

Title: Best practices for MRI Systematic Reviews and Meta-Analyses

Authors:

1. Trevor A McGrath BSc. The University of Ottawa Department of Radiology.
tmcgr043@uottawa.ca
2. Patrick M Bossuyt PhD, Department of Clinical Epidemiology, Biostatistics and Bioinformatics,
Academic Medical Center, Amsterdam. p.m.bossuyt@amc.nl
3. Paul Cronin MD MS. University of Michigan Department of Radiology.
pcronin@med.umich.edu
4. Jean-Paul Salameh BSc. The University of Ottawa Department of Clinical Epidemiology and
Public Health. The Ottawa Hospital Research Institute Clinical Epidemiology Program.
jsala016@uottawa.ca
5. Noémie Kraaijpoel, MD. Department of Vascular Medicine, Academic Medical Center,
Amsterdam, the Netherlands. n.kraaijpoel@amc.uva.nl
6. Nicola Schieda MD. The University of Ottawa Department of Radiology. The Ottawa Hospital
Research Institute Clinical Epidemiology Program. nschieda@toh.ca
7. Matthew DF McInnes MD. The University of Ottawa Department of Radiology. The Ottawa
Hospital Research Institute Clinical Epidemiology Program (Corresponding Author).
mcinnes.matt@gmail.com

Article Type: Review (Invited)

Disclosure: The authors confirm that this work has not been submitted not has it been published elsewhere. The authors have no relevant conflicts of interest to declare.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as doi: [10.1002/jmri.26198](https://doi.org/10.1002/jmri.26198)

Title: Best practices for MRI Systematic Reviews and Meta-Analyses

Abstract

As defined by the Cochrane Collaboration, a systematic review is a review of evidence with a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyze data from the studies that are included in the review. Meta-analysis is a statistical method to combine the results from primary studies that accounts for sample size and variability to provide a summary measure of the studied outcome. Systematic reviews of diagnostic test accuracy present unique methodological and reporting challenges not present in systematic reviews of interventions. This review provides guidance and further resources highlighting current best practices in methodology and reporting of systematic reviews of diagnostic test accuracy, with a specific focus on challenges and opportunities for MRI imaging.

Key Words: Systematic Review; Meta-Analysis; Diagnostic Test Accuracy

Introduction

Systematic reviews, when well reported and performed with methodological rigor, represent valuable summaries of existing evidence about risk factors for specific medical conditions, the effectiveness of interventions or the performance of medical tests (1-3). The Cochrane Collaboration defines a systematic review as “a review of the evidence on a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyze data from the studies that are included in the review” (1).

Five key characteristics of a systematic review are: (a) a clearly stated set of objectives with pre-defined eligibility criteria for studies; (b) an explicit, reproducible methodology; (c) a systematic search that attempts to identify all studies that would meet the eligibility criteria; (d) an assessment of the validity of the findings of the included studies, for example through the assessment of risk of bias; (e) a systematic presentation, and synthesis, of the characteristics and findings of the included studies (1). Frequently, systematic reviews will apply meta-analysis. Meta-analysis is a statistical method to combine the results from primary studies that accounts for sample size and variability to provide a

summary measure of the studied outcome (1). The term systematic review will be used henceforth to describe systematic reviews that may or may not contain a meta-analysis.

Systematic reviews can be used to aggregate data from multiple relatively small or underpowered studies to provide a more precise and more informative estimate of the effectiveness of an intervention or the accuracy of a diagnostic test (4,5). Additionally, meta-analysis may explore sources of heterogeneity among the observed effects of an intervention or the accuracy of a diagnostic test (2,3).

There are several fundamental differences between systematic reviews of randomized trials of interventions and reviews of test accuracy studies. When reviewing effectiveness studies, summary measures in meta-analysis usually estimate the benefit from treatment, in terms of, for example, a reduction in morbidity or mortality. In systematic reviews of test accuracy, the summary measure represents the ability of an index test to detect the target condition: a disease, or disease stage. This is typically expressed in terms of the index test's sensitivity and specificity.

Magnetic resonance imaging (MRI) is a powerful diagnostic tool and with a wide array of clinical, mainly diagnostic, applications (2-10). While MRI guided procedures (interventional MRI), began in the 1990s, image-guided interventions are still dominated by ultrasound, fluoroscopy and computed tomography (CT) (11). There are several challenges to MR-guided interventions, such as longer acquisition times, increased cost of MRI safe equipment, and the technical complexity of performing a procedure within a superconducting MRI system. These challenges presently limit their widespread use (11). Due to these limitations, the overwhelming majority of systematic reviews within MRI have focused on diagnostic test accuracy.

In this paper, we discuss specific features of systematic reviews of MR studies, with focus on diagnostic test accuracy systematic review methods.

Author Resources

Published systematic reviews of test accuracy studies have been shown to be often of questionable quality and prone to shortcomings in methodology and reporting (12-15). Several resources exist for authors for both methodological guidance and reporting best practices. The [Cochrane Handbook for DTA \(Diagnostic Test Accuracy\) Reviews](#) (freely available on-line) provides methodological guidance in the areas of protocol development, developing inclusion criteria, execution of a search strategy, assessment of methodological quality of included studies, analyzing and presenting results along with the interpretation of results (16). For reporting of DTA systematic reviews, [PRISMA-DTA](#) and PRISMA-DTA for abstracts should be used. However, if a systematic review is being performed for an MRI guided intervention, it would be more appropriate to use the Cochrane Handbook for Systematic Reviews of Interventions to guide methodology, and the original PRISMA statement to guide reporting (1,17).

Forming the clinical question

A critical step in planning a systematic review is forming a relevant clinical question. In order for a systematic review to advance knowledge, the question should be one for which there is clinical equipoise. As such, if there are multiple well-executed randomized trials, or several well-executed (low risk of bias) diagnostic accuracy studies on a topic with similar conclusions, there may be little to be

gained from performing a systematic review. Similarly, if a high quality systematic review has recently been published on the topic (or a protocol is registered and one is underway), duplicating these efforts may not be contributory (12). The ideal scenario is one in which: a) there is clinical equipoise b) there are a large number of well designed, well reported studies with varying conclusions c) there is no recent systematic review on the topic (or one underway). These conditions provide not only the relevant clinical question, but also the substrate with which to answer it. Identifying a relevant question can be challenging, and requires consideration of the present body of evidence, including primary research, as well as clinical practice guidelines. Often, the 'future research' identified in the discussion section of studies, or 'uncertainty- low level evidence' cited in guidelines can be helpful indices regarding topics that are ripe for systematic review.

Once a question is identified, a detailed, structured format allows readers to best understand the question. The 'PICO' (Patient, Intervention, Comparator, Outcome) format for systematic reviews of interventions, or 'Patients-Index Test-Target Condition' for diagnostic accuracy systematic reviews provides a useful framework with which to structure review questions (1,18-20). Each category of the 'PIT' question should contain sufficient detail such that the readers can interpret whether the study findings are generalizable to their practice.

Consider a systematic review aiming to evaluate the diagnostic accuracy of MRI in detecting extra-prostatic extension (EPE) in patients with prostate cancer. For the 'Patient' category relevant details might include: species (human); age (adult); gender (male); prior testing (prostate cancer detected by trans-rectal biopsy with Gleason score of ≥ 7) (21). For the 'Index Test' category, the general principle is to provide enough detail such that the reader could replicate the test in clinical practice (22). Relevant 'index test' details for the prostate study might include, but not be limited to: field strength

(e.g. 1.5 versus 3T), contrast agent used (e.g. dynamic contrast enhancement performed with extracellular contrast agent with temporal resolution of at least 10 seconds), slice thickness for T2 WI (≤ 4 mm), use of endo-rectal coil, use of diffusion weighed imaging (DWI), and which b-values used (23,24). For the ‘target condition’ category, readers need to be aware whether the disease, or target condition is similar to that encountered in their patient population. As such, for the prostate example, the ‘target condition’ is prostate cancer with extension beyond the capsule of the prostate detected at histopathology (25).

Part of defining a clinical question relates to deciding what the main outcome of the systematic review will be. There are two major categories of diagnostic accuracy systematic reviews—those evaluating a single index test, and those comparing multiple_index tests. Evaluating accuracy of a single test can aim to: a) gain a greater understanding of test precision (via more narrow confidence intervals) b) to identify sources for variability in accuracy. For evaluations of comparative accuracy, it is important to consider tests that might be used at a similar point in the clinical pathway (26). For example, it may not be optimal to compare prostate specific antigen (PSA) with MRI for screening of prostate cancer when each might be positioned at separate places in the diagnostic pathway—PSA being a triage test, with MR being a confirmatory test. Rule-out triage tests generally value sensitivity over specificity, and confirmatory tests require greater specificity. As such, head to head comparison of PSA to MRI would not be relevant since different components of accuracy are emphasized, and the intended use may be different (21).

Updating Systematic Reviews

Updating systematic reviews is an essential part of keeping knowledge synthesis contemporary. Deciding when to update a systematic review can be challenging and depends on the nature and extent of research done since a systematic review was completed (1). For MRI systematic reviews, indicators that an update may be necessary are typically linked to advances in technical parameters. For example, if a systematic review was performed on studies with lower field strength, or without the benefit of multi-parametric acquisition, and subsequent studies have demonstrated substantially different data from those of the original systematic review, an update may be fruitful.

Search and inclusion

Best practices for searching for MRI systematic reviews have much in common with systematic reviews in general. Searching multiple databases (e.g. MEDLINE, Embase, and Google Scholar), and searching the reference lists of included studies are recommended. Collaboration with an experience librarian to ensure that search terms are constructed based on the index test and target condition is recommended.

Use of methodologic search filters for diagnostic accuracy studies is not recommended since they have been shown to miss important studies (16,27). The need to search for unpublished data is controversial and may not always be necessary in the case of well-studied, well-established MRI techniques. However, in cases where a technique is new, or rapidly evolving, search of relevant conference abstracts (e.g. ISMRM, RSNA), and clinical trial databases may be appropriate in order to ensure that all relevant studies are included. Inclusion of studies in languages other than English can be a very labor-intensive process re: translation. As such, the risk of missing relevant studies if English is

used as a search limitation should be considered (28). For topics with a very regional focus (e.g. diseases endemic to certain parts of the world) it may be necessary to ensure that studies from these regions are included.

Eligibility of studies retrieved during the search process should be guided by the ‘PIT’ question. The inclusion criteria should ensure that the appropriate patients, index test and target condition are being evaluated. Returning to the prostate MRI example used above, the following inclusion criteria based on the ‘PIT’ question could be applied:

To be eligible for inclusion, studies must fulfill all of the following criteria:

1. Human male patients with biopsy proven Gleason ≥ 7 prostate cancer.
2. Index test applied was MRI with 3T magnet, dynamic contrast enhancement of ≤ 10 s temporal resolution, ≥ 2 planes of T2WI with ≤ 4 mm slice thickness, application of DWI using ≥ 3 b-values.
3. Target condition is extra-prostatic extension of tumor identified by surgical pathology following prostatectomy.

Additional criteria for inclusion can go beyond the ‘PIT’ question and consider aspects such as study quality (e.g. prospective design, consecutive or random selection of patients, and date of publication).

Authors need to be cognizant of striking a balance between inclusion criteria that are too broad vs. strict. Criteria that are too broad may include studies that are so heterogeneous that meaningful comparison is not possible. In contrast, criteria that are too narrow risks limiting the pool of potential studies, thereby restricting opportunity to evaluate for sources of variability (29).

Data extraction

Ideally data are extracted from primary studies using data extraction forms defined and piloted throughout the development of the study protocol. It is recommended to perform data extraction independently by two or more extractors with disagreements reconciled by mutual agreement or through discussion with a third reviewer to reach a final decision. Data extraction forms should be piloted on a sample of primary studies and refined as necessary. This is intended to assess both completeness of the forms, ensure clarity of user instructions and optimize inter-extractor consistency. The collected data can then be compared across extractors to assess inter-extractor consistency (1).

Identification of relevant data items, methods of the data extraction (e.g. independently or in duplicate), along with definitions of the information to be extracted and the processes for obtaining missing data from the investigators of the eligible studies should all be specified. Characteristics of the participants, clinical setting, study design, and classification of the target conditions, index tests, and reference standards are crucial for the assessment of test accuracy and possible sources of heterogeneity (19).

Quality assessment

Interpretation of systematic reviews is largely dependent on the quality of the included studies. Although quality assessment is essential in any systematic review, diagnostic accuracy systematic reviews require a specific tool, as study design of test accuracy research differs from interventional studies. The QUADAS-2 tool is the recommended tool for systematic quality assessment of diagnostic accuracy studies (30,31). QUADAS-2 consists of four domains: (1) patient selection, (2) index test, (3) reference standard, and (4) flow and timing. All domains are evaluated for risk of bias, and the first three

domains are also assessed for concerns regarding applicability. Risk of bias refers to flaws or limitations in the design or conduct of a study. Concerns with regard to applicability refer to differences between the clinical features of the included study compared with the review question, including patient characteristics, setting, definitions of the target condition, and application or interpretation of the index test. For example, in a recent review of DWI to assess treatment response in locally advanced uterine cervical cancer, the authors found the spectrum of patients included in three of nine studies was not representative of the patients who would receive the test in clinical practice (greater than 75% of patients with stage IIb or higher disease) (32).

As each review question may require different approaches for quality assessment, it is important to tailor the QUADAS-2 tool by adding or omitting signalling questions. Review authors should consider omitting any item that does not apply to the review question. For example, in imaging studies, if the index test is not interpreted based on a specific threshold, it may not be worthwhile considering this particular issue. The apparent diffusion coefficient (ADC) threshold would be relevant in DWI to indicate the presence or absence or degree of restricted diffusion and thus this signalling question would be relevant, however in detection of labrum tears threshold effects may not be relevant (5). Upon agreement of the reviewers with regard to the content of the tool, two independent investigators should perform a quality assessment pilot process.

Preferably, quality assessment should be performed by at least two independent authors. Risk of bias can be designated as either 'low', if all signaling questions within the same domain are answered with 'yes'; 'high', if any of the signaling question is answered with 'no'; or 'unclear', if risk of bias assessment is hampered by a lack of reported data (22). Concerns with regard to applicability are not

based on signaling questions, but represent an overall judgment for a specific domain, rated as ‘low’, ‘high’, or ‘unclear’.

It is not recommended to calculate an overall quality score based on the QUADAS-2 results, as the relevance of different sources of bias and concerns of applicability may differ between review questions (33). In general, a study may be regarded as having a low risk of bias or low concern of applicability when judged as ‘low’ on all risk of bias and applicability domains, respectively. If one of the domains was designated ‘high’ or ‘unclear’, the study may be judged to have risk of bias or as having concerns regarding applicability.

The results of the QUADAS-2 assessment are typically presented in a tabular fashion, listing the results of the individual studies for each domain. To better guide the reader, it is key to also provide for a narrative summary of the quality assessment, explaining how the quality of the included studies may affect the overall interpretation of the review results.

Authors may only include studies at low risk of bias without concerns regarding applicability for the primary analysis of the review. Subgroup or sensitivity analyses may be used to explore whether the diagnostic accuracy estimates vary across studies judged as ‘low’, ‘high’, or ‘unclear’ for all or for separate domains.

Data analysis

In order to ensure reproducibility of a review and transparency of reporting, any decisions made with respect to data handling need to be reported (19). How studies were grouped for meta-analysis should be reported (e.g. whether included studies were stratified by field strength and/or sequences used)

along with how primary study level data was handled (e.g. multiple thresholds or multiple index test readers).

In most meta-analyses of diagnostic test accuracy, data are dichotomized to “disease present” and “disease absent” to produce a single two by two table from each study. However, there are circumstances when a test report results are on a continuous scale, i.e. ADC map values, a cut point or threshold can be chosen above which the test results are “positive” and below which test results are classed as “negative”. For example, in ADC maps values less than 1.0 to 1.1×10^{-3} mm/s (or 1000 - 1100×10^{-6} mm²/s) are often selected to indicate restriction in adults. However, this can be problematic if different studies report different thresholds, i.e. studies report test performance at multiple thresholds (34). In this scenario, one option is to pool all studies regardless of threshold and perform subgroup-analysis or meta-regression to determine if threshold is a contributor to variability in accuracy. Alternatively, Riley *et al.* have extended the bivariate-normal meta-analysis model first proposed by Reitsma *et al.* (35). Their model accounts for within-study correlations in the sensitivities and specificities at various thresholds (34). In addition, their model allows for relationships between test performance metrics at the between-study level (34).

It has been shown that systematic reviews of test accuracy studies in imaging journals infrequently report how they handle primary studies with multiple readers (36). This deficiency makes it difficult to determine exactly how primary study data was input into the meta-analysis and thus difficult to reproduce results. Optimal methods for handling multiple reader data are currently not available but multi-level hierarchical models accounting for between-observer variability within studies, and between-study variability, provided multiple reader data is reported consistently at the primary study level are

needed. (36) Using such models, all readers would be included in the meta-analysis, inter-observer variability at the primary study level would not be lost, and a single study would not be over-represented (36). Until such optimal methods become available, authors are encouraged to report how they handle multiple reader data for primary studies in their meta-analysis.

The statistical model and software package used for meta-analysis should be explicitly reported (19). Unlike a meta-analysis of randomized trials, which typically produces a single summary effect measure, meta-analyses of diagnostic test accuracy studies typically produce two summary measures, such as sensitivity and specificity. These two summary measures are correlated and statistical methods used for meta-analysis need to account for this. Due to this unique challenge in diagnostic test accuracy meta-analyses, hierarchical methods, which account for this correlation, have been developed (35,37). In a comparison between traditional univariate methods and the recommended hierarchical methods, the univariate methods were found to overestimate diagnostic accuracy and provide narrower confidence intervals compared to hierarchical methods (13). A recent review by Cronin *et al.* provides a useful appendix describing software for diagnostic test accuracy meta-analysis (38).

Assessing variability (heterogeneity)

Students typically learn that the negative and positive predictive value of a test vary, depending on the prevalence of the target condition in those being tested, but that sensitivity and specificity are stable. The diagnostic accuracy is not a fixed property of a test, as the performance of test varies. Sensitivity and specificity differ, sometimes dramatically, depending on where the test is used: the setting (primary care or tertiary care, for example), the type of patients (young or old, obese or thin), or

whether they have had prior tests (39). These characteristics will also vary depending on the definition of the target condition, and may vary with the type of clinical reference standard researchers rely on.

This variability poses two types of challenges in systematic reviews: how best to express the variability, and how best to explain the variability. To express the variability, a meta-analysis of diagnostic accuracy studies will almost always rely on a random-effects model. Such a random-effects model assumes that the sensitivity and specificity vary between studies, not just by chance, but in a more systematic way, where each application has its own sensitivity and specificity. A random-effects does not try to estimate that setting-specific sensitivity and specificity, but aims to describe it in terms of a distribution, with a mean and a variance. In fact, there are two, correlated distributions: one for the sensitivity and one for the specificity. The magnitude of the respective variances is an expression of the variability in accuracy.

For explaining and handling variability, several approaches exist. One approach is to limit the variability by having a narrow review question. Rather than having a review question that focuses on a general target condition, such as “detecting metastases” one could add a specification of the type of patients (e.g. patients with colorectal cancer), the setting (e.g. tertiary care cancer center), and maybe even the specific type of MRI (e.g. MR liver using a hepatobiliary contrast agent) that one wants to evaluate in the review.

A second approach is to group included studies in such a way that they highlight likely sources of between-study variability beyond chance. Methodological guidance can be found in the Cochrane Handbook for DTA Reviews, specifically in Chapter 10 (16). One method is to perform a subgroup analysis: performing a meta-analysis of the results from on a selected group of studies, once again

defined by a narrower definition of the type of test, patients, setting, target condition, or a combination thereof. An example of a recently published subgroup analysis is presented in **Figure 1** (3). An alternative is meta-regression. Here a single meta-analysis is performed, based on all included studies, but indicator variables are used to mark studies that differ in an identifiable way. By including these variables in the random-effects model, one can estimate systematic differences in the mean, variance and covariance between studies. An example of a recently published meta-regression from a review of MRI in paediatric patients with suspected appendicitis is presented in **Figure 2** (4). This meta-regression showed that the addition of contrast has no impact on diagnostic accuracy and the use of DWI was detrimental to diagnostic accuracy, perhaps guiding future selection of MRI protocols for this target condition.

There are several potentially serious limitations to subgroup analysis and meta-regression for exploring sources of variability. In many systematic reviews, the number of studies is limited, and any results from such additional analyses may be imprecise. Second, studies may differ in multiple ways, and it can be difficult to pinpoint the most likely source of variability. Third, these approaches only work for between-study differences. Within study sources of variability cannot easily be incorporated.

Results from studies can also differ in a systematic way when some studies were designed or conducted with shortcomings, whereas others were not. In that case, studies with methodological deficiencies may yield biased results. Bias should not be a reflection of variability in accuracy, and the best way of handling risk of bias is probably to identify such studies using QUADAS-2 and to perform meta-analysis only on studies that are at low risk of bias (29).

Presenting results

Recently an extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) has been developed specific to diagnostic test accuracy: PRISMA-DTA (19). [PRISMA-DTA and PRISMA-DTA for abstracts](#) user-friendly checklists for authors and reviewers are freely available on-line from the EQUATOR network website (40). This reporting guideline was created to account for the challenges posed by systematic reviews of diagnostic test accuracy studies. Such reviews should be using the PRISMA-DTA checklist as their reporting guideline rather than the original PRISMA statement (17,19).

The number of search results screened, assessed for eligibility at the full-text level, excluded at full-text review (with reasons), and included in the systematic review and meta-analysis should be included. This information should be displayed in a flow diagram. A template is provided in **Figure 3** (17).

Results from the risk of bias and concerns regarding applicability assessment (e.g. QUADAS-2) should be presented granularly for each domain for each study, either in the full-text of the manuscript or online as supplemental information (30). An example is provided in **Figure 4**. Plots showing the proportion of studies in low, unclear, and high risk of bias categories per domain are less informative. An overall 'quality score' should not be used (33).

For each included primary study information should be provided: a) participant characteristics (presentation, previous testing); b) clinical setting; c) study design; d) target condition definition; e) index test(s); f) reference standard(s); g) sample size; h) funding sources (19). This provides a clear summary of the key characteristics included primary studies. For each analysis in each primary study

(e.g. unique combination of index test, reference standard and positivity threshold) 2x2 data should be provided along with estimates of diagnostic accuracy (e.g. sensitivity and specificity) along with associated confidence intervals. The estimates of diagnostic accuracy and confidence intervals should be displayed graphically, as points in forest plots or in receiver operating characteristic (ROC) space (19). An example is shown in **Figure 5** (41). Displaying the accuracy estimates with confidence intervals for each primary study in a forest plot allows for a quick visual assessment of heterogeneity. From the data presented in **Figure 5** it is clear the primary studies are quite homogeneous and it is unsurprising that meta-regression found no significant differences in accuracy among several chosen covariates.

Results from each meta-analysis including, at minimum, summary estimates of diagnostic accuracy and associated confidence intervals should be presented. This data can be presented on a forest plot or an HSROC curve. The forest plot (**Figure 5**), familiar to many readers, will display the sensitivity and specificity separately, although we know they are correlated accuracy measures. The HSROC plot will display the summary estimate and confidence region in ROC space, showing sensitivity and specificity together in two-dimensional space.

If using the bivariate random-effects or HSROC models, parameter estimates can be used to calculate and plot the summary ROC curve, the summary operating point (the summary estimates of sensitivity and specificity in ROC space), a 95% confidence region around the summary point, and a 95% prediction region (16). The 95% confidence region in ROC space expresses the uncertainty about the location of the mean diagnostic performance will fall. The 95% prediction region in ROC space is the area within which 95% of all estimates from primary study diagnostic accuracies can be expected

(16,19). An example of a HSROC curve displaying the summary estimate, along with 95% confidence and prediction regions is shown in **Figure 6** (2).

The Cochrane Handbook for DTA reviews recommends providing a summary of findings table. (16) This is a concise method to summarize the findings of a review. The table should include participants, clinical setting, index test(s), reference standard(s) and role in the clinical pathway. Any concerns arising from the assessment of risk of bias and applicability or from excessive heterogeneity should be noted within the table. For each unique combination of index test or positivity threshold a unique row should include the number of studies and participants, estimates of diagnostic accuracy generated by the review along with associated uncertainty estimates (e.g. confidence intervals) and information regarding disease prevalence from either the studies in the review or external sources (16). A sample summary of findings table is shown in **Figure 7** (42).

In comparative diagnostic accuracy reviews, comparing two or more index tests against a reference standard, the table should include the number of studies and participants arising from direct and indirect comparisons, estimates of diagnostic accuracy for each test with associated uncertainty estimates, and P-values for the comparison of the index tests being compared. This will allow to reader to determine significant differences in test accuracy (16).

Discussion and conclusions

Any systematic review should provide a balanced discussion and conclusions. The main findings of the review should be summarized including the diagnostic accuracy of the reviewed index test(s) and factors affecting variability in test accuracy. Limitations of the review should be discussed both from the

perspective of included studies (e.g. risk of bias and concerns regarding applicability) and the review process (e.g. incomplete retrieval of identified research, limited search strategy).

Authors should be cautioned only to form conclusions that are justified by the results of the systematic review. It has been shown that overinterpretation (often referred to as “spin”) of diagnostic test accuracy systematic review results is problematic and not infrequent (14). Caution should be exercised when stating a positive conclusion about the index test reviewed in the presence of concerns regarding risk of bias (and/or applicability) or variable test performance (heterogeneity); these limitations of the evidence should be explicitly mentioned, both in the abstract and the full text. In the setting of a comparative systematic review, test superiority should only be claimed when results of statistical comparison (e.g. meta-regression) identify that the index test accuracy differs significantly.

Future directions for imaging systematic reviews

Currently, most diagnostic accuracy studies estimate the accuracy of a single study compared to a reference standard. However, what is required are studies that compare the accuracy of competing tests, performed in all study participants, against the reference standard (43). Such studies allow strong inferences about whether one test is as accurate as or more accurate than other tests. This then enables clear recommendations to be made about test selection and consideration of the tests suitable for use (43). Because of the limitations of the primary literature, most systematic reviews and meta-analyses of diagnostic test accuracy estimate only the accuracy of an individual test (43). Reviews of comparative diagnostic accuracy would allow comparisons between tests which can then inform decision-making

(43). In addition, more studies that evaluate tests in well-defined clinical diagnostic pathways are also needed (43).

Diagnostic Test Accuracy Network Meta-analysis

Network meta-analysis (NMA) allows an indirect comparison between two treatments or tests of interest obtained through more than one common comparator. Recently, a method for quantitatively addressing both direct and indirect comparisons of several competing interventions has been developed by Lu and Ades (44). This has further improved NMA techniques, with the advantages of strengthening inference of the relative efficacy of two treatments (or accuracy of two tests), by including both “direct” and “indirect” comparisons and facilitating simultaneous inference regarding all treatments or tests, in order, for example, to select the best treatment or test, i.e. a ranking of the treatments or tests (44).

Prior to 2008, very few systematic reviews containing NMAs were published (45). The hypothetical example shown in **Figure 8A**, shows a review with multiple direct comparison meta-analyses (i.e. an umbrella review), whereas **Figure 8B** shows a network plot. NMA can be performed within either a frequentist or a Bayesian framework (45). Bayesian analyses are performed with Markov Chain Monte Carlo (MCMC) simulations. This allows repeated reproduction of the model until convergence. The Bayesian approach has several advantages. These include a straightforward way of making predictions, and the possibility of incorporating different sources of uncertainty. In addition, Bayesian analyses are more flexible statistical models, and are probably more applicable to NMA of diagnostic imaging studies (45).

Network meta-analyses are increasingly popular in comparative effectiveness research. However, they can be difficult to understand and interpret (46). Graphical tools can present results of statistical

analyses in a way that is more easily understood (46). These include network plots as described below and in **Figure 8B** (46). In a network plot, the nodes (circles) represent the interventions or technologies under evaluation (45). The lines that connect the interventions represent the comparisons. The set of direct and indirect statistical comparisons is the NMA (45). Node size is dependant on the number of studies for each intervention (45). Line width is dependant on the number of direct evidence studies (45). A contribution plot is used to show the influence of each direct piece of evidence. The size of each square is proportional to the weight attached to each direct summary effect, usually shown on the horizontal axis, for the estimation of each network summary effects, usually shown on the vertical axis (46). In an inconsistency plot, each closed loop in the network is assessed. Triangular networks formed by three treatments/technologies all compared with each other are assessed (46). An inconsistency factor (IF) is calculated with a 95% confidence interval. This is the absolute difference between direct and indirect estimates (46). A z-test for the IF can also be calculated. The IF is the logarithm of the ratio of two odds ratios (RoR) from direct and indirect evidence in the loop (46). Values close to 1 for the RoR mean that the two sources are in agreement (46). The comparison-adjusted funnel plot is similar to the funnel plot that assesses the presence of small-study effects in a meta-analysis. The 'comparison-adjusted' funnel plot presents the difference between the study-specific effect sizes from the corresponding comparison-specific summary effect (46). Heterogeneity within a NWM can be visually displayed using a predictive intervals plot. Instead a forest plot of the estimated summary effects along with their confidence intervals and their corresponding predictive intervals (PrI) for all comparisons can be displayed in one plot. This forest plot summarizes the relative mean effects, predictions and the impact of heterogeneity on each comparison (46). Visual tools can also be used to rank interventions. The ranking of the treatments or tests should be done using probabilistic methods, for example using the

surface under the cumulative ranking curve (SUCRA). These methods take into account the estimated effect sizes and their accompanying uncertainty (47). The SUCRA is used to provide a hierarchy of the interventions (46). The larger the SUCRA value, the better the rank of the intervention (46).

Multidimensional scaling (MDS) is an alternative approach to rank the competing interventions (46). A clustered ranking plot of the interventions in a network based on cluster analysis of SUCRA values for two different outcomes such as efficacy and acceptability can also be performed (46).

An example of a diagnostic test accuracy network meta-analysis to compare the diagnostic value of four imaging methods (MRI, positron emission tomography (PET), CT, and DWI) for diagnosing lymph node metastases in cervical cancer (48). In this study, the authors performed a traditional pairwise meta-analysis of studies that directly compared different diagnostic modalities (48). Secondly, the authors drew a network evidence diagram (network plot), whereby each node represented a different imaging method, node size reflected sample size, and the thickness of the line between nodes represented the number of included studies. A node splitting analysis showing a pair-wise comparison of the imaging methods is shown in **Figure 9** (48). Thirdly, the authors conducted a Bayesian NMA comparing different diagnostic modalities (48). A Bayesian approach adopting probability values summarized as surface under the cumulative ranking curve (SUCRA) was the most effective method (48). Another example of a NMA studied the outcomes of non-invasive diagnostic modalities for the detection of coronary artery disease (49). In this study, again, the authors performed a traditional pair-wise meta-analysis of studies that compared the different diagnostic modalities with associated network plots (49).

Individual Patient/Participant Data (IPD) Meta-analysis

Meta-analyses of individual patient/participant data (IPD) have been performed for therapeutic studies often using data from randomized controlled trials (RCTs). It has been shown that some of these IPD meta-analyses have influenced the selection of comparators and participants, sample size calculations, analysis and interpretation of subsequent trials, and the conduct and analysis of ongoing trials (50). Systematic reviews and meta-analyses of IPD have also been used to inform clinical guideline recommendations (51). Individual patient/participant data meta-analyses of test accuracy studies can also be performed and have advantages over conventional meta-analyses (52). They acquire the raw data from the studies (52). They can help elucidate the incremental information provided by testing over and above that already known from history and examination (52). Conventional meta-analyses usually assess a single diagnostic test compared to a clinical reference standard, often in isolation from the previous tests. Meta-analyses of IPD potentially allow assessment of for a complete diagnostic sequence starting with history, examination and testing, considering all the testing. This takes into account the redundancy of information (52). An IPD meta-analysis can be performed retrospectively or prospectively. In a retrospective IPD meta-analyses, authors are contacted and invited to supply raw data from their primary study. Ideally, an IPD meta-analysis should be performed prospectively as this ensures uniformity of data and its quality. Prospective IPD meta-analysis are often referred to the as the gold standard of meta-analyses.

Multivariable analyses of IPD meta-analysis allow for the redundancy of information in tests and is less likely to overestimate incremental test accuracy (52). These analyses can also help determine the optimal sequence in which tests especially the more advanced tests should be performed (52). Because the unit of analysis is at the patient level rather than the study level, there is greater power to explore heterogeneity and perform meta-regression analyses. Association across patient-level characteristics or

between patient level and study level characteristics can be explored reducing ecological bias (53). In addition, IPD meta-analyses may allow the development or evaluation of diagnostic algorithms for individual patients. They also allow the analysis of continuous test results rather than the dichotomous classification (with loss of information) that is generally used in reports of diagnostic tests (53).

Conclusion

Herein we have provided an overview of methodological and reporting best practices for MRI systematic reviews. Systematic reviews are regarded as high-level evidence, which can influence clinical decision-making and healthcare policy making. Methodological rigor and complete reporting are crucial to ensure the systematic reviews these decisions are based on are of the highest possible quality. The methodological guidance in this review is by no means exhaustive and we encourage authors to refer to the Cochrane Handbook for DTA Reviews if further guidance is sought (16). Similarly the PRISMA-DTA statement should be referenced for complete reporting requirements of DTA systematic reviews (19).

References

1. Higgins JPT, Green S, Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions. Chichester, England ; Hoboken, NJ: Wiley-Blackwell: 2008. xxi, 649 p. p.
2. Choi SH, Kim SY, Park SH, et al. Diagnostic performance of CT, gadoxetate disodium-enhanced MRI, and PET/CT for the diagnosis of colorectal liver metastasis: Systematic review and meta-analysis. *J Magn Reson Imaging* 2017.
3. Duncan JK, Ma N, Vreugdenburg TD, Cameron AL, Maddern G. Gadoteric acid-enhanced MRI for the characterization of hepatocellular carcinoma: A systematic review and meta-analysis. *J Magn Reson Imaging* 2017;45(1):281-290.
4. Kim JR, Suh CH, Yoon HM, et al. Performance of MRI for suspected appendicitis in pediatric patients and negative appendectomy rate: A systematic review and meta-analysis. *J Magn Reson Imaging* 2018;47(3):767-778.
5. Symanski JS, Subhas N, Babb J, Nicholson J, Gyftopoulos S. Diagnosis of Superior Labrum Anterior-to-Posterior Tears by Using MR Imaging and MR Arthrography: A Systematic Review and Meta-Analysis. *Radiology* 2017;285(1):101-113.
6. Poot DHJ, van der Heijden RA, van Middelkoop M, Oei EHG, Klein S. Dynamic contrast-enhanced MRI of the patellar bone: How to quantify perfusion. *J Magn Reson Imaging* 2018;47(3):848-858.
7. Gordon LG, James R, Tuffaha HW, Lowe A, Yaxley J. Cost-effectiveness analysis of multiparametric MRI with increased active surveillance for low-risk prostate cancer in Australia. *J Magn Reson Imaging* 2017;45(5):1304-1315.
8. Weng HH, Noll KR, Johnson JM, et al. Accuracy of Presurgical Functional MR Imaging for Language Mapping of Brain Tumors: A Systematic Review and Meta-Analysis. *Radiology* 2018;286(2):512-523.
9. Yokoo T, Serai SD, Pirasteh A, et al. Linearity, Bias, and Precision of Hepatic Proton Density Fat Fraction Measurements by Using MR Imaging: A Meta-Analysis. *Radiology* 2018;286(2):486-498.
10. McInnes MD, Hibbert RM, Inácio JR, Schieda N. Focal Nodular Hyperplasia and Hepatocellular Adenoma: Accuracy of Gadoteric Acid-enhanced MR Imaging-A Systematic Review. *Radiology* 2015;277(3):927.
11. Kaye EA, Granlund KL, Morris EA, Maybody M, Solomon SB. Closed-Bore Interventional MRI: Percutaneous Biopsies and Ablations. *AJR Am J Roentgenol* 2015;205(4):W400-410.
12. McInnes MD, Bossuyt PM. Pitfalls of Systematic Reviews and Meta-Analyses in Imaging Research. *Radiology* 2015;277(1):13-21.
13. 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(1):78-85.

14. 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.
15. 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(2):413-426.
16. Deeks J, Bossuyt P, Gatsonis C. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: The Cochrane Collaboration: 2013.*
17. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
18. 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(1):194.
19. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* 2018;319(4):388-396.
20. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Radiology* 2015;277(3):826-832.
21. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6(11):e012799.
22. Hong PJ, Korevaar DA, McGrath TA, et al. Reporting of imaging diagnostic accuracy studies with focus on MRI subgroup: Adherence to STARD 2015. *J Magn Reson Imaging* 2018;47(2):523-544.
23. Schieda N, Lim CS, Idris M, et al. MRI assessment of pathological stage and surgical margins in anterior prostate cancer (APC) using subjective and quantitative analysis. *J Magn Reson Imaging* 2017;45(5):1296-1303.
24. Lim CS, McInnes MDF, Lim RS, et al. Prognostic value of Prostate Imaging and Data Reporting System (PI-RADS) v. 2 assessment categories 4 and 5 compared to histopathological outcomes after radical prostatectomy. *J Magn Reson Imaging* 2017;46(1):257-266.
25. Krishna S, Lim CS, McInnes MDF, et al. Evaluation of MRI for diagnosis of extraprostatic extension in prostate cancer. *J Magn Reson Imaging* 2018;47(1):176-185.
26. Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Defining the clinical pathway in cochrane diagnostic test accuracy reviews. *BMC Med Res Methodol* 2016;16(1):153.
27. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol* 2006;59(3):234-240.
28. Grégoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? *J Clin Epidemiol* 1995;48(1):159-163.
29. Leeflang M, Reitsma J, Scholten R, et al. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem* 2007;53(2):164-172.
30. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529-536.
31. 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.

32. Schreuder SM, Lensing R, Stoker J, Bipat S. Monitoring treatment response in patients undergoing chemoradiotherapy for locally advanced uterine cervical cancer by additional diffusion-weighted imaging: A systematic review. *J Magn Reson Imaging* 2015;42(3):572-594.
33. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol* 2005;5:19.
34. Riley RD, Ahmed I, Ensor J, et al. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. *Syst Rev* 2015;4:12.
35. 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(10):982-990.
36. McGrath T, McInnes M, Langer F, Hong J, Korevaar D, Bossuyt P. Treatment of multiple test readers in diagnostic accuracy systematic reviews of imaging studies. *European Journal of Radiology* 2017;93:59-64.
37. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20(19):2865-2884.
38. Cronin P, Kelly AM, Altaee D, Foerster B, Petrou M, Dwamena BA. How to Perform a Systematic Review and Meta-analysis of Diagnostic Imaging Studies. *Acad Radiol* 2018.
39. Leeftang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* 2013;185(11):E537-544.
40. Network E. Reporting Guidelines under development. Volume 2014.
41. Repplinger MD, Levy JF, Peethumnongsin E, et al. Systematic review and meta-analysis of the accuracy of MRI to diagnose appendicitis in the general population. *J Magn Reson Imaging* 2015.
42. Giljaca V, Gurusamy KS, Takwoingi Y, et al. Endoscopic ultrasound versus magnetic resonance cholangiopancreatography for common bile duct stones. *Cochrane Database Syst Rev* 2015(2):CD011549.
43. Deeks JJ. Raising the Bar: Further Improvement is Required to Make More Test Accuracy Research Fit for Decision-making. *Clin Chem* 2017;63(8):1315-1317.
44. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004;23(20):3105-3124.
45. Tonin FS, Rotta I, Mendes AM, Pontarolo R. Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. *Pharm Pract (Granada)* 2017;15(1):943.
46. Chaimani A, Higgins JP, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS One* 2013;8(10):e76654.
47. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. *PLoS One* 2014;9(7):e99682.
48. Luo Q, Luo L, Tang L. A Network Meta-Analysis on the Diagnostic Value of Different Imaging Methods for Lymph Node Metastases in Patients With Cervical Cancer. *Technol Cancer Res Treat* 2018;17:1533034617742311.
49. Siontis GC, Mavridis D, Greenwood JP, et al. Outcomes of non-invasive diagnostic modalities for the detection of coronary artery disease: network meta-analysis of diagnostic randomised controlled trials. *BMJ* 2018;360:k504.
50. Tierney JF, Pignon JP, Gueffroy F, et al. How individual participant data meta-analyses have influenced trial design, conduct, and analysis. *J Clin Epidemiol* 2015;68(11):1325-1335.
51. Vale CL, Rydzewska LH, Rovers MM, et al. Uptake of systematic reviews and meta-analyses based on individual participant data in clinical practice guidelines: descriptive study. *BMJ* 2015;350:h1088.

52. Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. *Eur J Obstet Gynecol Reprod Biol* 2003;108(2):121-125.
53. Broeze KA, Opmeer BC, Bachmann LM, et al. Individual patient data meta-analysis of diagnostic and prognostic studies in obstetrics, gynaecology and reproductive medicine. *BMC Med Res Methodol* 2009;9:22.

Figure Captions

Figure 1: Example of subgroup analysis results. Figure reprinted from: Duncan JK, Ma N, Vreugdenburg TD, Cameron AL, Maddern G. Gadoteric acid-enhanced MRI for the characterization of hepatocellular carcinoma: A systematic review and meta-analysis. *J Magn Reson Imaging* 2017;45(1):281-290.

Figure 2: Example of meta-regression results. Figure reprinted from: Kim JR, Suh CH, Yoon HM, et al. Performance of MRI for suspected appendicitis in pediatric patients and negative appendectomy rate: A systematic review and meta-analysis. *J Magn Reson Imaging* 2018;47(3):767-778.

Figure 3: PRISMA flow diagram template

Figure 4: Sample QUADAS-2 risk of bias and applicability results table

Figure 5: Sample table showing 2x2 data, estimates of sensitivity and specificity (with confidence

intervals) and associated forest plots for included primary studies. Figure reprinted from: Replinger MD, Levy JF, Peethumnongsin E, et al. Systematic review and meta-analysis of the accuracy of MRI to diagnose appendicitis in the general population. *J Magn Reson Imaging* 2015.

Figure 6: HSROC curve displaying summary estimate, 95% confidence region and 95% prediction region. Figure printed from: Choi SH, Kim SY, Park SH, et al. Diagnostic performance of CT, gadoxetate disodium-enhanced MRI, and PET/CT for the diagnosis of colorectal liver metastasis: Systematic review and meta-analysis. *J Magn Reson Imaging* 2017.

Figure 7: Summary of findings table. Figure reprinted from: Giljaca V, Gurusamy KS, Takwoingi Y, et al. Endoscopic ultrasound versus magnetic resonance cholangiopancreatography for common bile duct stones. *Cochrane Database Syst Rev* 2015(2):CD011549.

Figure 8A: A schematic representation of a review with a direct comparison meta-analyses (otherwise known as an umbrella review). In this hypothetical example, there are 7 index test options (ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single-photon emission computed tomography (SPECT), PET-MRI and SPECT-MRI) compared to one reference test, catheter angiography. This results in 13 comparisons options.

Figure 8B: A schematic representation of a network review with the same data. Each test is shown by a node (circle) of different color, which represents the test being evaluated. Node size is dependant on the number of studies for each test. More studies result in a larger node. The lines or links that connect the tests (nodes) represent the comparisons. Line width is dependant on the number of direct evidence studies. The thicker the line the more direct evidence studies.

Figure 9: Node splitting plot of the diagnostic value of 4 imaging methods in the diagnosis of lymph node metastasis in patients with CC. A, Magnetic resonance imaging. B, Positron emission tomography. C, Computer tomography; D, diffusion-weighted imaging. CC indicates cervical cancer. Figure reprinted from: Luo Q, Luo L, Tang L. A Network Meta-Analysis on the Diagnostic Value of Different Imaging Methods for Lymph Node Metastases in Patients With Cervical Cancer. *Technol Cancer Res Treat* 2018;17:1533034617742311.

TABLE 2. Summary of Results From the Subgroup Analyses

| Subgroup | Number of studies | All sized lesions GA-MRI | | All sized lesions CE-CT | |
|------------------------------------|-------------------|-----------------------------------|-------------------------------|-------------------------------|-------------------------------|
| | | Sensitivity estimate [95% CI] | Specificity estimate [95% CI] | Sensitivity estimate [95% CI] | Specificity estimate [95% CI] |
| All studies | 7 | 0.881 [0.766, 0.94] | 0.926 [0.829, 0.97] | 0.713 [0.577, 0.819] | 0.918 [0.829, 0.963] |
| Retrospective studies ^a | 5 | 0.872 [0.733, 0.944] | 0.941 [0.857, 0.977] | 0.689 [0.539, 0.807] | 0.935 [0.859, 0.971] |
| Consecutive recruitment | 4 | 0.901 [0.834, 0.943] ^b | 0.951 [0.91, 0.974] | 0.698 [0.571, 0.801] | 0.956 [0.917, 0.977] |
| NR if consecutive recruitment | 3 | 0.853 [0.501, 0.971] | 0.894 [0.538, 0.984] | 0.738 [0.403, 0.922] | 0.876 [0.583, 0.973] |
| RS: Surgical specimen only | 2 | 0.667 [0.35, 0.879] | 0.972 [0.866, 0.995] | 0.616 [0.228, 0.898] | 0.963 [0.893, 0.988] |
| RS: mixed | 5 | 0.919 [0.875, 0.949] ^b | 0.894 [0.757, 0.958] | 0.751 [0.916, 0.848] | 0.894 [0.755, 0.96] |
| Reported conflict of interest | 2 | 0.934 [0.836, 0.975] | 0.764 [0.665, 0.841] | 0.852 [0.740, 0.921] | 0.764 [0.665, 0.841] |
| NR if conflict of interest | 4 | 0.826 [0.603, 0.937] | 0.959 [0.92, 0.979] | 0.637 [0.440, 0.798] | 0.965 [0.931, 0.982] |

| Subgroup | Number of studies | All sized lesions GA-MRI | | All sized lesions D-MRI | |
|----------------------------|-------------------|-----------------------------------|-------------------------------|-------------------------------|-------------------------------|
| | | Sensitivity estimate [95% CI] | Specificity estimate [95% CI] | Sensitivity estimate [95% CI] | Specificity estimate [95% CI] |
| All studies | 6 | 0.907 [.870, 0.934] | 0.929 [0.877, 0.961] | 0.820 [0.776, 0.857] | 0.934 [0.881, 0.964] |
| Consecutive recruitment | 4 | 0.905 [0.863, 0.936] | 0.942 [0.882, 0.972] | 0.821 [0.759, 0.870] | 0.943 [0.878, 0.975] |
| NR if consecutive | 2 | 0.935 [0.853, 0.973] | 0.927 [0.749, 0.982] | 0.831 [0.738, 0.896] | 0.968 [0.803, 0.995] |
| RS: mixed ^a | 5 | 0.908 [0.869, 0.936] ^b | 0.941 [0.886, 0.97] | 0.812 [0.759, 0.856] | 0.934 [0.875, 0.966] |
| No conflict of interest | 2 | 0.927 [0.846, 0.967] ^b | 0.946 [0.809, 0.987] | 0.761 [0.658, 0.840] | 0.948 [0.813, 0.987] |
| NR if conflict of interest | 4 | 0.903 [0.860, 0.934] | 0.931 [0.869, 0.965] | 0.842 [0.793, 0.881] | 0.93 [0.866, 0.965] |

^aIf there was only one study in a particular subgroup, then results from this group were not reported. Instead, analysis for the remaining studies was reported as a sensitivity analysis removing the unique study and should be compared to the entire cohort of studies. Estimates of sensitivity and specificity are from the bivariate model.

^bResults are statistically significant and favor GA-MRI.

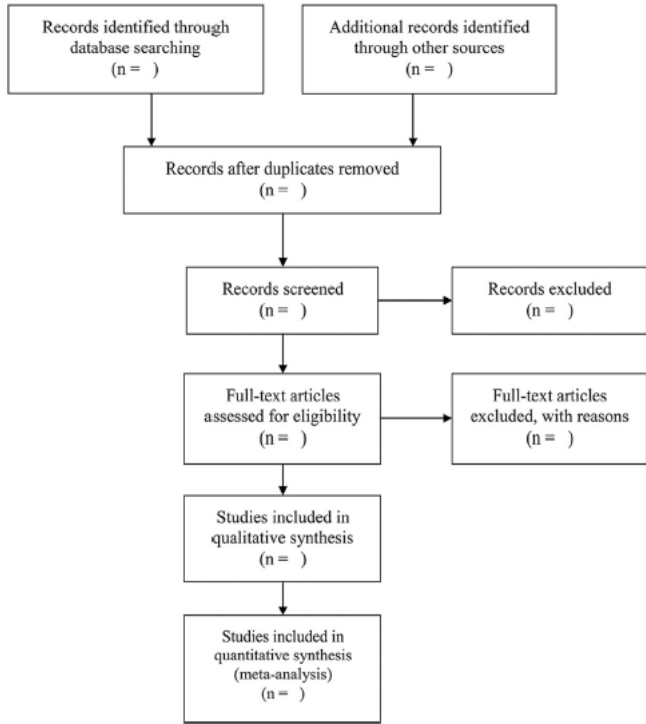
CE-CT = contrast-enhanced computed tomography. CI = confidence interval. D-MRI = dynamic MRI, GA-MRI = MRI with gadoteric acid contrast, NR = not reported, RS = reference standard.

TABLE 3. Results of the Meta-regression of MRI for the Diagnosis of Acute Appendicitis in Pediatric Patients

| Covariates | Subgroup | No. of study | Meta-analytic summary estimates | | | AUROC (95% CI) |
|------------------------------------|----------------------------------|--------------|---------------------------------|----------------------|----------|------------------|
| | | | Sensitivity (95% CI) | Specificity (95% CI) | <i>P</i> | |
| Study design | Prospective | 5 | 97% (95–100%) | 95% (89–100%) | 0.76 | 1.00 (0.99–1.00) |
| | Retrospective | 8 | 97% (95–100%) | 97% (95–100%) | | |
| Study population | Suspected appendicitis | 10 | 97% (95–99%) | 96% (92–99%) | 0.43 | 0.98 (0.97–0.99) |
| | US inconclusive results | 3 | 98% (94–100%) | 99% (96–100%) | | |
| Age, years | ≥12 | 6 | 98% (96–100%) | 95% (91–99%) | <0.01 | 0.98 (0.96–0.99) |
| | <12 | 6 | 96% (94–98%) | 98% (96–100%) | | |
| Proportion of appendicitis | ≥30% | 6 | 97% (95–99%) | 95% (89–100%) | 0.54 | 0.99 (0.98–1.00) |
| | <30% | 7 | 97% (95–99%) | 98% (95–100%) | | |
| Magnetic field strength | 1.5-T only | 8 | 97% (95–99%) | 94% (89–98%) | 0.07 | 0.99 (0.97–0.99) |
| | 3-T only, or combined with 1.5-T | 5 | 97% (94–99%) | 99% (97–100%) | | |
| Use of DWI | Yes | 4 | 97% (93–100%) | 96% (92–100%) | <0.01 | 0.99 (0.98–1.00) |
| | No | 8 | 99% (96–100%) | 97% (95–100%) | | |
| Addition of contrast enhanced scan | Yes | 2 | 97% (93–100%) | 97% (90–100%) | 0.95 | NA |
| | No | 11 | 97% (95–99%) | 97% (94–100%) | | |
| Blinding to pathologic report | Yes | 8 | 98% (96–100%) | 97% (94–100%) | 0.65 | 1.00 (0.99–1.00) |
| | No | 5 | 96% (94–99%) | 97% (93–100%) | | |
| Use of structured reporting | Yes | 7 | 98% (95–100%) | 96% (91–100%) | 0.65 | 0.98 (0.96–0.99) |
| | No | 6 | 96% (94–98%) | 97% (95–100%) | | |

NA = not available due to small numbers of studies; DWI = diffusion-weighted imaging.




Identification
Screening
Eligibility
Included

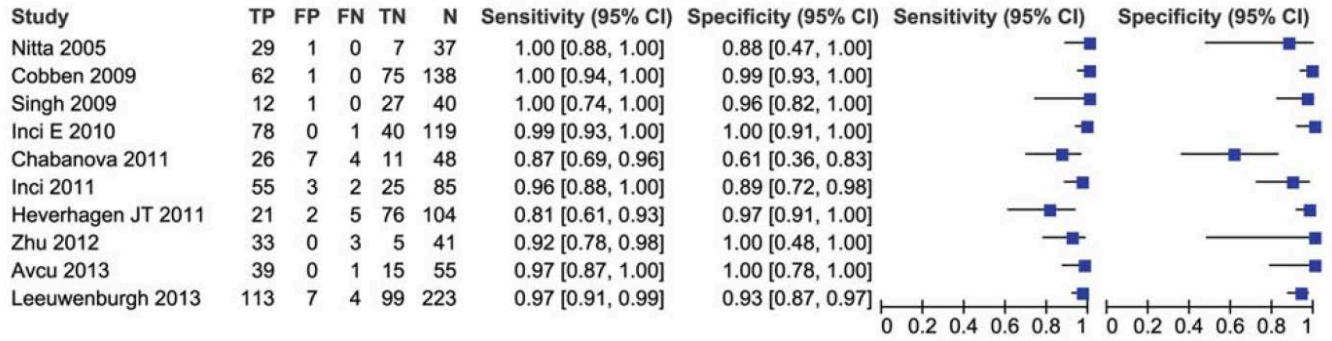


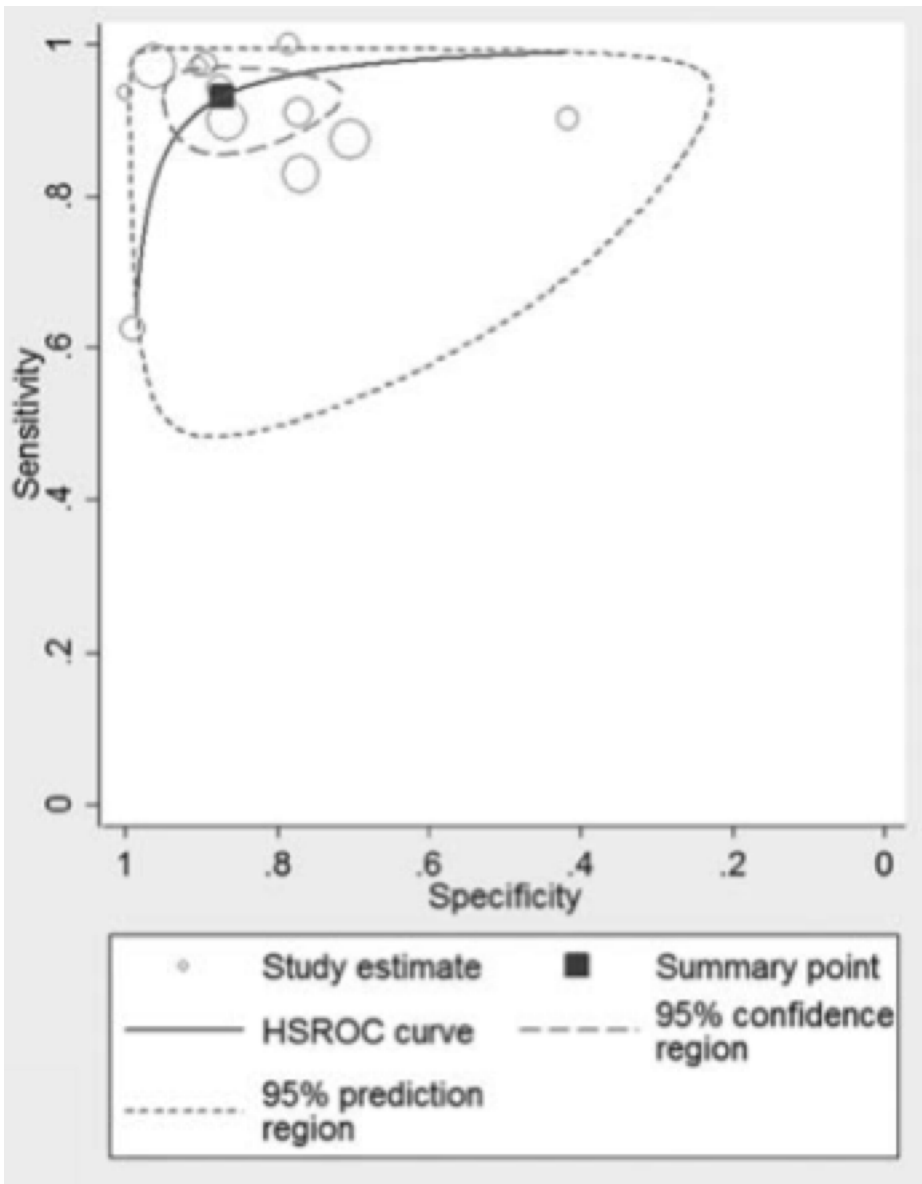
Risk of Bias

Applicability Concerns

| | Patient Selection | Index Test | Reference Standard | Flow and Timing | Patient Selection | Index Test | Reference Standard |
|---------|-------------------|------------|--------------------|-----------------|-------------------|------------|--------------------|
| Study 1 | + | + | + | + | + | + | + |
| Study 2 | - | ? | ? | + | - | ? | ? |
| Study 3 | + | + | + | ? | + | + | + |
| Study 4 | ? | ? | ? | - | ? | ? | ? |

| | | |
|---|--|--|
|  High |  Unclear |  Low |
|---|--|--|





Summary of findings

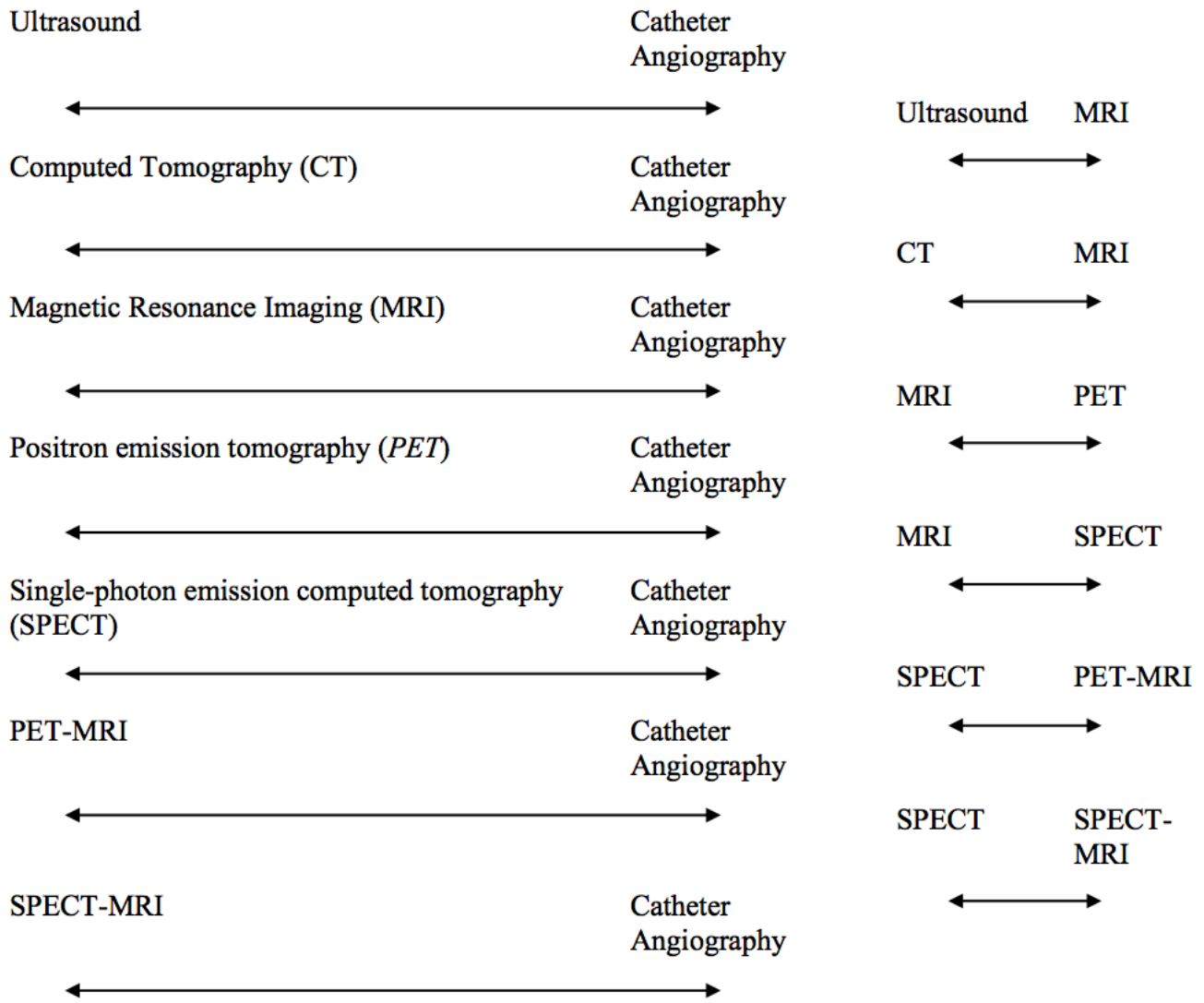
| | | | | | |
|---|---|--|-------------------------------------|--|--|
| Population | Patients suspected of having common bile duct stones based on symptoms, liver function tests, and ultrasound | | | | |
| Settings | Secondary and tertiary care setting in different parts of the world | | | | |
| Index tests | Endoscopic ultrasound (EUS) and magnetic resonance cholangiopancreatography (MRCP) | | | | |
| Reference standard | Endoscopic or surgical extraction of stones in patients with a positive index test result or clinical follow-up (minimum 6 months) in patients with a negative index test result | | | | |
| Target condition | Common bile duct stones | | | | |
| Number of studies | A total of 18 studies were included. Thirteen studies (686 cases, 1537 participants) evaluated EUS and 7 studies (361 cases, 996 participants) evaluated MRCP. Two of the studies evaluated both tests in the same patients | | | | |
| Methodological concerns | quality | All the studies were of poor methodological quality; most studies were at high risk of bias or gave high concern about applicability across all domains of quality assessment, or both | | | |
| Pre-test probability¹ | Test | Summary sensitivity (95% CI) | Summary specificity (95% CI) | Positive post-test probability (95% CI)² | Negative post-test probability (95% CI)³ |
| 0.14 | EUS | 0.95 (0.91 to 0.97) | 0.97 (0.94 to 0.99) | 0.85 (0.72 to 0.93) | 0.01 (0.01 to 0.02) |
| | MRCP | 0.93 (0.87 to 0.96) | 0.96 (0.89 to 0.98) | 0.79 (0.61 to 0.90) | 0.01 (0.01 to 0.02) |
| 0.30 | EUS | 0.95 (0.91 to 0.97) | 0.97 (0.94 to 0.99) | 0.94 (0.87 to 0.97) | 0.02 (0.01, 0.04) |
| | MRCP | 0.93 (0.87 to 0.96) | 0.96 (0.89 to 0.98) | 0.90 (0.80 to 0.96) | 0.03 (0.02, 0.06) |
| 0.41 | EUS | 0.95 (0.91 to 0.97) | 0.97 (0.94 to 0.99) | 0.96 (0.92 to 0.98) | 0.03 (0.02, 0.06) |
| | MRCP | 0.93 (0.87 to 0.96) | 0.96 (0.89 to 0.98) | 0.94 (0.87 to 0.97) | 0.05 (0.03 to 0.09) |
| 0.48 | EUS | 0.95 (0.91 to 0.97) | 0.97 (0.94 to 0.99) | 0.97 (0.93, 0.99) | 0.05 (0.03 to 0.08) |
| | MRCP | 0.93 (0.87 to 0.96) | 0.96 (0.89 to 0.98) | 0.95 (0.90 to 0.98) | 0.06 (0.04 to 0.11) |
| 0.68 | EUS | 0.95 (0.91 to 0.97) | 0.97 (0.94 to 0.99) | 0.99 (0.97 to 0.99) | 0.10 (0.06 to 0.16) |
| | MRCP | 0.93 (0.87 to 0.96) | 0.96 (0.89 to 0.98) | 0.98 (0.95 to 0.99) | 0.13 (0.08 to 0.23) |
| <p>Comparison of the diagnostic accuracy of EUS and MRCP: at pre-test probabilities of 14%, 41%, and 58%, out of 100 people with positive EUS, common bile duct stones will be present in 85, 96, and 99 people respectively; while out of 100 people with positive MRCP, common bile duct stones will be present in 79, 94, and 98 people. For the same pre-test probabilities, out of 100 people with negative EUS, common bile duct stones will be present in 1, 3, and 10 people respectively; while out of 100 people with negative MRCP, common bile duct stones will be present in 1, 5, and 13 people respectively</p> | | | | | |
| <p>Conclusions: the performance of EUS and MRCP appears to be comparable for diagnosis of common bile duct stones. The strength of the evidence for the test comparison was weak because the studies were methodologically flawed, and only two studies made head-to-head comparisons of EUS and MRCP</p> | | | | | |

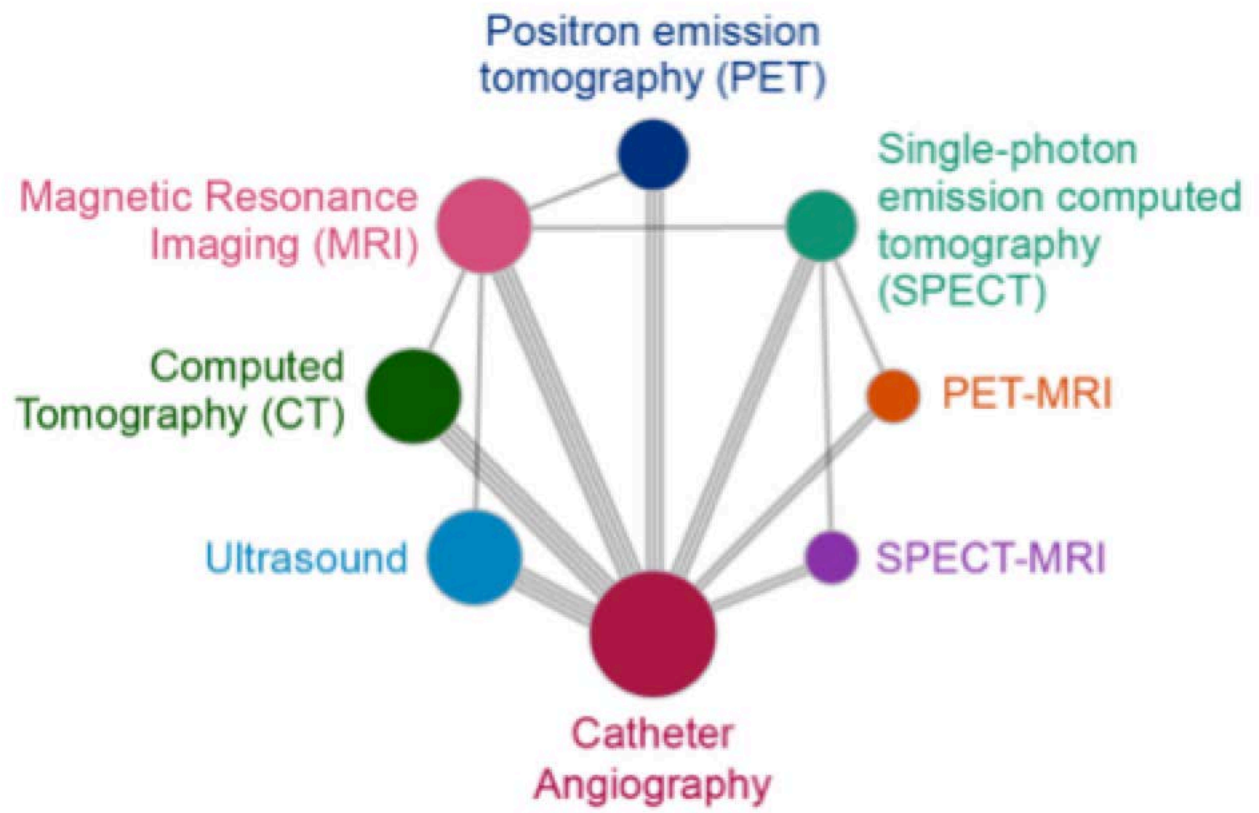
¹ The pre-test probability (proportion with common bile duct stones out of the total number of participants) was computed for each included study. These numbers represent the minimum, lower quartile, median, upper quartile and the maximum values from the 18 studies.

² Post-test probability of common bile duct stones in people with positive index test results.

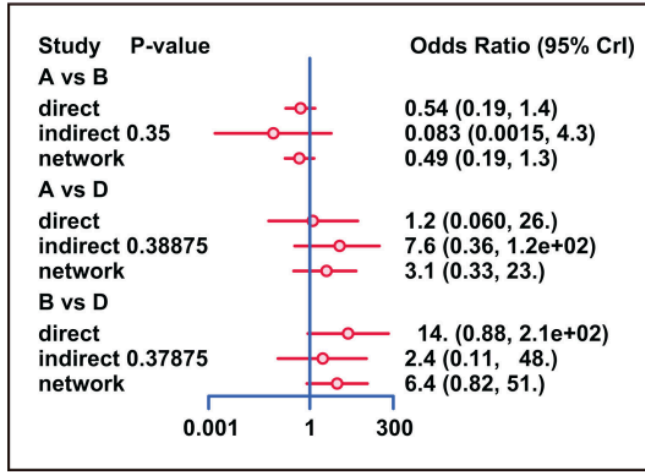
³ Post-test probability of common bile duct stones in people with negative index test results.

Author Manuscript

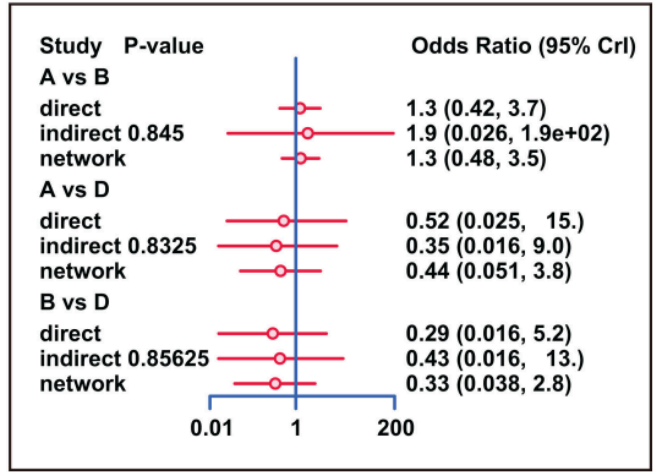




A (positive likelihood ratio)



B (negative likelihood ratio)



C (diagnostic odds ratio)

