies and percentages using the total number of included comparative DTA reviews as the denominator for most analyses. In cases where the total number of reviews was not the denominator, we explicitly stated this. We computed a 95% confidence interval for the proportion of reviews that conducted RoB assessment for test comparisons using the Clopper-Pearson (''exact'') approach.

3. Results

3.1. Search results

Our search retrieved 4,085 records published in 2017. After removal of duplicates, we screened 4,014 titles and abstracts and included 416 records for full-text assessment. We excluded 178 records at this phase, most common reasons being (1) no comparison between index tests (n = 67); (2) not a systematic review (n = 43); and (3) not a DTA review (n = 19). We included 238 comparative DTA reviews in this overview, of which 13 were Cochrane systematic reviews. Figure 2 displays the review inclusion flow chart and Appendix C contains the list of included reviews.

3.2. Characteristics of included comparative DTA reviews

Table 2 describes characteristics of the 238 included comparative DTA reviews. Of the 238 reviews, the index tests in the comparison were most often imaging (n = 119; 50%) or biochemical (n = 80; 34%) tests; target conditions were most frequently neoplasms (n = 107; 45%), infectious diseases (n = 25; 11%), or diseases of the digestive system (n = 25; 11%). Most of the 238 reviews compared two tests in the first comparison (n = 150; 63%), but comparisons involving 5 or more tests were also common (n = 43; 18%).

Around half of the reviews (n = 134; 56%) were explicit about their intention to compare tests. For the remaining 104 (44%) reviews, the comparative nature of the review only became apparent in the results, discussion, or conclusion section of the article. Of the 134 reviews with an explicit comparative question, 115 (86%) reviews stated that test comparison was an objective of the review, and 19 (14%) reviews described a test comparison in their methods.

The intended role of the index tests in the clinical pathway (triage, replacement, or add-on) was specified in 2 of 238 (1%) reviews. In other reviews, this was not specified, or the specified role was not relevant for the comparison (e.g., the role of the index tests was to replace another test not in the comparison).

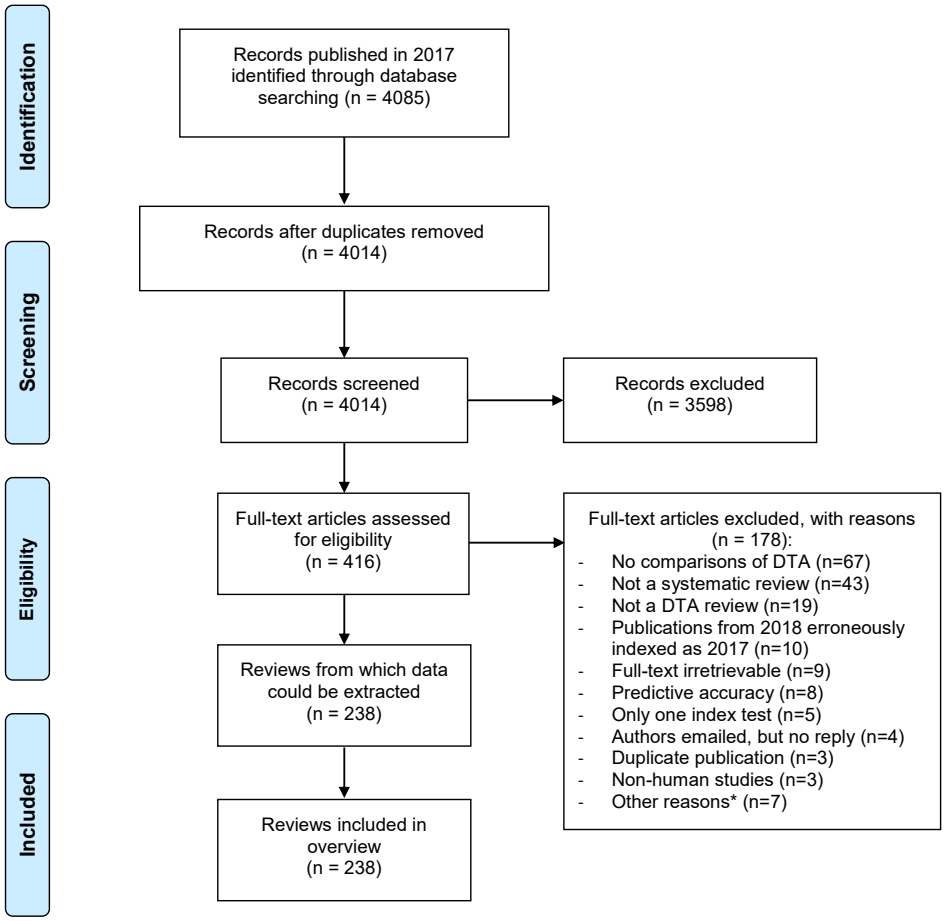

Fig. 2. Review inclusion flow chart. *Other reasons: not clinically relevant target condition, publication from 2016, review of prediction models using CHARMS checklist, regression modeling of individual patient data. Abbreviations: DTA, diagnostic test accuracy.

3.3. Type of studies in comparative DTA reviews

In 54 (23%) reviews, the comparison was based on data from comparative DTA studies only. The comparison was based on noncomparative DTA studies in 35 (15%) reviews and most frequently, a combination of comparative and noncomparative DTA studies in 123 (52%) reviews. In 22 (9%) reviews, we were unable to determine the type of studies included in the review and 4 (2%) reviews did not include any studies. Fifty-six (24%) reviews reported comparative DTA studies or a description of comparative DTA studies as one of the criteria for inclusion, and of these, 30 planned to solely include comparative DTA studies for comparing tests.

Of the 123 reviews in which the comparison was based on both comparative and noncomparative DTA studies, 19 (15%) also reported results of comparative DTA studies separately from noncomparative studies, whereas 104 (85%) reviews did not. Among the 19 reviews reporting results of comparative and noncomparative DTA studies separately, 12 reviews performed a separate meta-analysis for only comparative DTA studies, 2 reviews performed sensitivity analysis with only comparative DTA studies, and 5 reviews narratively reported the results of comparative DTA studies separately.

3.4. Risk of bias assessment of comparative DTA studies

Most reviews performed RoB assessment for the accuracy of an individual test ($n = 213$; 90%) with 163 reviews using the QUADAS-2 tool, 36 reviews using the original QUADAS tool, and 14 reviews using other methods.

Two of the 238 (0.8%; 95% CI, 0.1 to 3.0%) reviews conducted RoB assessment of test comparisons undertaken in comparative DTA studies. Neither review was a Cochrane systematic review. Both reviews restricted inclusion to studies that randomized participants to index test A or B and subsequently performed the reference standard, allowing estimation of the accuracy of each index test. The first review [20] included randomized studies of chromoendoscopy versus other endoscopic techniques for dysplasia surveillance in inflammatory bowel disease, using histopathology as the reference standard. This review used the Cochrane RoB tool for RCTs [21] to assess random sequence generation, allocation concealment, blinding of participants, investigators and outcome assessors, incomplete outcome data, and selective reporting. The second review [22] included randomized studies that compared the accuracy of 22 versus 25 gauge needles for fine needle aspiration of pancreatic lesions using a poorly specified

Table 2. Characteristics of 238 comparative DTA reviews

Characteristic	N	%
Total	238	100
Comparative objective or methods		
Comparison is objective of review	115	48
Comparison is planned in methods	19	8
Comparison inferred in results, discussion or conclusion	104	44
Number of index tests in the comparison		
2	150	63
3	22	9
4	14	6
≥5	43	18
Unclear	9	4
Type of studies in the comparison		
Comparative DTA studies	54	23
Noncomparative DTA studies	35	15
Combination of comparative and noncomparative DTA studies	123	52
Unclear	22	9
No studies included	4	2
Type of index tests compared[a]		
Imaging	119	50
Biochemical	80	34
Clinical	26	11
Questionnaire	13	6
Pathology	12	5
Combination of multiple types	8	3
Other[b]	15	6
Target conditions (ICD-11 classification)		
Neoplasms	107	45
Infectious diseases	25	11
Digestive system	25	11
Musculoskeletal	13	6
Mental/behavioral disorders	11	5
Other	57	24
Role of the index tests in the comparison		
Replacement	2	1
Not reported	236	99

Abbreviations: DTA, diagnostic test accuracy; ICD-11, International Classification of Diseases, Eleventh Revision.

As a review may include multiple comparisons, we collected data on the first reported comparison in the review.

[a] There can be multiple types of index tests per review.

[b] Auditory brainstem response, audiometric tests, dermatological tests, dental tests, electronystagmography, evoked potentials, intelligent systems, optical examinations, optical spectroscopy, portable recording device, urological tests.

reference standard that included surgical pathological diagnosis. It also reported using the Cochrane RoB tool for RCTs, but further details of quality assessment were not provided. Neither review undertook quality assessment of the estimation of individual test accuracy.

Among the 236 reviews without RoB assessment of test comparisons, 39 (17%) did not contain comparative DTA studies. In the remaining 197 reviews that included comparative DTA studies (or where the type of included studies was unclear), 8 of 197 (4%) reviews discussed potential RoB in test comparisons in comparative DTA studies as a limitation of the evidence, whereas 189 (96%) reviews did not discuss any potential RoB issues specific to the comparison. Potential RoB issues highlighted by the 8 reviews are listed in Table 3.

4. Discussion

4.1. Summary of the evidence

In this overview, we aimed to identify whether and how RoB assessment of test comparisons was conducted in recent comparative DTA reviews. We found that while most reviews conducted RoB assessment for estimation of the accuracy of a single test, reviews rarely considered RoB assessment for test accuracy comparisons.

In our sample of 238 reviews, only two (0.8%; 95% CI, 0.1 to 3.0%) performed RoB assessment of test comparisons. In these two reviews [20,22], authors used the Cochrane RoB tool for RCTs [21], which assesses the adequacy of randomization, but does not address other aspects of validity pertinent to DTA studies. Notably, neither review conducted RoB assessment of issues specific to estimation of individual test accuracy (for example, by using QUADAS-2) or comparative accuracy.

We also addressed the question: if reviews do not assess the RoB of test comparisons, do they discuss potential RoB as a limitation of the evidence? We found that the potential RoB issues in test comparisons were only discussed in 8 of 197 (4%) reviews that included, or potentially included comparative DTA studies.

Our findings are concerning because the full extent of the validity of the comparative evidence in these reviews

Table 3. Risk of bias issues in test comparisons in comparative DTA studies discussed by review authors as limitations of the evidence

Potential risk of bias issues	Reviews (n = 8)
Interpreting test A with knowledge of test B [12–14]	3
Index tests not compared in the same patients [15]	1
Nonrandom allocation to index tests [16]	1
Learning effects of the endoscopist by always performing test A before B [17]	1
Noncomparability of endoscopies (in vivo versus image interpretation) [12]	1
Reference standard interpreted with knowledge of the index tests [18]	1
Inclusion of only patients who can tolerate both index tests [19]	1

Abbreviations: DTA, diagnostic test accuracy.

is unclear. Furthermore, in the absence of RoB assessment for test comparisons, review authors may erroneously assume that the validity of the estimated accuracy of individual tests also applies to the comparison of their accuracy. Additional sources of bias need to be considered when assessing the validity of test comparisons (Table 1) and ignoring these ultimately carries the risk that clinicians and health care providers may make decisions based on systematic reviews of DTA without being properly informed about the limitations of the comparative evidence.

Although there are no clear reasons why comparative DTA reviews poorly consider test comparisons in RoB assessment, an explanation may be the lack of appropriate tools and guidance on assessing comparative DTA evidence. While some guidance is available in the literature [5,6,9,23], RoB assessment of test comparisons has not yet been adequately addressed or developed. Our overview demonstrates that, in the increasingly important field of comparative accuracy, the development of dedicated tools and guidance for RoB assessment of test comparisons is necessary.

4.2. Limitations

To gain an unbiased, representative sample of recent comparative DTA reviews, we included all comparative DTA reviews published in 2017. Although our cohort of reviews is large, we may have missed other examples of RoB assessments in test comparisons. In this overview, we defined comparative DTA reviews as reviews that report a comparison of index tests. Hence, we also included reviews that did not explicitly state comparative objectives or methods, but contained a statement in which the tests were compared. It may be argued that these are not comparative DTA reviews, because it is not always clear whether the intention of the review authors was to compare tests. We decided to take this approach as the reader may interpret the evidence as comparative, regardless of the intention of the review authors. It is also possible that review authors did plan to perform RoB assessments for test comparisons, but decided it was unfeasible or did not report the assessment. We did not check the protocols of included reviews, nor did we contact the review authors. Finally, there is inherent subjectivity in assessing textual data. Aspects of review selection and data extraction required judgment (e.g., whether a sentence contains a comparison, whether the role of the index tests was reported) which was complicated due to frequent unclear reporting of review characteristics. We tried to reduce subjectivity by duplicate and independent review selection and data extraction.

4.3. Implications for practice and research

We present some initial considerations on how prospective authors of comparative DTA reviews could assess RoB in test comparisons. Comparative DTA studies represent potentially higher quality evidence compared with noncomparative DTA studies [24] when comparing multiple index tests. While comparative DTA studies are not always available, reviewers should aim to include these studies and separately report their results when possible. A validated tool to assess RoB in test comparisons in comparative DTA studies is not yet available. In the absence of such a tool, reviewers should consider which sources of bias are important in their primary studies. For example, reviewers could add signaling questions to QUADAS-2 asking whether index test A was interpreted without knowledge of index test B and vice versa if interpretation of one or both tests is subjective. We have collated examples of potential sources of bias reported by reviews in this overview (Table 3). An effort is currently underway to extend the QUADAS-2 tool to assess RoB in comparative DTA studies, preliminary named QUADAS-C [25].

5. Conclusion

In this overview, almost all reviews assessed RoB of individual test accuracy but seldom assessed the RoB in test comparisons undertaken in comparative DTA studies. The minority of reviews that considered the RoB of test comparisons focused mainly on the adequacy of random allocation to index tests and did not address the RoB in the estimation of test accuracy. This demonstrated lack of RoB assessment of test comparisons is an important limitation in comparative DTA reviews. Our overview highlights the need to develop appropriate methods and guidance on how to assess RoB in test comparisons.

CRediT authorship contribution statement

Bada Yang: Project administration, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Yasaman Vali:** Investigation, Writing - review & editing. **Anahita Dehmoobad Sharifabadi:** Investigation, Writing - review & editing. **Isobel Marion Harris:** Investigation, Writing - review & editing. **Sophie Beese:** Investigation, Writing - review & editing. **Clare Davenport:** Conceptualization, Writing - review & editing. **Christopher Hyde:** Conceptualization, Writing - review & editing. **Yemisi Takwoingi:** Conceptualization, Writing - review & editing. **Penny Whiting:** Conceptualization, Writing - review & editing. **Miranda W. Langendam:** Investigation, Writing - review & editing, Supervision. **Mariska M.G. Leeflang:** Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision.

Acknowledgments

Y.T. is supported by a National Institute for Health Research (NIHR) Postdoctoral Fellowship. Y.T. and C.D.

are supported by the NIHR Birmingham Biomedical Research Centre. This article presents independent research supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Appendix A

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.08.007.

References

[1] Takwoingi Y. Meta-analytic approaches for summarising and comparing the accuracy of medical tests. University of Birmingham Research Archive; 2016.
[2] Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration 2011. Available at www.handbook.cochrane.org. Accessed January 17, 2019.
[3] McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies the PRISMA-DTA statement. JAMA 2018;319(4):388−96.
[4] Whiting P, Rutjes A, Westwood M, 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.
[5] Leeflang M, Davenport C, Takwoingi Y. Sources of bias. Lesson 6.1: Cochrane Collaboration DTA Online Learning Materials. The Cochrane Collaboration 2014. Available at http://training.cochrane.org/path/diagnostic-test-accuracy-dta-reviews-pathway. Accessed March 21, 2019.
[6] Takwoingi Y, Leeflang M, Davenport C. How to convert your review to QUADAS-2. Lesson 9.2: Cochrane Collaboration DTA Online Learning Materials. The Cochrane Collaboration 2014. Available at http://training.cochrane.org/path/diagnostic-test-accuracy-dta-reviews-pathway. Accessed March 21, 2019.
[7] Dinnes J, Matin R, Moreau J, Patel L, Chan SA, Chuchu N, et al. Tests to assist in the diagnosis of cutaneous melanoma in adults: a generic protocol. Cochrane Database Syst Rev 2015;10.
[8] Rai N, Champaneria R, Snell K, Mallett S, Bayliss SE, Neal RD, et al. Symptoms, ultrasound imaging and biochemical markers alone or in combination for the diagnosis of ovarian cancer in women with symptoms suspicious of ovarian cancer. Cochrane Database Syst Rev 2015;12.
[9] Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies : our experience using a modified. version of the QUADAS-2 tool. Res Synth Methods 2013;4.
[10] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089−92.
[11] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015351.
[12] Backes Y, Moss A, Reitsma JB, Siersema PD, Moons LM. Narrow band imaging, magnifying chromoendoscopy, and gross morphological features for the optical diagnosis of T1 colorectal cancer and deep submucosal invasion: a systematic review and meta-analysis. Am J Gastroenterol 2017;112(1):54−64.
[13] Samim M, Molenaar IQ, Seesing MFJ, van Rossum PSN, van den Bosch MAAJ, Ruers TJM, et al. The diagnostic performance of18F-FDG PET/CT, CT and MRI in the treatment evaluation of ablation therapy for colorectal liver metastases: a systematic review and meta-analysis. Surg Oncol 2017;26(1):37−45.
[14] Zhang QW, Teng LM, Zhang XT, Zhang JJ, Zhou Y, Zhou ZR, et al. Narrow-band imaging in the diagnosis of deep submucosal colorectal cancers: a systematic review and meta-analysis. Endoscopy 2017;49:564−80.
[15] Yun SJ, Ryu C-W, Choi NY, Kim HC, Oh JY, Yang DM. Comparison of Low- and Standard-Dose CT for the Diagnosis of Acute Appendicitis: A Meta-Analysis. Am J Roentgenol 2017;208(6):W198−207.
[16] Berger N, Luparia A, Di Leo G, Carbonaro LA, Trimboli RM, Ambrogi F, et al. Diagnostic performance of MRI versus galactography in women with pathologic nipple discharge: a systematic review and meta-analysis. Am J Roentgenol 2017;209(2):465−71.
[17] Xiong Y, Li J, Ma S, Ge J, Zhou L, Li D, et al. A meta-analysis of narrow band imaging for the diagnosis and therapeutic outcome of non-muscle invasive bladder cancer. PLoS One 2017;12:1−14.
[18] Duncan JK, Ma N, Vreugdenburg TD, Cameron AL, Maddern G. Gadoxetic acid-enhanced MRI for the characterization of hepatocellular carcinoma: a systematic review and meta-analysis. J Magn Reson Imaging 2017;45(1):281−90.
[19] Singnurkar A, Poon R, Metser U. Comparison of 18F-FDG-PET/CT and 18F-FDG-PET/MR imaging in oncology: a systematic review. Ann Nucl Med 2017;31(5):366−78.
[20] Iannone A, Ruospo M, Wong G, Principi M, Barone M, Strippoli GF, et al. Chromoendoscopy for surveillance in ulcerative colitis and crohn's disease: a systematic review of randomized trials. Clin Gastroenterol Hepatol 2017;15(11):1684−1697.e11.
[21] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:1−9.
[22] Facciorusso A, Stasi E, Di Maso M. Endoscopic ultrasound-guided fine needle aspiration of pancreatic lesions with 22 versus 25 Gauge needles: a meta-analysis. United Eur Gastroent J 2017;5(6):846−53.
[23] Reitsma H, Rutjes A, Whiting P, Vlassov VV, Leeflang MM, Deeks JJ, et al. Chapter 9: Assessing Methodological Quality. Cochrane Database Syst Rev; 2009.
[24] Leeflang MMG, Reitsma JB. Systematic reviews and meta-analyses addressing comparative test accuracy questions. Diagn Progn Res 2018;2(17).
[25] QUADAS-2C Group. Development of QUADAS-2C, a quality assessment tool for comparative diagnostic accuracy studies: a delphi study protocol 2018. Available at https://osf.io/tmze9. Accessed September 17, 2019.