Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist

Jean-Paul Salameh,[1] Patrick M Bossuyt,[2] Trevor A McGrath,[3] Brett D Thombs,[4] Christopher J Hyde,[5] Petra Macaskill,[6] Jonathan J Deeks,[7,8] Mariska Leeflang,[9] Daniël A Korevaar,[10] Penny Whiting,[11] Yemisi Takwoingi,[7,8] Johannes B Reitsma,[12] Jérémie F Cohen,[13] Robert A Frank,[3] Harriet A Hunt,[5] Lotty Hooft,[12] Anne W S Rutjes,[14] Brian H Willis,[15] Constantine Gatsonis,[16] Brooke Levis,[17] David Moher,[18] Matthew D F McInnes[19]

For numbered affiliations see end of the article.

Correspondence to:
M D F McInnes
mmcinnes@toh.ca
(ORCID 0000-0001-8404-4075)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;370:m2632
http://dx.doi.org/10.1136/bmj.m2632

Accepted: 20 May 2020

Systematic reviews of diagnostic test accuracy (DTA) studies are fundamental to the decision making process in evidence based medicine. Although such studies are regarded as high level evidence, these reviews are not always reported completely and transparently. Suboptimal reporting of DTA systematic reviews compromises their validity and generalisability, and subsequently their value to key stakeholders. An extension of the PRISMA (preferred reporting items for systematic review and meta-analysis) statement was recently developed to improve the reporting quality of DTA systematic reviews. The PRISMA-DTA statement has 27 items, of which eight are unmodified from the original PRISMA statement. This article provides an explanation for the 19 new and modified items, along with their meaning and rationale. Examples of complete reporting are used for each item to illustrate best practices.

SUMMARY POINTS

PRISMA DTA is a checklist of 27 items to guide the reporting of diagnostic accuracy systematic reviews

The checklist contains two new items, two omitted items, and 17 modified items from the original PRISMA checklist

PRISMA DTA enables transparent and complete reporting that will enhance both reproducibility and the ability to determine quality of evidence

The understanding of diagnostic test performance can be enhanced through diagnostic test accuracy (DTA) systematic reviews. When performed following rigorous methodology, systematic reviews can improve our understanding of a specific intervention or diagnostic test.[1-3] However, published systematic reviews, including DTA reviews, are often insufficiently informative and therefore of limited use.[4-6] Incomplete reporting of systematic reviews prevents stakeholders who rely on health research from critically assessing the quality of evidence and could lead to patient harm, misallocation of resources, and research waste.[7-9]

An extension of the PRISMA (preferred reporting items for systematic review and meta-analysis) statement was recently developed to facilitate complete and transparent reporting of DTA systematic reviews, along with another PRISMA extension for abstracts.[10-12] The PRISMA-DTA statement includes 27 items; eight of the 27 original PRISMA items were unmodified, 17 original items were modified, two new items were added, and another two were omitted.

This article is modelled after similar explanation and elaboration documents for other reporting guidelines.[13-17] This document should be used concurrently with the PRISMA-DTA statement, which includes the PRISMA-DTA checklist (table 1).[10] Box 1 also explains terminology used throughout the checklist. PRISMA-DTA is not meant to be a comprehensive guide on how to perform a DTA systematic review; readers are directed towards other resources for such guidance, such as the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.[18]

PRISMA-DTA items that were either added or modified (relative to the PRISMA statement) are discussed in this document, followed by published examples of complete reporting for each item. Elaboration on the rationale for the inclusion of the item, reporting deficiencies, and relevant supporting evidence are presented. Items from the original PRISMA statement that were not modified for PRISMA-DTA are listed but not discussed. An independent explanation and elaboration document for the PRISMA-DTA checklist for abstracts (in preparation) will expand on the rationale for the addition or modification of new items and provides examples of optimal reporting of abstracts of DTA systematic reviews.

PRISMA-DTA item 1: title

Identify the report as a systematic review (with or without meta-analysis) of DTA studies.

Examples

1. Diagnostic accuracy of saline contrast sonohysterography in detecting endometrial

polyps in women with postmenopausal bleeding: systematic review and meta-analysis.[19]
2. Diagnostic accuracy of segmental enhancement inversion for diagnosis of renal oncocytoma at biphasic contrast enhanced CT [computed tomography]: systematic review.[20]

Explanation

A clear title identifying the work as a systematic review and, if conducted, a meta-analysis, of DTA studies serves two purposes. It allows readers to immediately identify that the study purpose is to evaluate diagnostic accuracy, rather than other measures of diagnostic performance, and it allows for easy identification when searching for or indexing systematic reviews.

Authors are encouraged to include relevant terms regarding the study participants, index test, target condition, and comparisons made, if applicable, such that readers can easily locate the study when performing a search, and rapidly identify whether the systematic review is pertinent to their clinical query.

PRISMA-DTA item 2: abstract

The PRISMA-DTA for abstracts checklist (table 2) and explanation and elaboration document describe what should be reported in the abstract of a DTA review.

PRISMA-DTA item 3 (not modified from original PRISMA): rationale

Describe the rationale for the review in the context of what is already known.

PRISMA-DTA item D1 (new item): introduction

State the scientific and clinical background, including the intended use and clinical role of the index test, and if applicable, the rationale for minimally acceptable test accuracy (or minimum difference in accuracy for a comparative design).

Examples

1. "[S]putum induction is time-consuming, needs experienced laboratory personnel, and many patients are unable to produce adequate samples. Several minimally invasive markers of eosinophilic airway inflammation . . . could have potential as a surrogate to replace sputum induction, but their accuracy to distinguish between patients with and without airway eosinophilia remains controversial. We did a systematic review and meta-analysis to obtain summary estimates of the diagnostic accuracy of markers for airway eosinophilia in patients with asthma."[21]
2. "An a priori minimum diagnostic accuracy for DECT [dual energy computed tomography] to be considered sufficient was defined as an area under the receiver operating characteristic (AUROC) curve of 0.95 and a minimum specificity of 0.95. These consensus values were based on consultation with three fellowship trained endourologists who practice in a tertiary care center. False positive test results are problematic as effective treatment would be delayed due to failed dissolution therapy of non-uric acid stones, which may cause patient harm. No minimum sensitivity has been determined as patients with false negative test results would receive current standard treatment and are considered at lower risk of potential harm than patients with false positive test results."[22]

Explanation

If the intended use and clinical role of the index test being evaluated have not yet been completely defined, explicitly stating the exploratory use of the index test is recommended, because it will limit just how definitive the review can be to support decisions. The clinical background in the introduction explains the choices that will be made later in the review in formulating the review question (item 4), defining eligibility criteria (item 6), identifying potential applicability concerns (item 12), and interpreting the results (item 26).

When evaluating a potential replacement test, a systematic review might aim to evaluate whether a test confers improved accuracy; in other situations, the benefit of a test might be its greater ease of use (as in example 1), and the purpose of the review is to evaluate whether accuracy is compromised relative to more complex alternatives. If possible, the minimally acceptable test accuracy of the index test (example 2), or difference in test accuracy relative to comparator tests that might be used, to detect a condition should be provided, with a rationale.

In example 1, the target condition is eosinophilic airway inflammation in patients with asthma, because patients with eosinophilic airway inflammation are more likely to respond to corticosteroid treatment. The intended use is treatment selection and the potential clinical role is replacement: sputum induction is recommended by clinical guidelines (as an add-on test to clinical criteria) because applying this test in clinical practice has been shown to reduce the number of asthma exacerbations but is insufficiently feasible. The review aims at identifying minimally invasive markers that might replace this test in the existing clinical pathway, thereby saving time, costs, and effort. The authors do not define minimally acceptable test accuracy in this example, but a replacement test should generally be at least as accurate as the existing test. However, other properties might have a role in defining the minimally clinically important differences in accuracy. For instance, when replacing an invasive test with a non-invasive one, some loss of accuracy could be tolerated. Similarly, when introducing a point-of-care diagnostic test, the benefit of increased access and timing might be traded against lower accuracy. Whatever the choice made by authors regarding minimum accuracy, the rationale for the decision should be clearly stated.

Table 1 | PRISMA-DTA checklist

Section/topic	Item No	PRISMA-DTA checklist item
Title/abstract		
Title	1	Identify the report as a systematic review (+/−meta-analysis) of diagnostic test accuracy (DTA) studies
Abstract	2	Abstract: see PRISMA-DTA checklist for abstracts
Introduction		
Rationale	3	Describe the rationale for the review in the context of what is already known
Clinical role of index test	D1	State the scientific and clinical background, including the intended use and clinical role of the index test, and if applicable, the rationale for minimally acceptable test accuracy (or minimum difference in accuracy for comparative design)
Objectives	4	Provide an explicit statement of question(s) being addressed in terms of participants, index test(s), and target condition(s)
Methods		
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (eg, web address), and, if available, provide registration information including registration number
Eligibility criteria	6	Specify study characteristics (participants, setting, index test(s), reference standard(s), target condition(s), and study design) and report characteristics (eg, years considered, language, publication status) used as criteria for eligibility, giving rationale
Information sources	7	Describe all information sources (eg, databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched
Search	8	Present full search strategies for all electronic databases and other sources searched, including any limits used, such that they could be repeated
Study selection	9	State the process for selecting studies (that is, screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis)
Data collection process	10	Describe method of data extraction from reports (eg, piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators
Definitions for data extraction	11	Provide definitions used in data extraction and classifications of target condition(s), index test(s), reference standard(s) and other characteristics (eg, study design, clinical setting)
Risk of bias and applicability	12	Describe methods used for assessing risk of bias in individual studies and concerns regarding the applicability to the review question
Diagnostic accuracy measures	13	State the principal diagnostic accuracy measure(s) reported (eg, sensitivity, specificity) and state the unit of assessment (eg, per patient, per lesion)
Synthesis of results	14	Describe methods of handling data, combining results of studies and describing variability between studies. This could include, but is not limited to: a) handling of multiple definitions of target condition. b) handling of multiple thresholds of test positivity, c) handling multiple index test readers, d) handling of indeterminate test results, e) grouping and comparing tests, f) handling of different reference standards
Meta-analysis	D2	Report the statistical methods used for meta-analyses, if performed
Additional analyses	16	Describe methods of additional analyses (eg, sensitivity or subgroup analyses, meta-regression), if done, indicating which were prespecified
Results		
Study selection	17	Provide numbers of studies screened, assessed for eligibility, included in the review (and included in meta-analysis, if applicable) with reasons for exclusions at each stage, ideally with a flow diagram
Study characteristics	18	For each included study provide citations and present key characteristics including: a) participant characteristics (presentation, prior testing), b) clinical setting, c) study design, d) target condition definition, e) index test, f) reference standard, g) sample size, h) funding sources
Risk of bias and applicability	19	Present evaluation of risk of bias and concerns regarding applicability for each study
Results of individual studies	20	For each analysis in each study (eg, unique combination of index test, reference standard, and positivity threshold) report 2×2 data (TP, FP, FN, TN) with estimates of diagnostic accuracy and confidence intervals, ideally with a forest or receiver operator characteristic (ROC) plot
Synthesis of results	21	Describe test accuracy, including variability; if meta-analysis was done, include results and confidence intervals
Additional analysis	23	Give results of additional analyses, if done (eg, sensitivity or subgroup analyses, meta-regression; analysis of index test: failure rates, proportion of inconclusive results, adverse events)
Discussion		
Summary of evidence	24	Summarise the main findings including the strength of evidence
Limitations	25	Discuss limitations from included studies (eg, risk of bias and concerns regarding applicability) and from the review process (eg, incomplete retrieval of identified research)
Conclusions	26	Provide a general interpretation of the results in the context of other evidence. Discuss implications for future research and clinical practice (eg, the intended use and clinical role of the index test)
Funding		
Funding	27	For the systematic review, describe the sources of funding and other support and the role of the funders

TP=true positives; FP=false positives; FN=false negatives; TN=true negatives. Original version of checklist is also included in the supplementary material.

In comparative reviews, the clinical role must be specified for each of the index tests. This ensures that primary studies are selected with diagnostic pathways that specifically address the intended roles of the index tests being compared.[23]

PRISMA-DTA item 4: objectives

Provide an explicit statement of question being addressed in terms of participants, index test, and target conditions.

Examples

1. "We did a systematic review and meta-analyses of studies evaluating the diagnostic accuracy of the Elecsys Troponin T high-sensitive assay . . . for early diagnosis of acute myocardial infarction in patients presenting to the emergency department with chest pain".[24]
2. "To summarise and compare the accuracy of transabdominal ultrasound (TAUS) and endoscopic ultrasound (EUS) for the detection of

Box 1: Terminology

Systematic review
Synthesis of all relevant primary research studies using a rigorous methodological approach to answer a clearly defined research question. With a well documented search strategy, identified articles are included in the review if they meet prespecified eligibility criteria. Systematic reviews can provide high quality evidence to guide decision making in healthcare, owing to the reliability of the findings derived through systematic approaches that minimise bias.

Meta-analysis
Statistical approach for combining results from multiple studies included in a systematic review. Meta-analysis is a common but not a necessary component of a systematic review.

Diagnostic test accuracy (DTA) studies
Studies that evaluate the ability of an index test to distinguish between participants with and those without a prespecified target condition. DTA studies estimate the sensitivity and specificity of a test. These summary statistics allow for comparisons between the accuracy of different tests.

Index test
Test of interest evaluated in a DTA study. The sensitivity and specificity of the index test are estimated by comparing results of the index test to those of a reference standard applied to the same participants.

Reference standard
Test (or combination of tests/procedures) that is deemed to be the best available method to categorise participants as having or not having a target condition.

Target condition
Clearly defined health or disease state of participants which the test is used to identify. Evaluation of the performance of an index test depends on how accurately it identifies the target condition in study participants.

Risk of bias
Systematic errors that threaten the validity of the findings. In DTA systematic reviews, bias can be due to methodological or clinical misconduct in four areas of the included studies, as highlighted in the QUADAS-2 tool: patient selection (eg, were participants enrolled consecutively), index test (eg, was the assessment of the index test blinded to the reference standard results), reference standard (eg, is the reference standard sufficiently accurate), or flow and timing (eg, is the time between the index test and the reference standard short enough).

Applicability concerns
In a DTA systematic review, concerns regarding applicability can arise when the selection of participants, implementation of the index test, or target condition of the primary studies differ from those specified in the review question.

Quality assessment of diagnostic accuracy studies (QUADAS)-2 tool
Tool for the quality assessment of diagnostic accuracy studies that evaluates the quality of individual studies included in a systematic review in terms of potential risk of bias, and concerns about applicability to the review question.

Publication bias
Publication bias is when the decision to publish part or all of the results of a study depends on the study findings.

Table 2 | PRISMA-DTA for abstracts checklist

Section/topic	Item	PRISMA-DTA for abstracts checklist item
Title and purpose		
Title	1	Identify the report as a systematic review (+/−meta-analysis) of diagnostic test accuracy (DTA) studies
Objectives	2	Indicate the research question, including components such as participants, index test, and target conditions
Methods		
Eligibility criteria	3	Include study characteristics used as criteria for eligibility
Information sources	4	List the key databases searched and the search dates
Risk of bias and applicability	5	Indicate the methods of assessing risk of bias and applicability
Synthesis of results	A1	Indicate the methods for the data synthesis
Results		
Included studies	6	Indicate the number and type of included studies and the participants and relevant characteristics of the studies (including the reference standard)
Synthesis of results	7	Include the results for the analysis of diagnostic accuracy, preferably indicating the number of studies and participants. Describe test accuracy including variability; if meta-analysis was done, include summary results and confidence intervals
Discussion		
Strengths and limitations	9	Provide a brief summary of the strengths and limitations of the evidence
Interpretation	10	Provide a general interpretation of the results and the important implications
Other		
Funding	11	Indicate the primary source of funding for the review
Registration	12	Provide the registration number and the registry name

Original version of checklist is also included in the supplementary material.

gallbladder polyps, for differentiating between true and pseudo gallbladder polyps . . . in adults".[25]

Explanation

The central focus of this item is to describe all components of the review questions, with explicit reference to participants, index tests, and target conditions (PIT), which differs from the traditional PICO approach (participants, intervention, control, outcome) used in systematic reviews of intervention studies. Criteria for considering studies eligible for including in a review and search methods for identification rely on the PIT criteria.

Tests could have a different accuracy in different populations; hence the characteristics of the included participants are important. Also, the selection of patients, and preceding and subsequent patient care steps followed before and after testing might differ between settings. Therefore, a description of the participants should also include the setting in which they were tested. The type of index tests should be clearly described, including sufficient detail to ensure that readers can understand whether findings are generalisable to their practice, and any other details specifying the precise nature and application of the index tests. The target condition should, if applicable, include international standardised terminology (eg, the World Health Organization's International Classification of Diseases). All details relating to staging, severity, and symptomatology of the condition should be included here in order to clearly differentiate the target condition being addressed from other, possibly similar, conditions.

The comparator should be defined carefully (as often done in the PICO approach, for interventional reviews) in the review objective because of ambiguity about whether this refers to an alternate index test, current diagnostic practice, or the reference standard.

PRISMA-DTA item 5 (not modified from original PRISMA): protocol and registration

Indicate whether a review protocol exists, indicate whether and where it can be accessed (eg, web address), and, if available, provide registration information including the registration number.

PRISMA-DTA item 6: eligibility criteria

Specify study characteristics (participants, setting, index test, reference standards, target conditions, and study design) and report characteristics (eg, years considered, language, publication status) used as criteria for eligibility and providing rationale.

Example

"Patients living in enteric fever-endemic areas attending a healthcare facility with fever were eligible . . . All rapid diagnostic tests (RDTs) specifically designed to detect enteric fever cases [were eligible] . . . Studies may have compared one or more RDT against one or more reference standards . . . Studies were required to diagnose enteric fever using one of the following reference standards: (1) bone marrow culture; (2) peripheral blood culture, peripheral blood PCR [polymerase chain reaction], or both . . . [Target conditions included] typhoid fever caused by *Salmonella enterica* serovar Typhi [and] paratyphoid fever caused by *Salmonella enterica* serovar Paratyphi A."[26]

Explanation

Eligibility criteria are expected to involve both study characteristics and report characteristics, because one report might describe more than one study, and one study might be described in multiple reports. Each of these eligibility criteria should be sufficiently described to allow replication, and a rationale should be provided when alternatives exist. A clear set of inclusion and exclusion criteria successfully guides the screening process, and ultimately the final selection of what is included in the review, in a systematic and reproducible manner. It also informs the development of the literature search strategy and allows for an appraisal of the validity, applicability, and comprehensiveness of the systematic review itself.

For participant and setting characteristics, authors are advised to describe any requirements for the presentation (eg, specific signs and symptoms such as fever), previous diagnostic testing, and, if applicable, the clinical settings (eg, healthcare facilities located in areas where enteric fever is endemic).

Details on the type of index tests should be provided, along with the comparator tests, if applicable. Additional details can include a description of who is doing the test, and aspects of the testing process such as specimen type and handling and transport of specimens. For study design, authors should describe which type of design is considered, specifically, if both comparative and single test accuracy designs will be considered, and if any restriction applies for the study sample size or the number of diseased participants included in a study.

Authors should be explicit on the inclusion of studies with multiple groups (also known as multiple gate studies, and previously often referred to as diagnostic case-control studies).[27] These multiple group studies can lead to biased estimates of accuracy.[28 29] Authors should provide a clear definition of the target condition and the reference standard(s) that will be considered for inclusion. If the topic of interest concerns a target condition that can only be established after a reasonable length of time, authors are expected to specify the length of follow-up required for the reference standard.

For reference standards and index tests with multiple categories or continuous results, authors should specify whether studies are required to report outcome data at specific positivity thresholds or result categories, or whether data from all thresholds reported in primary studies will be included. Comparative DTA reviews largely rely on non-comparative primary studies, where only one of the index tests has been investigated. Inclusion of this study type can lead to comparisons made between study populations

with different characteristics, varying diagnostic pathways, and reference standards.[30 31] To reduce this source of bias in comparative DTA reviews, authors should consider including only comparative primary study designs (all patients get all tests, or patients randomised to tests)

Eligibility criteria related to study reports typically concern language of publication, publication status (eg, published, unpublished, in press, or ongoing), and year of publication. Complete reporting of eligibility criteria for reports ensures reproducibility and generalisability.

PRISMA-DTA item 7 (not modified from original PRISMA): information sources

Describe all information sources (eg, databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.

PRISMA-DTA item 8: search

Present full search strategies for all electronic databases and other sources searched, including any limits used so that they can be repeated.

Example

"The search included MEDLINE, Embase, and Cochrane Central Register of Controlled Trials (CENTRAL). No date restrictions were applied. Language of publication was limited to English. Full details of the database search including coverage dates for each database are presented in appendix 1 [https://link.springer.com/article/10.1007/s00330-019-06559-0#Sec10]."[22]

Explanation

Replicability of a systematic review includes replicability of the search strategy. The review report should provide a complete description of the methods used for study retrieval (eg, electronic, grey literature, expert contact, reference lists), who did the searches, which electronic databases were used, and the dates when the searches were performed. This information should include the actual search terms for at least one of the common bibliographic databases. If the string of search terms is too lengthy, it can be reported in the appendices of the review (as supplementary material), where authors can also indicate how it was modified for other databases. Authors should report whether the search strategy was reviewed by independent information specialists using the evidence based guideline for peer review of electronic search strategies,[32] or using the guidance for describing search strings for systematic reviews in the form of PRISMA-S (available on the open science framework, https://osf.io/ygn9w/).

PRISMA-DTA item 9 (not modified from original PRISMA): study selection

State the process for selecting studies (that is, screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).

PRISMA-DTA item 10 (not modified from original PRISMA): data collection process

Describe method of data extraction from reports (eg, piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.

PRISMA-DTA item 11: definitions for data extraction

Provide definitions used in data extraction and classifications of target conditions, index tests, reference standards, and other characteristics (eg, study design, clinical setting).

Example

"TP [true positive] was considered a diagnosis of solid renal mass on DECT [dual energy computed tomography] confirmed by the reference standard (including RCC [renal cell carcinoma], AML [angiomyolipoma], oncocytoma, and renal abscess). TN [true negative] was considered a diagnosis of a non-solid renal lesion on DECT confirmed by the reference standard. FP [false positive] was considered a diagnosis of solid renal mass on DECT confirmed to be a benign cyst by the reference standard, and FN [false negative] was considered a diagnosis of non-solid lesion on DECT confirmed to be a solid renal mass by the reference standard".[33]

Explanation

To facilitate the interpretation of the review findings and to allow replication, clear definitions should be given for extracting data for all critical components of the review. This includes the patient population and setting, index test and target condition, and reference standard, but also the methods used to identify patients with the target condition.

Authors are encouraged to report the different thresholds for test positivity (whether numerical or based on a specific finding) and the different stages and grades of disease (or target condition), when applicable (eg, tumours). Transparency in the definitions of test positivity and target condition is not only fundamental for any effort of reproducibility but is also necessary for defining disease positivity (in the example, through providing the positivity thresholds used based on the grade of the tumour).

Authors can refer readers to the study protocol or record in a trial registry,[34] provide a detailed summary of the relevant definitions within their methods section, and include extraction forms as a supplementary file.

In many cases when extracting data, review authors will need to contact the study investigators with the request to provide additional details or to confirm extracted data that were not clearly stated in the report (eg, method of sampling, the overall number of participants with and without the target condition). The authors should report in their review which studies and variables were identified through this approach. The review authors should indicate if any outcome data were imputed, for which study and for which elements this was done.

PRISMA-DTA item 12: risk of bias and applicability

Describe methods used for assessing risk of bias in individual studies and concerns regarding the applicability to the review question.

Example

"Quality assessment of studies was performed using the QUADAS [quality assessment of diagnostic accuracy studies]-2 tool, examining bias and applicability of the studies with respect to four separate domains: patient selection, index test, reference standard and the flow and timing of patients through the study. No overall summary score was calculated, but for each domain, any concern with regards to bias and applicability were qualified as 'low', 'high' or 'unclear'. These results were then presented in graph and table form."[35]

Explanation

Limitations in the design or conduct of a primary DTA study can result in estimates of diagnostic accuracy that differ systematically from the truth; this is known as "bias" (box 1). Sources of variation are also very important to consider when interpreting the results of a diagnostic accuracy study. Estimates of accuracy might vary because of differences in study populations, how the test was conducted or in how the target condition is defined. Although estimates of accuracy could be unbiased, they might not apply directly to the specific review question.[36] Potential sources of bias and concerns regarding applicability should be considered when interpreting the results of a DTA study.

When reporting the results of a DTA systematic review, the criteria used to assess the risk of bias and concerns regarding the applicability of included primary studies should be clearly defined to facilitate the interpretation and make replication and update possible. This clear definition will allow readers of the review to determine whether appropriate criteria were used and whether all potential sources of bias and applicability were considered.

Authors must also provide details of the selected tool and how it was applied. For example, QUADAS (quality assessment of diagnostic accuracy studies)-2 is a systematically developed, evidence based tool comprised of four domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of risk of bias, and the first three domains are also assessed in terms of concerns regarding applicability.[37]

If QUADAS-2 is used, any modifications to the signalling questions should be reported. QUADAS-2 encourages users to adapt the guidance to make it specific to the review, helping reviewers determine what would be considered high risk of bias in the context of their review question. Any modifications to the guidance should also be reported. Space in the review text is often insufficient to provide full details on rating guidance or modifications, but this can be provided as supplementary material.

The process used for assessing risk of bias and applicability should also be reported. This information includes details such as the number of reviewers involved (eg, two independent reviewers), the process for resolving disagreements (eg, through discussion or referral to a third reviewer), and whether any piloting was conducted to achieve consensus on rating guidance before assessing all studies.

A description in the methods section of the review of how the results of the quality assessment were summarised and incorporated into the review is recommended. The use of quality scores (scales that numerically summarise multiple components into a single score) is discouraged; these quality scores have been shown to be misleading.[38 39] Instead, a description of the methods used for an overall assessment of risk of bias for one study is preferred. For example, if QUADAS-2 is used, guidance suggests that any domain judged at high risk of bias makes the whole study at high risk of bias.

PRISMA-DTA item 13: diagnostic accuracy measures

State the principal diagnostic accuracy measures reported (eg, sensitivity, specificity) and state the unit of assessment (eg, per patient *v* per lesion).

Example

1. "We used the data from the two-by-two tables to calculate sensitivity and specificity for each study. We present individual study results graphically by plotting the estimates of sensitivity and specificity (and their 95% confidence intervals) in both forest plots and the receiver operating characteristic (ROC) space."[40]
2. "In our primary meta-analyses, we used the individual participant as the unit of analysis (that is, any abnormal finding versus none) and not individual ultrasound findings. Clinically, it is also useful to know the accuracy of individual ultrasound findings, as it is plausible that some findings are better indicators of tuberculosis than others. We therefore determined the accuracy of individual ultrasound findings in secondary analyses."[41]

Explanation

Diagnostic accuracy metrics summarise the performance of the test as evaluated against a reference standard, which captures the presence of the target condition. Many different metrics can be used to express a test's accuracy.[42] The most commonly used metrics in meta-analyses are sensitivity (the probability of the test correctly identifying those with disease) and specificity (correctly excluding disease in those without disease).[40] Occasionally meta-analyses summarise positive and negative predictive values (probabilities that positive and negative test results correctly indicate or exclude disease, respectively), diagnostic odds ratios (the ratio of the odds of a positive test among individuals with the disease relative to those without), or areas under receiver operating characteristic (ROC) curves.

The choice of the most appropriate metric should be guided by the review question and the preferred

study design of the test accuracy studies included in the review. For example, the use of a different reference standard in test positives and in test negatives can limit meaningful calculation of sensitivity and specificity when the reference standards have important differences in the misclassification rates; instead, authors might prefer to calculate and report positive and negative predictive values. Diagnostic odds ratios are somewhat limited because they do not provide information on the numbers of false positives and false negatives, for which the consequences typically differ.

The unit of analysis and the type of collected data will affect estimates of these metrics. Usually, the presence or absence of the target condition is analysed on a per patient basis; occasionally a per lesion classification is more relevant (eg, where the intervention is delivered at lesion level).

PRISMA-DTA item 14: synthesis of results

Describe the methods of handling the data, combining the results of the studies, and describing the variability between studies. The descriptions could include handling of multiple definitions of the target condition, handling of multiple thresholds of test positivity, handling multiple index test readers, handling of indeterminate test results, grouping and comparing tests, and handling of different reference standards.

Examples

1. "For each index test, algorithm or checklist under consideration, we plotted estimates of sensitivity and specificity on coupled forest plots and in receiver operating characteristic (ROC) space. Where missing or indeterminate results were reported, study authors usually did not provide sufficient details to allow us to include these data in our analyses. Where study authors reported missing or indeterminate results in more detail, these results were excluded by us for consistency."[43]
2. "We included studies that defined macrosomia [target condition] using either birthweight >90th centile or >4000 g in the same meta-analysis because both are generally considered to be similar. However, we also performed subgroup analyses considering each definition independently."[44]
3. "The comparisons made in this review can be considered in a hierarchy. The highest level comparison groups tests by antibody type (HRP-2 versus pLDH) and is formed by combining the test types into two groups: HRP-2 antibody-based (Types 1, 2, 3 and 6) and pLDH antibody-based (Types 4 and 5). However, the data on each test type is classified in the primary studies according to commercial brands. In order to provide a coherent description of the studies contributing to each analysis, the results are structured first by grouping studies according to their commercial brand, then grouping brands to form test types, and finally grouping test types by antibody. The analytical strategy thus compared the test accuracy of commercial brands within each test type before making comparisons between test types, and then between antibodies. Comparative analyses first included all studies with relevant data, and were then restricted to studies that made direct comparisons between tests with the same participants, where such studies existed."[45]

Explanation

Choices made regarding handling of data (eg, how to combine results of tests with different positivity thresholds) in a DTA systematic review may be potential sources of bias and variability, as illustrated by the three examples. For instance, using the same data to select an optimal threshold for positivity and to estimate test accuracy, rather than estimating test accuracy at a threshold defined a priori, generally overestimates test accuracy.[30] To obtain clinically meaningful results from the narrative or statistical synthesis, factors such as multiple thresholds, multiple reference standards, multiple target conditions, and multiple index tests should be carefully considered during the review process, and where relevant, reported with clear justification for decisions made (examples 1-2).

For comparative DTA systematic reviews of multiple index tests, direct comparisons using head-to-head comparisons of multiple index tests are likely to have lower risk of bias and higher internal validity. However, such comparisons might not be feasible owing to the paucity of comparative DTA studies.[46] The alternative strategy of including all eligible studies (that is, indirect comparison) should acknowledge the potential for differences between test accuracy to be confounded by differences in study characteristics. As such, reporting whether direct or indirect comparisons were used in the review will allow readers to better consider the risk of bias when comparing the accuracy of multiple index tests (example 3).

Deleted items

Deleted PRISMA-DTA item 15: methods—risk of bias across studies

Specify any assessment of risk of bias that might affect the cumulative evidence (eg, publication bias, selective reporting within studies).

Deleted PRISMA-DTA item 22: results—risk of bias across studies

Present results of any assessment of risk of bias across studies.

Explanation

Empirical evidence indicates that publication bias exists for randomised trials, mainly driven by non-publication of statistically non-significant results.[47] Publication bias (box 1) or small study effects can be identified from funnel plots and tests assessing the association between effect estimates and their precision.[48] Hence, items for reporting investigations of the risk of bias across studies are included in PRISMA.

For DTA studies, although delayed and incomplete publication is likely, the determinants and magnitude of the bias resulting from the failure to report are unclear. Non-comparative DTA studies rarely test hypotheses or report P values,[49] and no simple driver for non-publication exists that is equivalent to statistical non-significance. Non-publication of findings is likely linked to study results, but definitions of low accuracy vary by test and context.

Studies of the link between observed accuracy and publication have produced mixed results.[50-53] The Deeks test for detecting publication bias can be used while standard tests such as the Egger test are not appropriate in the DTA context.[54 55] The Deeks test has low power to detect publication bias and small study effects.[54 56] For these reasons, statistical investigation of publication and reporting bias is not routinely recommended in DTA systematic reviews[57] and these items have been dropped for PRISMA-DTA. However, registration and availability of protocols for prospective DTA studies is encouraged,[34 58] and the review should report studies for which results are unavailable.

PRISMA-DTA item D2 (new item): meta-analysis

Report the statistical methods used for meta-analyses if performed.

Example

"For tests where commonly used thresholds were reported we estimated summary operating points (summary sensitivities and specificities), with 95% confidence and prediction regions using the bivariate hierarchical model. Where inadequate data were available for the model to converge, we simplified it, first by assuming no correlation between estimates of sensitivity and specificity and secondly by setting estimates of near zero variance terms to zero. Where all studies reported 100% sensitivity (or 100% specificity), we summed the number with disease (or no disease), across studies and used it to compute a binomial exact 95% confidence interval."[43]

Explanation

Multiple approaches to meta-analysis exist that have different limitations and can yield different estimates for the same statistic. Therefore, the model used for meta-analysis should be reported so that readers can consider whether the model selected was suitable.[59]

DTA systematic reviews will frequently perform meta-analysis to aggregate the available evidence into a summary measure of sensitivity and specificity, or to estimate an underlying summary receiver operating characteristic (ROC) curve. Key concepts of meta-analytical methods include appropriate modelling of within study uncertainty in estimates of sensitivity and specificity, simultaneous modelling of paired sensitivity and specificity statistics allowing for the likely negative correlation in estimates across studies, and estimation of unexplained variability (heterogeneity) in key parameters.[60-63]

Stating the name of a particular model (eg, the bivariate model or the hierarchical summary ROC model) with a reference is sufficient, because how such models deal with these concepts is well known. But if adaptations are made to a standard method (such as using a mixed, or a novel model), full descriptions and justification are needed.[18 59 64]

Additional considerations that might apply in some reviews include the structure of meta-regression models used to investigate sources of variability, methods for incorporating multiple thresholds from the same study, methods allowing for misclassification in the reference standard, and methods for explicit comparisons of the accuracy of multiple index tests.[18]

PRISMA-DTA item 16 (not modified from original PRISMA): additional analyses

Describe methods of additional analyses (eg, sensitivity or subgroup analyses, meta-regression), if done, indicating which were prespecified.

PRISMA-DTA item 17 (not modified from original PRISMA): study selection

Provide numbers of studies screened, assessed for eligibility, included in the review (and included in meta-analysis, if applicable) with reasons for exclusions at each stage, ideally with a flow diagram.

PRISMA-DTA item 18: study characteristics

For each included study, provide citations and present key characteristics, including participant characteristics (presentation, previous testing), clinical setting, study design, target condition definition, index test, reference standard, sample size, and funding sources.

Examples

Supplementary table 1 provides broad detail on the nature of the included studies.[24] Supplementary table 2 gives detail on the nature of the tests in each included study.[65]

Explanation

Diagnostic accuracy is not a fixed property of a test. A test's accuracy might vary between settings, patient populations, and findings on previous testing. Meta-analyses of DTA studies often show substantial heterogeneity between studies in sensitivity, specificity, or both. To assist interpretation and applicability of a systematic review's results, authors should provide sufficient details of the key study characteristics that might influence test accuracy.

The expected characteristics to be reported relate to elements captured in the review's objective because they might depend on previous evidence about sources of variability in accuracy and clinical reasons for false positive and false negative test results, in addition to characteristics considered when assessing the risk of bias and concerns about applicability in primary studies (see items 12 and 19). Describing the characteristics of primary studies is important because

it helps readers get a better sense of the variability of the studies included in the review.

Supplementary table 1 (and others) are useful to present key characteristics of the participants (presentation of the target condition, prior testing), setting (eg, general practice setting or hospital setting, single or multicentre settings), tests (technical descriptions of the index test(s), comparator test(s) and reference standard(s), including thresholds applied), severity of the target condition (eg, locally advanced breast cancer or metastatic breast cancer), study design, sampling methods (eg, consecutive or convenience), avoidance of inappropriate exclusions, blinding procedures, and verification procedures (eg, complete verification of test results *v* partial or random verification of a sample of test negatives, the time interval between execution of the index tests and reference standard, and whether all participants were included in the analyses).

The fraction of excluded participants and reasons for exclusion are of interest, to assess the risk of bias. Many of these items are also required reporting for risk-of-bias assessment (item 19); authors are encouraged to consider efficient presentation of these items to avoid redundancy in reporting. Languages of published papers, years of publication, and geographical origins of the included studies can be summarised. The funding sources should be stated here, in case of any association between sponsorship and estimates of DTA that might favour the interests of that sponsor.[66]

Authors are expected to transparently report the source of the included data and how it was accessed. For each included study, both published and unpublished, authors should provide a full citation for the source of their information. Unpublished reports could be posted on web repositories (eg, online conference abstracts book).

PRISMA-DTA item 19: risk of bias and applicability

Present evaluation of risk of bias and concerns regarding applicability for each study (this corresponds to item 12 regarding risk of bias and applicability methods).

Examples

The following examples illustrate three different strategies for presentation of results. Supplementary table 3 shows a tabular presentation of results,[67] and supplementary figure 1 provides a graphical presentation of results.[67] The following text gives a narrative summary of results:

"Risk of bias with respect to the index test was rated high in one study because it was not reported whether the radiologist who interpreted US (index test) was blinded to herniography (reference standard), which was performed by this same radiologist immediately after US. Risk of bias with respect to reference standard was rated high in all studies: in all studies, there was concern that the reference standard was not blinded to US findings, whereas in four studies there was also concern that the reference standard could not correctly classify the presence or absence of groin hernia. Risk of bias with respect to flow and timing was rated high in 13 studies because not all patients received the same reference standard (i.e., presence of verification bias) and/or not all patients were included in the analysis. Risk of bias with respect to flow and timing was rated unclear in two studies because time interval between US and reference standard was not reported. There were no applicability concerns."[67]

Explanation

Reviewers should report the results of their assessment of risk of bias and concerns regarding applicability. Authors should use graphical displays and figures such as those shown in examples 1 and 2 to summarise the results of the risk of bias and applicability assessments. These results provide an overview of the risk of bias and applicability across each domain and within individual studies, and should be supplemented by a narrative summary of the risk of bias and applicability assessment (example 3). As well as providing an overall summary of the risk of bias and applicability across studies, authors should highlight particular domains or signalling questions that were problematic in the included studies and highlight studies that were at high risk of bias or had concerns regarding applicability. Rather than simply specifying which domains were at high or unclear risk of bias, reviewers are encouraged to provide a more detailed explanation as to why they were judged at high or unclear risk of bias, and describe the methodological issues specific to the review topic that caused concern (example 3).

Results of the risk of bias and applicability assessment can be incorporated into the results of the review in various ways. These results range from a descriptive summary, supported by tables and graphs, to statistical incorporation as a means of investigating variability, such as stratifying the analysis according to risk of bias or applicability concerns, restricting inclusion into the review or primary analysis based on risk of bias or applicability concerns, or using covariates in meta-regression. Each reporting method can be done by considering overall study level ratings of bias or applicability, or by prespecifying individual domains or signalling questions considered particularly important to the review topic. Risk-of-bias evaluation in comparative accuracy studies remains a challenge, because the QUADAS-2 tool does not yet include criteria to assess studies comparing multiple index tests.[37]

PRISMA-DTA item 20: results of individual studies

For each analysis in each study (eg, unique combination of index test, reference standard, and positivity threshold), report 2×2 data (true positives, false positives, false negatives, true negatives) with estimates of diagnostic accuracy and confidence intervals, ideally with a forest plot or a ROC plot. Note that the original PRISMA-DTA publication used the term "curve" in item 20, but this is incorrect; the correct term is "plot."[10]

Examples

A variety of strategies that can be used to report the results of individual studies.

1. Supplementary figure 2 shows forest plots for detection of sputum eosinophils of 2% or more in adults.[21]
2. Supplementary figure 3 shows an summary ROC plot of magnetic resonance imaging, estimated fetal weight on two dimensional ultrasound using any Hadlock formula at threshold weight higher than the 90th centile or more than 4000 g), and abdominal circumference more than 35 cm for prediction of macrosomia.[44]
3. Supplementary figure 4 shows summary ROC plots of results for the bipolar spectrum diagnostic scale, hypomania checklist 32, and mood disorder questionnaire, for the detection of bipolar disorder in a mental health centre setting.[68]
4. Supplementary figure 5 shows a summary ROC plot of direct comparisons.[69]

Explanation

Systematic reviews collect the available evidence based on previously reported accuracy studies. Access to results from individual studies allows readers to examine the variability and distribution of test accuracy statistics across studies, inspect individual study features, verify meta-analysis results, and identify potential data extraction errors. Presentation of findings from individual studies allows interested readers to reproduce the analyses and also apply alternative methods (eg, direct pooling of predictive values).[70] Access to the 2×2 data for each study also allows additional analyses to be performed that are not specifically considered in the review, such as sensitivity analyses and explorations of variability (see item 23).

Essential data to report for each included study are complete 2×2 data (true positives, true negatives, false positives, and false negatives) and diagnostic accuracy statistics of interest with corresponding 95% confidence intervals; a table or forest plot might be an appropriate method for presentation.

A scatter plot of sensitivity versus specificity (summary ROC plot) provides an informative visual display that illustrates variability between studies in test accuracy. Use of colours and symbols allows comparisons between subgroups or test comparisons, as shown in supplementary figure 5.[69] In systematic reviews and meta-analyses that report results for multiple thresholds, presenting 2×2 data for all primary studies might not be feasible; in such cases, authors should consider reporting the complete 2×2 in appendices or supplementary materials.

Another useful method of displaying data in comparative accuracy systematic reviews is a coupled forest plot for sensitivity and specificity[57] or coupled summary ROC plot as in example 4.[71] Appropriate grouping and ordering of studies can enhance any plot. In supplementary figure 2, for example, studies in each subgroup are ordered by the threshold used to define test positivity in a forest plot.[21]

PRISMA-DTA item 21: synthesis of results

Describe test accuracy, including variability; if meta-analysis was done, include results and confidence intervals.

Examples

1. "Substantial heterogeneity was observed as shown by the extent of the 95% prediction region around the summary point on the summary receiver operating characteristic plot . . . The summary sensitivity and specificity of 2D ultrasound EFW were 0.56 (95% CI [confidence interval] 0.49–0.61) and 0.92 (95% CI 0.90–0.94), respectively."[44]
2. "Direct comparisons were based on few head-to-head studies. The ratios of diagnostic odds ratios (DORs) were 0.68 (95% CI 0.12 to 3.70; P = 0.56) for urea breath test-13C versus serology (seven studies), and 0.88 (95% CI 0.14 to 5.56; P = 0.84) for urea breath test-13C versus stool antigen test (seven studies). The 95% CIs of these estimates overlap with those of the ratios of DORs from the indirect comparison."[69]
3. "Sensitivities and specificities for differentiating FTD from non-FTD ranged from 0.73 to 1.00 and from 0.80 to 1.00, respectively, for the three multiple-headed camera studies. Sensitivities were lower for the two single-headed camera studies; one reported a sensitivity and specificity of 0.40 (95% confidence interval (CI) 0.05 to 0.85) and 0.95 (95% CI 0.90 to 0.98), respectively, and the other a sensitivity and specificity of 0.36 (95% CI 0.24 to 0.50) and 0.92 (95% CI 0.88 to 0.95), respectively."[72]

Explanation

The generation of summary estimates of the accuracy of a diagnostic test (ideally based on all applicable studies at low risk of bias) is one of the main objectives of DTA systematic reviews. A meta-analysis can produce these summary estimates, as means, variances, and their covariance. Estimates—especially those of the means—should always be accompanied by indicators of statistical imprecision, such as 95% confidence intervals.

Meta-analysis of DTA studies should ideally rely on random effects models, because variability between studies is often considerable and cannot be explained by chance only. In this case, only presenting summary sensitivity and summary specificity with confidence intervals can be misleading, because these confidence intervals do not reflect the variability between studies. Prediction intervals and regions can be used as statistics that indicate both the likely location of the summary accuracy statistics and the effects of variability between studies when enough studies are available, and the distributional assumptions are met.

An ROC plot with the individual study estimates can include summary ROC curves (supplementary figures 4-5)[68 69] or summary points with corresponding confidence and prediction regions, to visually illustrate

statistical uncertainty and variability (example 1). In addition, for test comparisons, relative or absolute differences can be presented along with confidence intervals and P values (example 2). When a meta-analysis is not possible, the range of results can be presented (example 3).

Methods for quantifying or describing heterogeneity in DTA systematic reviews used in intervention reviews cannot all be applied in DTA reviews. The I^2 statistic[73] is not informative for DTA systematic reviews because it does not account for potential correlation between sensitivity and specificity, for example, owing to threshold effects. Multivariate and DTA specific I^2 statistics have been proposed to quantify heterogeneity, but they are not well established.[71]

PRISMA-DTA item 23: additional analyses

Give results of additional analyses, if done (eg, sensitivity or subgroup analyses, meta-regression, analysis of index test failure rates, proportion of inconclusive results, and adverse events).

Examples

1. "A sensitivity analysis including only the five studies ... that used any Hadlock formula incorporating HC [head circumference], AC [abdominal circumference] and FL [femur length] to compute estimated fetal weight gave similar results to the analysis that included studies using any version of the Hadlock formula."[44]
2. "Subgroup analyses were conducted to investigate heterogeneity in sensitivity, and to a lesser degree, in specificity . . . Rapid influenza diagnostic tests showed a higher pooled sensitivity in children (66.6% [CI [confidence interval], 61.6% to 71.7%]) than in adults (53.9% [CI, 47.9% to 59.8%]) that was statistically significant (P<0.001), whereas specificities in the 2 groups were similar. The difference in pooled sensitivity between children and adults remained statistically significant when adjusted for brand of RIDT, specimen type, or reference standard."[74]
3. "The included studies reported an inconclusive result rate of 0.32–5.30%. This issue was further compounded by a myriad of varying quality control (QC) standards...Some studies investigated the reasons for their false and inconclusive results and reported these clearly, accounting for all samples. Other studies reported inconclusive results as false negatives or did not report them at all."[75]
4. "Serious adverse events from colonoscopy in asymptomatic persons included perforations (4/10000 procedures, 95% CI, 2-5 in 10000) and major bleeds (8/10000 procedures, 95%CI, 5-14 in 10 000)."[76]

Explanation

Sensitivity analyses are used to assess whether the results of the primary analysis are robust to changes in decisions regarding which studies and data are included in the meta-analysis, such as the impact of using more stringent inclusion criteria for the index test[44] or excluding studies at high or unclear risk of bias.[40] Not all sensitivity analyses can be prespecified because many issues only become apparent during the systematic review process, but authors should clarify which analyses were prespecified and which were not.[1]

Investigations of variability are often conducted using subgroup analyses and meta-regression. Subgroups are typically defined by study level characteristics (eg, clinical setting) with summary estimates of test accuracy computed for each subgroup.[77] Statistical comparisons can be made using meta-regression by including covariate(s) in models of test accuracy.[57 78]

In the example, subgroup analysis followed by a meta-regression identified differences in sensitivity, but not in specificity, between adults and children.[74] Prespecified analyses can be problematic or not feasible when the number of studies is small; any necessary simplifying assumptions should be described.[79] Individual participant data allow more refined stratification of patients and greater power to investigate heterogeneity, but only for characteristics that vary at the patient level.[77]

The presence and nature of inconclusive test results might be critical for assessing the usefulness of a test in practice. However, such information is often not reported or poorly described in primary studies,[80 81] and inconsistency in how such results are handled adds to apparent heterogeneity between studies.

Adverse events might occur as a result of the index test or reference test,[82] and could vary in severity from minor discomfort to life threatening complications.[76] The frequency and severity of adverse events might influence the clinical usefulness of a test and should therefore also be summarised and reported.

PRISMA-DTA item 24: summary

Summarise the main findings including the strength of the evidence.

Examples

1. Supplementary table 4 presents a summary of findings.[83]
2. "The principal findings of this systematic review were that the diagnostic accuracy of the three main groups of commercially available rapid diagnostic tests . . . for enteric fever . . . was moderate. There was no statistically significant difference in the average sensitivity between Typhidot, TUBEX, or Test-It Typhoid tests."[26]
3. "If the point estimates of the tests for S. haematobium are applied to hypothetical cohort of 1000 individuals suspected of having active S. haematobium infection, among whom 410 actually have the infection, the strip for microhaematuria would be expected to miss (102) and falsely identify (77) the least number of cases. This test would identify 384 positive cases in total."[84]

Explanation

The main findings of the review are typically summarised in the first part of the discussion section and might also be reported in a summary of findings table (example 1). Such structured tables are useful for summarising the main study objectives, setting, index tests, reference standards, key findings, and other information of relevance to readers. Relevant information encompasses the summary sensitivity and specificity from the primary analysis, which might be a comparison between index tests (example 2). The main findings should also cover any other objectives of the review.

Application of the summary estimates to a hypothetical cohort of patients, with a translation of the findings using absolute numbers, has been shown to help readers in understanding the findings (example 3).[85] This approach requires specification of a prevalence that would be used to re-express the sensitivity and specificity estimates in terms of predictive values, if required. Care incorporating uncertainty about the summary accuracy estimate arising from imprecision and heterogeneity would also need to be exercised.[86]

A tool for assessing the quality of the evidence and grading the strength of recommendations in health care was developed by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group.[87-89] However, application of the GRADE criteria to DTA systematic reviews is challenging, because a clear distinction is needed between patient important outcomes and test accuracy as the choice outcome.[90] Concrete guidance regarding translating the QUADAS-2 assessment to the corresponding GRADE criteria of indirectness and risk of bias could facilitate the use of the GRADE approach in DTA systematic reviews; however, this area remains a work in progress.[90]

PRISMA-DTA item 25: limitations

Discuss limitations from included studies (eg, risk of bias and concerns regarding applicability) and from the review process (eg, incomplete retrieval of identified research).

Examples

1. Risk of bias: "There was a high proportion of studies at high risk of bias and with high concern regarding applicability in all the four domains of the QUADAS-2 tool. This makes the validity and applicability of the results questionable."[69]
2. Applicability: "Furthermore, almost all in-person evaluations of dermoscopy used in conjunction with visual inspection had high concerns for the applicability of the included population and half had high concern for the applicability of the test. The restriction of including only excised lesions and the small number of studies conducted in a limited prior testing population mean that our results cannot be extrapolated to a primary care population."[43]
3. Review: "We were unable to perform all the investigations of heterogeneity that we had originally intended to because the data simply were not available."[91]

Explanation

The limitations section should include the validity of the findings (that is, risk of bias based on QUADAS-2), generalisability of the findings (that is, applicability based on QUADAS-2), and any limitations of the review process itself (eg, low number of included studies).

Incomplete reporting in primary studies could hamper interpretation of findings, and biases within the included publications (such as the reporting of accuracy results for only high-performing thresholds of continuous or ordinal tests) can distort meta-analytical results. Incomplete retrieval of relevant publications might also contribute to bias, if the omitted studies differ substantively from those included in the meta-analysis.

All threats to the validity and generalisability of the review should be discussed, with suggestions on how these factors could have influenced the reported synthesised results, including magnitude and direction of possible biases. Reviewers are encouraged to provide a more detailed explanation as to why certain domains were judged at high or unclear risk of bias, and to describe the methodological issues specific to the review topic that caused concern, rather than simply specifying which domains were at high or unclear risk of bias.

PRISMA-DTA item 26: conclusions

Provide a general interpretation of the results in the context of other evidence. Discuss implications for future research and clinical practice (eg, the intended use and clinical role of the index test).

Examples

1. "The most important conclusion from this review is that CEA [carcinoembryonic antigen] has inadequate sensitivity to be used as the sole method of detecting recurrence. Most national guidelines already recommend that it should be used in conjunction with another mode of diagnosis (such as CT [computed tomography] imaging of the thorax, abdomen, and pelvis at 12 to 18 months) to pick up the remaining cases. Our review supports this recommendation. If CEA is used as the sole triage test, a significant number of cases will be missed, whatever threshold is adopted for defining a positive test."[92]
2. "Future studies that evaluate the diagnostic accuracy of non-sputum-based tests for tuberculosis, such as LF-LAM [lateral flow urine lipoarabinomannan assay], in people living with HIV should use a reference standard that includes at least two different specimens (eg sputum, and urine), and in addition, for presumed extrapulmonary tuberculosis, appropriate specimens from the suspected sites of involvement."[93]

Explanation

The conclusions of a test accuracy systematic review should consider the results of the analyses, taking into account the intended use and clinical role of the index test in clinical practice, as well as limitations of the review, such as risk of bias and applicability concerns.[94]

In the discussion, authors should consider whether the index test is sufficiently accurate for the proposed role in the clinical pathway.[23] Conclusions will ideally reflect persistent uncertainty: Were the summary estimates after meta-analysis sufficiently precise? Were the included studies of sufficient quality? Could the results be applied to the clinical setting in which the test is likely to be used?[95]

Recent evidence suggests that systematic reviews of diagnostic accuracy studies often spin their results: authors, for example, arrive at strong recommendations regarding the use of a test in clinical practice despite having identified relatively low accuracy for the test under evaluation.[94] Such overinterpretation can be avoided by carefully taking into account the required accuracy for the destined role of the test in the clinical pathway.

Even if adequate accuracy of a test is demonstrated, the effectiveness (clinical utility) and cost effectiveness of the test when used in practice needs to be verified, and complementary non-accuracy evidence could already exist to answer these additional questions. Authors should note this condition particularly if they are making a strong recommendation for change to clinical practice.

PRISMA-DTA item 27 (not modified from original PRISMA): funding

For the systematic review, describe the sources of funding and other support and the role of the funders.

Additional considerations

The PRISMA-DTA reporting guideline is a minimum set of items to inform readers about the review process and its findings, and to enable quality appraisal and assessment of generalisability of the review findings.[10]

Although all DTA systematic reviews share basic methodological approaches, different subspecialties might have individual considerations to report. Therefore, authors are encouraged to include any additional information deemed necessary to allow readers to critically evaluate the findings and replicate the research. For example, interobserver variability is understood to be an important facet of imaging DTA research.[96] As such, reporting of statistics relevant to assessing this variability (eg, κ coefficients) could be relevant to imaging research.

DTA meta-analyses of ordinal index tests could have bias if included primary studies only report results from well performing thresholds, and if the thresholds reported differ across primary studies. This problem has been raised as a concern in mental health tests, and authors should report how they handle missing threshold data.[97]

With the growing evidence supporting the correlation between adherence to reporting guidelines and study quality, orchestrated strategies should be dedicated towards implementing PRISMA-DTA into research practices.[98] These approaches could be achieved on the journal level, by encouraging adoption of PRISMA-DTA and giving journal peer reviewers the option of using the PRISMA-DTA checklist as part of a manuscript peer review process, or on the author level, through organising workshops and raising awareness of PRISMA-DTA. Computerised analysis of manuscripts for compliance with PRISMA-DTA, as has been done for CONSORT (consolidated standards of reporting trials), would greatly decrease barriers to evaluating completeness of reporting.

With the increasing number of DTA systematic reviews, several emerging advances might be relevant to DTA systematic reviews. The implementation of machine learning in the identification of relevant DTA articles for inclusion in systematic reviews could increase efficiency, automate relatively daunting tasks, and yield a broader recall of identified articles.[99] However, the underlying algorithms of such processes are not yet fully understood. Whatever methods of article identification are used, readers will benefit from a complete description of the process. With the challenges arising from the poor reporting of artificial intelligence driven primary research in DTA,[100] the development of reporting guidelines specifying the minimum parameters to be reported for these algorithms could improve our understanding and allow for the use of their results in meta-research. Similar guidelines are available for individual participant data.[3 101]

Conclusion

This explanatory document aims to provide a resource for authors seeking guidance in what to include in a report of a DTA systematic review. We encourage authors to use this article when seeking a more comprehensive explanation of each item included in the PRISMA-DTA statement. We hope that these resources, along with the associated website (http://www.prisma-statement.org/Extensions/DTA), help improve the complete and transparent reporting of DTA systematic reviews.

Author affiliations

[1]Ottawa Hospital Research Institute, Clinical Epidemiology Program, Ottawa, ON, Canada

[2]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Centres, University Medical Centres, University of Amsterdam, Amsterdam, Netherlands

[3]University of Ottawa Department of Radiology, Ottawa, ON, Canada

[4]Lady Davis Institute of the Jewish General Hospital and Department of Psychiatry, McGill University, Montréal, QC, Canada

[5]Exeter Test Group, College of Medicine and Health, University of Exeter, Exeter, UK

[6]University of Sydney, Sydney, Australia

[7]Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[8]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK

[9]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Centers, University Medical Centres, University of Amsterdam, Amsterdam, Netherlands

[10]Department of Respiratory Medicine, Amsterdam University Medical Centres, University of Amsterdam, Amsterdam, Netherlands

[11]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[12]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Cochrane Netherlands, Utrecht, Netherlands

[13]Department of Paediatrics and Inserm UMR 1153 (Centre of Research in Epidemiology and Statistics), Necker-Enfants Malades Hospital, Assistance Publique-Hôpitaux de Paris, Paris Descartes University, Paris, France

[14]Institute of Social and Preventive Medicine, Berner Institut für Hausarztmedizin, University of Bern, Bern, Switzerland

[15]University of Birmingham, Birmingham, UK

[16]Brown University, Providence, RI, USA

[17]Lady Davis Institute of the Jewish General Hospital and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada

[18]Ottawa Hospital Research Institute Clinical Epidemiology Program (Centre for Journalology), Ottawa, ON, Canada

[19]Clinical Epidemiology Programme, Ottawa Hospital Research Institute, University of Ottawa, Ottawa, ON K1E 4M9, Canada

Contributors: All authors approve of the final submitted version. All authors meet ICMJE criteria for authorship. All authors met ICMJE requirements for authorship, and contributed to the conceptualisation, writing, editing, and approval process. MDFM is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: J-PS is supported by the Ontario Graduate Scholarship, and MDFM is supported by the University of Ottawa Department of Radiology Research Stipend Programme. MDFM is supported by the Canadian Institute for Health Research (grant No 375751), Canadian Agency for Drugs and Technologies in Health, and STAndards for Reporting of Diagnostic accuracy studies group (STARD). CJH is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula. BDT is supported by the Fonds de recherche du Québec-Santé researcher salary award. JJD is a United Kingdom NIHR Senior Investigator Emeritus, and is supported by the NIHR Birmingham Biomedical Research Centre. YT is funded by a UK NIHR postdoctoral fellowship, and is supported by the NIHR Birmingham Biomedical Research Centre. BHW is supported by a Medical Research Council Clinician Scientist Fellowship (grant No MR/N007999/1). BL is supported by a Canadian Institutes of Health Research Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award. None of the funding bodies listed had any role in the design of the document; management, preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The views expressed are those of the authors and not necessarily those of the UK NHS, NIHR, or the Department of Health and Social Care.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the Ontario Graduate Scholarship and the University of Ottawa Department of Radiology Research Stipend Programme, Canadian Institute for Health Research, Canadian Agency for Drugs and Technologies in Health, and STAndards for Reporting of Diagnostic accuracy studies group, National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula, Fonds de recherche du Québec-Santé, NIHR, and Medical Research Council for the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Higgins JPT, Thomas J, Chandler J, et al (eds). Cochrane Handbook for Systematic Reviews of Interventions version 6.0. 2019. https://training.cochrane.org/handbook.
2 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;45:281-90. doi:10.1002/jmri.25345
3 Alabousi M, Alabousi A, McGrath TA, et al. Epidemiology of systematic reviews in imaging journals: evaluation of publication trends and sustainability?*Eur Radiol* 2019;29:517-26. doi:10.1007/s00330-018-5567-z
4 Tunis AS, McInnes MD, Hanna R, Esmail K. Association of study quality with completeness of reporting: have completeness of reporting and quality of systematic reviews and meta-analyses in major radiology journals changed since publication of the PRISMA statement?*Radiology* 2013;269:413-26. doi:10.1148/radiol.13130273
5 Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. *BMC Med Res Methodol* 2011;11:27. doi:10.1186/1471-2288-11-27
6 Willis BH, Quigley M. The assessment of the quality of reporting of meta-analyses in diagnostic research: a systematic review. *BMC Med Res Methodol* 2011;11:163. doi:10.1186/1471-2288-11-163
7 Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q* 2016;94:485-514. doi:10.1111/1468-0009.12210
8 Page MJ, Shamseer L, Altman DG, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. *PLoS Med* 2016;13:e1002028. doi:10.1371/journal.pmed.1002028
9 Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383:267-76. doi:10.1016/S0140-6736(13)62228-X
10 McInnes MDF, Moher D, Thombs BD, et al, and the PRISMA-DTA Group. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA Statement. *JAMA* 2018;319:388-96. doi:10.1001/jama.2017.19163
11 McGrath TA, Alabousi M, Skidmore B, et al. Recommendations for reporting of systematic reviews and meta-analyses of diagnostic test accuracy: a systematic review. *Syst Rev* 2017;6:194. doi:10.1186/s13643-017-0590-8
12 Frank RA, Bossuyt PM, McInnes MDF. Systematic reviews and meta-analyses of diagnostic test accuracy: the PRISMA-DTA Statement. *Radiology* 2018;289:313-4. doi:10.1148/radiol.2018180850
13 Bossuyt PM, Reitsma JB, Bruns DE, et al, Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7-18. doi:10.1373/49.1.7
14 Altman DG, Schulz KF, Moher D, et al, CONSORT GROUP (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94. doi:10.7326/0003-4819-134-8-200104170-00012
15 Vandenbroucke JP, von Elm E, Altman DG, et al, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Int J Surg* 2014;12:1500-24. doi:10.1016/j.ijsu.2014.07.014
16 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700. doi:10.1136/bmj.b2700
17 Salameh JP, McInnes MDF, Moher D, et al. Completeness of reporting of systematic reviews of diagnostic test accuracy based on the PRISMA-DTA reporting guideline. *Clin Chem* 2019;65:291-301. doi:10.1373/clinchem.2018.292987
18 Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy version 1.0. 2010. https://methods.cochrane.org/sdt/.
19 Vroom AJ, Timmermans A, Bongers MY, van den Heuvel ER, Geomini PMAJ, van Hanegem N. Diagnostic accuracy of saline contrast sonohysterography in detecting endometrial polyps in women with postmenopausal bleeding: systematic review and meta-analysis. *Ultrasound Obstet Gynecol* 2019;54:28-34. doi:10.1002/uog.20229
20 Schieda N, McInnes MD, Cao L. Diagnostic accuracy of segmental enhancement inversion for diagnosis of renal oncocytoma at biphasic contrast enhanced CT: systematic review. *Eur Radiol* 2014;24:1421-9. doi:10.1007/s00330-014-3147-4
21 Korevaar DA, Westerhof GA, Wang J, et al. Diagnostic accuracy of minimally invasive markers for detection of airway eosinophilia in asthma: a systematic review and meta-analysis. *Lancet Respir Med* 2015;3:290-300. doi:10.1016/S2213-2600(15)00050-8
22 McGrath TA, Frank RA, Schieda N, et al. Diagnostic accuracy of dual-energy computed tomography (DECT) to differentiate uric acid from non-uric acid calculi: systematic review and meta-analysis. *Eur Radiol* 2020;30:2791-801. doi:10.1007/s00330-019-06559-0
23 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92. doi:10.1136/bmj.332.7549.1089

24 Zhelev Z, Hyde C, Youngman E, et al. Diagnostic accuracy of single baseline measurement of Elecsys Troponin T high-sensitive assay for diagnosis of acute myocardial infarction in emergency department: systematic review and meta-analysis. *BMJ* 2015;350:h15. doi:10.1136/bmj.h15
25 Wennmacker SZ, Lamberts MP, Di Martino M, Drenth JP, Gurusamy KS, van Laarhoven CJ. Transabdominal ultrasound and endoscopic ultrasound for diagnosis of gallbladder polyps. *Cochrane Database Syst Rev* 2018;8:CD012233. doi:10.1002/14651858.CD012233.pub2
26 Wijedoru L, Mallett S, Parry CM. Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever. *Cochrane Database Syst Rev* 2017;5:CD008892. doi:10.1002/14651858.CD008892.pub2
27 Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335-41. doi:10.1373/clinchem.2005.048595
28 Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76. doi:10.1503/cmaj.050090
29 Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6. doi:10.1001/jama.282.11.1061
30 Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working GCochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;149:889-97. doi:10.7326/0003-4819-149-12-200812160-00008
31 Dehmoobad Sharifabadi A, Leeflang M, Treanor L, et al. Comparative reviews of diagnostic test accuracy in imaging research: evaluation of current practices. *Eur Radiol* 2019;29:5386-94. doi:10.1007/s00330-019-06045-7
32 McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. *J Clin Epidemiol* 2016;75:40-6. doi:10.1016/j.jclinepi.2016.01.021
33 Salameh JP, McInnes MDF, McGrath TA, Salameh G, Schieda N. Diagnostic accuracy of dual-energy CT for evaluation of renal masses: systematic review and meta-analysis. *AJR Am J Roentgenol* 2019;212:W100-5. doi:10.2214/AJR.18.20527
34 Korevaar DA, Hooft L, Askie LM, et al. Facilitating prospective registration of diagnostic accuracy studies: a STARD initiative. *Clin Chem* 2017;63:1331-41. doi:10.1373/clinchem.2017.272765
35 Maynard-Smith L, Larke N, Peters JA, Lawn SD. Diagnostic accuracy of the Xpert MTB/RIF assay for extrapulmonary and pulmonary tuberculosis when testing non-respiratory samples: a systematic review. *BMC Infect Dis* 2014;14:709. doi:10.1186/s12879-014-0709-7
36 Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* 2013;66:1093-104. doi:10.1016/j.jclinepi.2013.05.014
37 Whiting PF, Rutjes AW, Westwood ME, et al, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009
38 Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, 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. doi:10.1186/1471-2288-3-25
39 Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054-60. doi:10.1001/jama.282.11.1054
40 Leeflang MM, Debets-Ossenkopp YJ, Wang J, et al. Galactomannan detection for invasive aspergillosis in immunocompromised patients. *Cochrane Database Syst Rev* 2015;(12):CD007394. doi:10.1002/14651858.CD007394.pub2
41 Van Hoving DJGR, Griesel R, Meintjes G, Takwoingi Y, Maartens G, Ochodo EA. Abdominal ultrasound for diagnosing abdominal tuberculosis or disseminated tuberculosis with abdominal involvement in HIV-positive individuals. *Cochrane Database Syst Rev* 2019;9:CD012777. doi:10.1002/14651858.CD012777.pub2
42 Bossuyt PM. Interpreting diagnostic test accuracy studies. *Semin Hematol* 2008;45:189-95. doi:10.1053/j.seminhematol.2008.04.001
43 Dinnes J, Deeks JJ, Chuchu N, et al, Cochrane Skin Cancer Diagnostic Test Accuracy Group. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database Syst Rev* 2018;12:CD011902. doi:10.1002/14651858.CD011902.pub2
44 Malin GL, Bugg GJ, Takwoingi Y, Thornton JG, Jones NW. Antenatal magnetic resonance imaging versus ultrasound for predicting neonatal macrosomia: a systematic review and meta-analysis. *BJOG* 2016;123:77-88. doi:10.1111/1471-0528.13517
45 Abba K, Kirkham AJ, Olliaro PL, et al. Rapid diagnostic tests for diagnosing uncomplicated non-falciparum or Plasmodium vivax malaria in endemic countries. *Cochrane Database Syst Rev* 2014;(12):CD011431. doi:10.1002/14651858.CD011431
46 Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013;158:544-54. doi:10.7326/0003-4819-158-7-201304020-00006
47 Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias G, Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One* 2013;8:e66844. doi:10.1371/journal.pone.0066844
48 Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002. doi:10.1136/bmj.d4002
49 Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology* 2013;267:581-8. doi:10.1148/radiol.12120527
50 Brazzelli M, Lewis SC, Deeks JJ, Sandercock PA. No evidence of bias in the process of publication of diagnostic accuracy studies in stroke submitted as abstracts. *J Clin Epidemiol* 2009;62:425-30. doi:10.1016/j.jclinepi.2008.06.018
51 Korevaar DA, Cohen JF, Spijker R, et al. Reported estimates of diagnostic accuracy in ophthalmology conference abstracts were not associated with full-text publication. *J Clin Epidemiol* 2016;79:96-103. doi:10.1016/j.jclinepi.2016.06.002
52 Korevaar DA, van Es N, Zwinderman AH, Cohen JF, Bossuyt PM. Time to publication among completed diagnostic accuracy studies: associated with reported accuracy estimates. *BMC Med Res Methodol* 2016;16:68. doi:10.1186/s12874-016-0177-4
53 Sharifabadi AD, Korevaar DA, McGrath TA, et al. Reporting bias in imaging: higher accuracy is linked to faster publication. *Eur Radiol* 2018;28:3632-9. doi:10.1007/s00330-018-5354-x
54 Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;58:882-93. doi:10.1016/j.jclinepi.2005.01.016
55 van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC Med Res Methodol* 2014;14:70. doi:10.1186/1471-2288-14-70
56 Begg CB. Systematic reviews of diagnostic accuracy studies require study by study examination: first for heterogeneity, and then for sources of heterogeneity. *J Clin Epidemiol* 2005;58:865-6. doi:10.1016/j.jclinepi.2005.03.006
57 Macaskill P, Gatsonis C, Deeks J, et al. Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy version 1.0* . https://methods.cochrane.org/sdt/resources-authors. 2010.
58 Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem* 2008;54:1101-3. doi:10.1373/clinchem.2008.108993
59 McGrath TA, McInnes MD, Korevaar DA, Bossuyt PM. Meta-Analyses of Diagnostic Accuracy in Imaging Journals: Analysis of Pooling Techniques and Their Effect on Summary Estimates of Diagnostic Accuracy. *Radiology* 2016;281:78-85. doi:10.1148/radiol.2016152229
60 Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982-90. doi:10.1016/j.jclinepi.2005.02.022
61 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865-84. doi:10.1002/sim.942
62 Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006;59:1331-2, author reply 1332-3. doi:10.1016/j.jclinepi.2006.06.011
63 Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol* 2016;16:97. doi:10.1186/s12874-016-0196-1
64 McGrath TA, Bossuyt PM, Cronin P, et al. Best practices for MRI systematic reviews and meta-analyses. *J Magn Reson Imaging* 2019;49:e51-64. doi:10.1002/jmri.26198
65 Pennant M, Takwoingi Y, Pennant L, et al. A systematic review of positron emission tomography (PET) and positron emission tomography/computed tomography (PET/CT) for the diagnosis of breast cancer recurrence. *Health Technol Assess* 2010;14:1-103. doi:10.3310/hta14500
66 Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. *BMJ* 2007;335:1202-5. doi:10.1136/bmj.39376.447211.BE

67 Kwee RM, Kwee TC. Ultrasonography in diagnosing clinically occult groin hernia: systematic review and meta-analysis. *Eur Radiol* 2018;28:4550-60. doi:10.1007/s00330-018-5489-9
68 Carvalho AF, Takwoingi Y, Sales PM, et al. Screening for bipolar spectrum disorders: A comprehensive meta-analysis of accuracy studies. *J Affect Disord* 2015;172:337-46. doi:10.1016/j.jad.2014.10.024
69 Best LM, Takwoingi Y, Siddique S, et al. Non-invasive diagnostic tests for *Helicobacter pylori* infection. *Cochrane Database Syst Rev* 2018;3:CD012080. doi:10.1002/14651858.CD012080.pub2
70 Leeflang MM, Deeks JJ, Rutjes AW, Reitsma JB, Bossuyt PM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol* 2012;65:1088-97. doi:10.1016/j.jclinepi.2012.03.006
71 Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med* 2014;33:2701-17. doi:10.1002/sim.6115
72 Archer HA, Smailagic N, John C, et al. Regional cerebral blood flow single photon emission computed tomography for detection of Frontotemporal dementia in people with suspected dementia. *Cochrane Database Syst Rev* 2015;(6):CD010896. doi:10.1002/14651858.CD010896.pub2
73 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60. doi:10.1136/bmj.327.7414.557
74 Chartrand C, Leeflang MM, Minion J, Brewer T, Pai M. Accuracy of rapid influenza diagnostic tests: a meta-analysis. *Ann Intern Med* 2012;156:500-11. doi:10.7326/0003-4819-156-7-201204030-00403
75 Mackie FL, Hemming K, Allen S, Morris RK, Kilby MD. The accuracy of cell-free fetal DNA-based non-invasive prenatal testing in singleton pregnancies: a systematic review and bivariate meta-analysis. *BJOG* 2017;124:32-46. doi:10.1111/1471-0528.14050
76 Lin JS, Piper MA, Perdue LA, et al. Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA* 2016;315:2576-94. doi:10.1001/jama.2016.3332
77 Reitsma JB, Moons KG, Bossuyt PM, Linnet K. Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers. *Clin Chem* 2012;58:1534-45. doi:10.1373/clinchem.2012.182568
78 Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. *Evid Based Ment Health* 2015;18:103-9. doi:10.1136/eb-2015-102228
79 Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res* 2017;26:1896-911. doi:10.1177/0962280215592269
80 Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. *BMJ* 2013;346:f2778. doi:10.1136/bmj.f2778
81 Korevaar DA, Wang J, van Enst WA, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology* 2015;274:781-9. doi:10.1148/radiol.14141160
82 Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799. doi:10.1136/bmjopen-2016-012799
83 Soares-Weiser K, Maayan N, Bergman H, et al. First rank symptoms for schizophrenia. *Cochrane Database Syst Rev* 2015;1:CD010653.
84 Ochodo EA, Gopalakrishna G, Spek B, et al. Circulating antigen tests and urine reagent strips for diagnosis of active schistosomiasis in endemic areas. *Cochrane Database Syst Rev* 2015;(3):CD009579. doi:10.1002/14651858.CD009579.pub2
85 Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. *Syst Rev* 2013;2:32. doi:10.1186/2046-4053-2-32
86 Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Interpreting results and drawing conclusions. In: Deeks J, Bossuyt P, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy version 0.9*. The Cochrane Collaboration, 2013: 24-5, https://methods.cochrane.org/sdt/2013.
87 Andrews J, Guyatt G, Oxman AD, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol* 2013;66:719-25. doi:10.1016/j.jclinepi.2012.03.013
88 Guyatt GH, Oxman AD, Vist GE, et al, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-6. doi:10.1136/bmj.39489.470347.AD
89 Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ, GRADE Working Group. What is "quality of evidence" and why is it important to clinicians?*BMJ* 2008;336:995-8. doi:10.1136/bmj.39490.551019.BE
90 Gopalakrishna G, Mustafa RA, Davenport C, et al. Applying grading of recommendations assessment, development and evaluation (GRADE) to diagnostic tests was challenging but doable. *J Clin Epidemiol* 2014;67:760-8. doi:10.1016/j.jclinepi.2014.01.006
91 Alldred SK, Takwoingi Y, Guo B, et al. First trimester ultrasound tests alone or in combination with first trimester serum tests for Down's syndrome screening. *Cochrane Database Syst Rev* 2017;3:CD012600. doi:10.1002/14651858.CD012600
92 Nicholson BD, Shinkins B, Pathiraja I, et al. Blood CEA levels for detecting recurrent colorectal cancer. *Cochrane Database Syst Rev* 2015;12:CD011134. doi:10.1002/14651858.CD011134.pub2
93 Shah M, Hanrahan C, Wang ZY, et al. Lateral flow urine lipoarabinomannan assay for detecting active tuberculosis in HIV-positive adults. *Cochrane Database Syst Rev* 2016;(5):CD011420. doi:10.1002/14651858.CD011420.pub2
94 McGrath TA, McInnes MDF, van Es N, Leeflang MMG, Korevaar DA, Bossuyt PMM. Overinterpretation of research findings: evidence of "spin" in systematic reviews of diagnostic accuracy studies. *Clin Chem* 2017;63:1353-62. doi:10.1373/clinchem.2017.271544
95 Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clin Chem* 2012;58:1636-43. doi:10.1373/clinchem.2012.182576
96 McGrath TA, McInnes MDF, Langer FW, Hong J, Korevaar DA, Bossuyt PMM. Treatment of multiple test readers in diagnostic accuracy systematic reviews-meta-analyses of imaging studies. *Eur J Radiol* 2017;93:59-64. doi:10.1016/j.ejrad.2017.05.032
97 Levis B, Benedetti A, Levis AW, et al. Selective cutoff reporting in studies of diagnostic test accuracy: a comparison of conventional and individual-patient-data meta-analyses of the Patient Health Questionnaire-9 Depression Screening Tool. *Am J Epidemiol* 2017;185:954-64. doi:10.1093/aje/kww191
98 van der Pol CB, McInnes MD, Petrcich W, Tunis AS, Hanna R. Is quality and completeness of reporting of systematic reviews and meta-analyses published in high impact radiology journals associated with citation rates?*PLoS One* 2015;10:e0119892. doi:10.1371/journal.pone.0119892
99 Bannach-Brown A, Przybyła P, Thomas J, et al. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst Rev* 2019;8:23. doi:10.1186/s13643-019-0942-7
100 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health* 2019;1:e271-e297. doi:10.1016/S2589-7500(19)30123-2.
101 Stewart LA, Clarke M, Rovers M, et al, PRISMA-IPD Development Group. Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;313:1657-65. doi:10.1001/jama.2015.3656

Web appendix: Supplementary material