Title:

Updating Diagnostic Test Accuracy Systematic Reviews: which, when and how should they be updated?

Authors:

Ersilia Lucenteforte[1,*] Alessandra Bettiol[1], Salvatore De Masi[2] , Gianni Virgili[3]

Affiliation:

1. Department of Neurosciences, Psychology, Drug Research and Child Health (NEUROFARBA), University of Florence, viale Gaetano Pieraccini, 6 – 50139 Florence, Italy

2. Clinical Trial Office, University hospital "Azienda Ospedaliero-Universitaria Meyer", Viale Pieraccini 24 - 50139 Florence, Italy

3. Department of Surgery and Translational Medicine (DCMT), University of Florence, Largo Brambilla, 3 – 50134 Florence, Italy

Corresponding author*:

Ersilia Lucenteforte, ScD, PhD,

Department of Neurosciences, Psychology, Drug Research and Child Health (NEUROFARBA), University of Florence, viale Gaetano Pieraccini, 6 – 50139 Florence, Italy

e-mail: ersilia.lucenteforte@unifi.it

tel: 00390552758197

Keywords:

checklist; currency; novelty; value of information; refreshing

ABSTRACT

Updating diagnostic test accuracy (DTA) systematic reviews (SR) is fundamental in order to avoid making clinical decisions based on out-of-date and/or incomplete information. The decision of which SR to update should be based on the quality of the SR and on the currency of its topic. If SRs are considered worthy of updating, priority should first be establisheddepending on the availability of elements of novelty (in terms of published studies, methodology, decisional approach, or standards of quality), as well as on the expected impact and value of information.
Before starting the updating process, a careful work plan is necessary, refreshing the state of the art, the aim, and the methods. Once the update has been performed, new findings and conclusions should be clearly displayed.

INTRODUCTION

The number of diagnostic test accuracy (DTA) studies has rapidly increased, especially over the last five years. A quick PUBMED search (sensitivity[tiab] OR specificity[tiab] ORaccuracy[tiab], Filters: Humans) revealed there were 15,772 published studies in 2000, 20,916 in 2005 (5,144 more than 2000), 28,723 in 2010 (7,807 more than 2005) and 39,110 in 2015 (10,387 more more than 2010).

Systematic reviews (SRs) represent a very useful tool to synthesize the most relevant findings from different studies regarding a specific DTA question, as well as to investigate the possible reasons for discrepancies among studies and to assess the efficacy and clinical impact of new tests [1].

In the diagnostic field, assessing the impact of a new test is particularly critical and much more complicated than assessing the impact of new treatments. In fact, differently from new therapies, which are directly connected to clinical outcomes (either therapeutic effect or adverse event), the relationship between a new diagnostic test and the final clinical outcome is much more complex and indirect [2] (**Figure 1**). In light of this, studies on new tests naturally tend to concentrate more on the performance of the test alone (sensitivity, specificity, safety and costs), rather than on its overall possible clinical impact, with DTA studies measuring only sensitivity and specificity (such as cross-sectional studies) being more common than Randomized Clinical Trials (RCTs). In diagnostic fields, in fact, RCTs are not required for marketing approval, and new diagnostic tests often enter clinical practice without having their impact tested on patient outcomes [3]. All this makes it particularly complex to write a DTA SR and summarize evidence on the clinical impact of a new test. Moreover, it is time- and labor-consuming to write a DTA SR, evaluate and re-elaborate the latest and most interesting works concerning the topic of interest.

Despite the effort, SRs in general are destined to become “out of date” due to ongoing progress and newly published works. In fact, new studies may contain relevant elements of novelty which could significantly affect the conclusions and validity of the previous review, making it not only incomplete but also misleading. This is particularly true in the field of DTA, given the high rate of development of new diagnostic tools, ongoing changes in current reference techniques, and clinical pathways.

Therefore, when elements of novelty are available, systematic reviews need to be either completely re-written or updated. Updating a SR is defined as a new version of a previously published review with novelties in terms of data, methods, or analyses compared to the previous edition [4]. Compared to an ex-novo edition, an update presents significant advantages, since it is generally more efficient and time-saving. However, the decision whether to update or completely re-write a new review

should be based on the quality of the already existing review. In fact, if the SR was imprecisely conducted using unsound methods (e.g. vague inclusion criteria, poorly developed outcomes, etc.), then starting all over again is probably the best choice. AMSTAR (A Measurement Tool to Assess Systematic Reviews) is an example of a useful instrument that can be adopted to assess the overall quality of a SR [5], while PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) is useful to check reporting [6]. If updating is more convenient than re-writing, authors need to choose the review to update and when.

WHICH SR SHOULD BE UPDATED?

Updating a SR is a challenging process and its real worth must be well-balanced in order to channel the efforts into those areas where new evidence is crucial. Therefore, choosing which reviews should be updated is a central problem. The systematic approach used to prioritize which SR to update may vary among guideline panels and review groups. However, some elements are common key points in the decisional approach (**Figure 2**).

Among them, **currency** of a systematic review is essential. In fact, only reviews addressing questions of current interest can be considered worthy of being updated. This is particularly true in the field of DTA, where diagnostic tests are rapidly and easily exceeded by more modern diagnostic techniques (particularly in the field of imaging).

To estimate the currency of a DTA SR, different strategies can be used. Among them, analysis of diagnostic approaches adopted in clinical practice is fundamental. In fact, reviews on diagnostic tests no longer used in routine diagnosis are probably out-of-date and unworthy of being updated.

Another valid strategy to assess currency is to evaluate whether the systematic review under consideration and/or other studies surrounding the DTA of interest receive good access, estimated through metrics for citations and downloads. In fact, widely cited or downloaded articles probably refer to topics of current interest, whereas reviews that are no longer cited or read probably refer to out-of-date diagnostic tools. In the latter case, updating of the review is generally considered unwarranted.

Even when currency is verified, the decision to update a SR must be based on the **quality** of the review itself; as reported above, when a SR addresses a question of current interest but is of poor quality, starting a completely new SR is the best choice.

WHEN SHOULD A SR BE UPDATED?

After having evaluated whether a review is worth being updated, choosing when to update it is a crucial point (**Figure 2**).

Theoretically, SRs should always be kept updated with the newest available evidence in order to avoid healthcare decisions being made on out-of-date or even misleading information. However, updating a SR every time a new study is published is both a utopian and methodologically incorrect approach. In fact, in addition to being extremely time-consuming, a “too-frequent” or “too-soon” update might lead to **randomness bias** since evidence from a single newer trial might completely modify the previous conclusions of the review. This is particularly true in light of the fact that studies with significant and particularly interesting results are more likely to be published and more quickly.

Some review groups arbitrarily decide to update the most relevant reviews with a fixed frequency (The Cochrane Library)[7], whereas others decide to update them according to the availability of elements of novelty[8], or to several other factors such as public health importance, rapidity of scientific developments or nature of the health condition in question (AHRQ-Agency for Healthcare Research and Quality)[9].

To assess when to update, routine surveillance for **newly published studies** around the topic of interest should be performed. Novelties in studies not only include completely original works, but also follow-up results of already included studies (although follow-up studies are uncommon in the diagnostic field since the majority of studies on diagnostic tests are designed as cross-sectional or observational studies).

Given the high rate of publication of DTA studies, a systematic approach is often useful for an exhaustive literature screening. Adopted approaches may vary among groups and are often based on the use of full or abbreviated search strategies, focusing on the review of either the overall literature or of selected groups of core journals in the field of interest [10]. Two valid search approaches are the RAND and Ottawa methods [11-13]. Together with the GRADE approach [14], with statistical prediction tools and value of information analysis (described below), these methods also represent valid tools to estimate how relevant new studies can be in changing or confirming the conclusions of a review.

The RAND method [11] performs an abbreviated search in five major journals to find new studies. Following this first step, the method abstracts the results of relevant articles and qualitatively assesses whether the new findings change or confirm the conclusion of the previous review. The RAND method also includes a subsequent step of consultation of the US Food and Drug Administration website and of external expert judgments to evaluate the currency and possible impact of these findings. Based on this approach, one of four levels indicating update necessity is attributed to the review: 1) Original conclusion is still valid and this portion of the original report does not need updating; 2) Original conclusion is possibly out of date and this portion of the original report may need updating; 3) Original conclusion is probably out of date and this portion of the original report may need updating; 4) Original conclusion is out of date.

On the other hand, the Ottawa method [12] is a full search approach that uses a PubMed search to identify new studies around a selected topic. If new studies are detected, this method performs quantitative and qualitative analysis to evaluate the possible impact of these findings on the conclusions of the review, without involving an expert judgment.

In addition to availability of newly published studies, **novelties in methodology** can affect the decision of when to update a review. Changes in methods are particularly important in the field of DTA SRs, where marked inhomogeneity in results of DTA studies often occurs due to differences in the applied methods [15].

Methodological changes usually involve one or more of the parameters considered in the PICO(S) tool (i.e. Population, Intervention, Comparison, Outcomes and Study design of an article), or in the SPIDER tool (i.e. Sample, Phenomenon of Interest, Design, Evaluation, and Research type of quantitative or mixed-methods studies) [16].

However, methodological novelties may also involve routine approaches and procedures adopted in clinical practice. In diagnostic fields, **changes in the decisional approach** in which the diagnostic test of interest is inserted could significantly affect the conclusions and level of certainty of a review. Therefore, an update of a DTA SR should be considered every time the decisional tree undergoes modifications.

Similarly, changes in the **standards of quality** requested for studies included in the SR could occur as well. In fact, removal of some studies included in the previous version of the review, due to a re-

classification as “poor quality studies” following variations in the reference standards, may lead to significant variations in the overall conclusions.

When elements of novelty have been found, assessing the possible **impact** of novelties on the conclusions and certainty of a systematic review is a crucial step in the decision of whether or not to update a review. Experts in the diagnostic field of interest, as well as editors or referees, can often provide an informed and critical estimate of this impact [11] (RAND method).

However, the consensus of experts is often limitative and not objective. Therefore, different tools have been developed to estimate the impact of an update. A possible approach is GRADE [14], which is based on the assessment of the level of certainty of the evidence reported in a review. According to this approach, the highest assigned level is the certainty of outcomes reported in the review, with the lowest referring to the probability that results from new not-yet-included studies will affect the conclusions of the review.

Assessment of the impact of a review cannot be considered only in terms of gains in scientific knowledge. In fact, the strongest is evidence reported in a review, the highest will be its probability of influencing the clinical practice, leading to both social and economic consequences.

In light of this consideration, also a **value of information analysis** should be performed before starting a review update [17]. This statistical prediction method allows calculation of the gain in terms of reduction of losses related to uncertainty compared to the cost measured in days required to update the SR. Those with significantly positive estimated value of information are worthy of being updated soon due to their probable relevant implications in clinical practice.

Along with the above-mentioned considerations, the moment of updating a review may also be influenced by the **aim** of the update itself. In fact, the objective of systematic reviews can go beyond a simple synthesis of evidence, aiming to estimate a ROC curve or to summarize evidence on the validity of a certain test in a specific setting (such as for a particular clinical condition, or in a particular range of values) [1].

HOW SHOULD A SR BE UPDATED?

If a review has been judged as worthy of updating, authors should carefully plan the work according to these suggested points:

1. Authorship must be updated: if authors differ from those of the first review, then the previous author team should be acknowledged in the update.

2. State of the art must be refreshed, including all background information and evidence already known about the topic.

3. The aim of the previous review should be re-considered to evaluate if it is still relevant to patients and clinical practice. If so, the question of the previous review can be re-addressed, otherwise a new question of current relevance should be formulated.

4. Inclusion criteria should be revisited: not all previously included studies should be included in the new edition. When better quality and larger studies are published, previously-included weaker and smaller studies should be excluded from the update. Similarly, studies comparing the test of interest with obsolete or no longer commonly used tests should be removed.

5. Methods should be revisited: authors are advised to use the latest and most accurate accepted methods, eventually repeating the whole data extraction for all studies.

6. A search for newly published studies should be started, taking into account the new inclusion criteria and aims of the update; thus, search strategies may vary compared to the previous version of the review.

7. A clear description of novelties, in terms of search strategy or methods, must be provided and well-documented to assure replicability. Moreover, given that newly-included studies can partially or radically change the overall conclusions of the review, it is crucial to clearly and attractively present the new findings, highlighting and discussing the differences compared to the previous edition. Users of reviews greatly benefit from a concise and easy to read synthesis of results and novelties, with possible explanations for changes. A valid choice is to use a stand-alone concise summary composed mainly of tables and figures providing a full report with a detailed description of all data and results, especially for those who need more accurate information on the topic [18].

8. Updating can be conducted manually, however this is both time-consuming and poorly efficient;various technological innovations have been developed to increase both the rapidity and efficacy of an update [4]. The implementation of the speed and rapidity of the update process through the already-existing and the under-development tools aims to allow, in a near future, the real-time update of knowledge with the results from new studies [4,19].

CASE STUDY

In order to assess the impact of updating DTA SRs, we searched Cochrane reviews reporting the terms “accuracy” or “sensitivity” or “specificity” in the title or abstract and labelled as updated by a “new search” in the Cochrane Library (**Table 1**). We found four SRs which fulfilled these requirements:

1) Galactomannan detection for invasive aspergillosis in immunocompromized patients [20,21]
2) Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy [22,23]
3) The diagnostic accuracy of the GenoType® MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs [24,25]
4) Diagnostic accuracy of laparoscopy following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer [26,27]

We assessed the following features: months since the original and the updated search and number of new studies found; any change in review objectives in the main text, including PICO, clinical pathway and test role; conclusions as presented in the Abstract, implications for practice and research in the main text; and methodological tools used as risk of bias tool and summary of findings (SoF) table.

The number of studies increased from 30 to 54 in 78 months [20,21], from 9 to 10 in 37 months [22,23], from 21 to 27 in 18 months [24,25], from 15 to 16 in 32 months [26,27].

The review objectives in the main text were unchanged, or rephrased with no substantive change, in three reviews, while Virgili 2015 [23] added details on the clinical pathway and the potential for the index test to replace the reference standard. Regarding PICO components, all reviews were unchanged in terms of index and reference tests, but two reviews [23,25] noticed that different index test versions were available. The test role was explicit in Allen 2013 and Allen 2016 [26,27]

(replacement), while other reviews referred more generically to estimating accuracy with no explicit role.

Regarding the main conclusions presented in the Abstract, Leeflang 2008 and Leeflang 2015 [20,21] used this section to present absolute frequencies of test performance and Allen 2013 and Allen 2014 were also unchanged with minimal rephrasing. On the other hand, Theron 2014 and Theron 2016 [24,25] used very different wording, suggesting a change in the clinical interpretation of results. This was also the case for Virgili 2011 and Virgili 2014 [22,23] who discussed discordances between the index and reference tests in support of the widely accepted dominance of the index test in modern clinical practice.

An update from QUADAS to QUADAS-2 was conducted in three reviews and a Summary of Results or Summary of Findings table was present in all reviews.

This survey of four updated Cochrane DTA reviews suggests no explicit reason for updating was used apart from time since the publication of the original version, except when a change of the index test role was expected. The number of new studies in the update was quite variable, probably reflecting different phases and importance of the test development with respect to the clinical question made in the review. Updating methodological tools, such as for QUADAS checklist version, was the main structural change to the review methodology.

CONCLUSIONS

Updates of DTA SR are a precious instrument for clinical practice as well as regulatory aspects, supporting the decision making approach based on current scientific evidence. Despite their high value, up to now only a few updates of DTA SR have been published. However, considering that the majority of DTA studies of current relevance have been conducted in the most recent years, the number of SR will probably increase significantly in the near future as new evidence becomes available.

REFERENCES

1. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working G. Systematic reviews of diagnostic test accuracy. *Annals of internal medicine.* Dec 16 2008;149(12):889-897.
2. Matchar DB. Chapter 1: Introduction to the Methods Guide for Medical Test Reviews. *Journal of general internal medicine.* Jun 2012;27 Suppl 1:S4-10.
3. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Annals of internal medicine.* Jun 06 2006;144(11):850-855.
4. Garner P, Hopewell S, Chandler J, et al. When and how to update systematic reviews: consensus and checklist. *Bmj.* Jul 20 2016;354:i3507.
5. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC medical research methodology.* Feb 15 2007;7:10.
6. Stovold E, Beecher D, Foxlee R, Noel-Storr A. Study flow diagrams in Cochrane systematic review updates: an adapted PRISMA flow diagram. *Systematic reviews.* May 29 2014;3:54.
7. Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011].* . 2011.
8. Chalmers I, Enkin M, Keirse MJ. Preparing and updating systematic reviews of randomized controlled trials of health care. *The Milbank quarterly.* 1993;71(3):411-437.
9. Tsertsvadze A, Maglione M, Chou R, et al. Updating comparative effectiveness reviews: current efforts in AHRQ's Effective Health Care Program. *Journal of clinical epidemiology.* Nov 2011;64(11):1208-1215.
10. Sagliocca L, De Masi S, Ferrigno L, Mele A, Traversa G. A pragmatic strategy for the review of clinical evidence. *Journal of evaluation in clinical practice.* Aug 2013;19(4):689-696.
11. Shekelle P, Eccles MP, Grimshaw JM, Woolf SH. When should clinical guidelines be updated? *Bmj.* Jul 21 2001;323(7305):155-157.
12. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Annals of internal medicine.* Aug 21 2007;147(4):224-233.
13. Chung M, Newberry SJ, Ansari MT, et al. Two methods provide similar signals for the need to update systematic reviews. *Journal of clinical epidemiology.* Jun 2012;65(6):660-668.
14. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology.* Apr 2011;64(4):401-406.
15. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* Oct 18 2011;155(8):529-536.
16. Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services research.* Nov 21 2014;14:579.
17. Wilson EC. A practical guide to value of information analysis. *PharmacoEconomics.* Feb 2015;33(2):105-121.
18. Newberry SJ, Shekelle PG, Vaiana M, Motala A. *Reporting the Findings of Updated Systematic Reviews of Comparative Effectiveness: How Do Users Want To View New Information?* Rockville (MD)2013.
19. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine.* Feb 2014;11(2):e1001603.
20. Leeflang MM, Debets-Ossenkopp YJ, Visser CE, et al. Galactomannan detection for invasive aspergillosis in immunocompromized patients. *The Cochrane database of systematic reviews.* Oct 08 2008(4):CD007394.
21. Leeflang MM, Debets-Ossenkopp YJ, Wang J, et al. Galactomannan detection for invasive aspergillosis in immunocompromised patients. *The Cochrane database of systematic reviews.* Dec 30 2015(12):CD007394.

22. Virgili G, Menchini F, Murro V, Peluso E, Rosa F, Casazza G. Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. *The Cochrane database of systematic reviews.* Jul 06 2011(7):CD008081.
23. Virgili G, Menchini F, Casazza G, et al. Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. *The Cochrane database of systematic reviews.* Jan 07 2015;1:CD008081.
24. Theron G, Peter J, Richardson M, et al. The diagnostic accuracy of the GenoType((R)) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs. *The Cochrane database of systematic reviews.* Oct 29 2014(10):CD010705.
25. Theron G, Peter J, Richardson M, Warren R, Dheda K, Steingart KR. GenoType(R) MTBDRsl assay for resistance to second-line anti-tuberculosis drugs. *The Cochrane database of systematic reviews.* Sep 08 2016;9:CD010705.
26. Allen VB, Gurusamy KS, Takwoingi Y, Kalia A, Davidson BR. Diagnostic accuracy of laparoscopy following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer. *The Cochrane database of systematic reviews.* Nov 25 2013(11):CD009323.
27. Allen VB, Gurusamy KS, Takwoingi Y, Kalia A, Davidson BR. Diagnostic accuracy of laparoscopy following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer. *The Cochrane database of systematic reviews.* Jul 06 2016;7:CD009323.

Figure 1: Testing the impact of a new diagnostic test

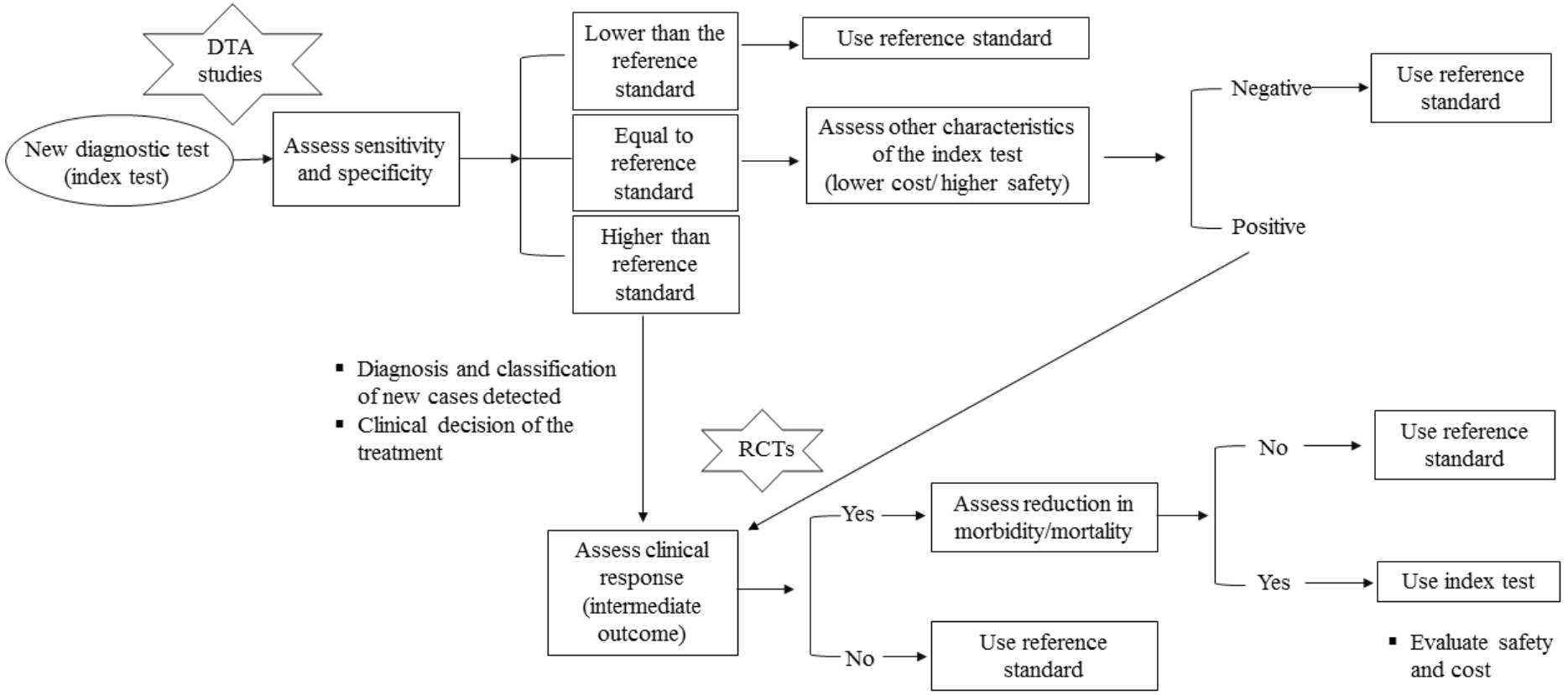

Figure 2: Which and when SRs should be updated

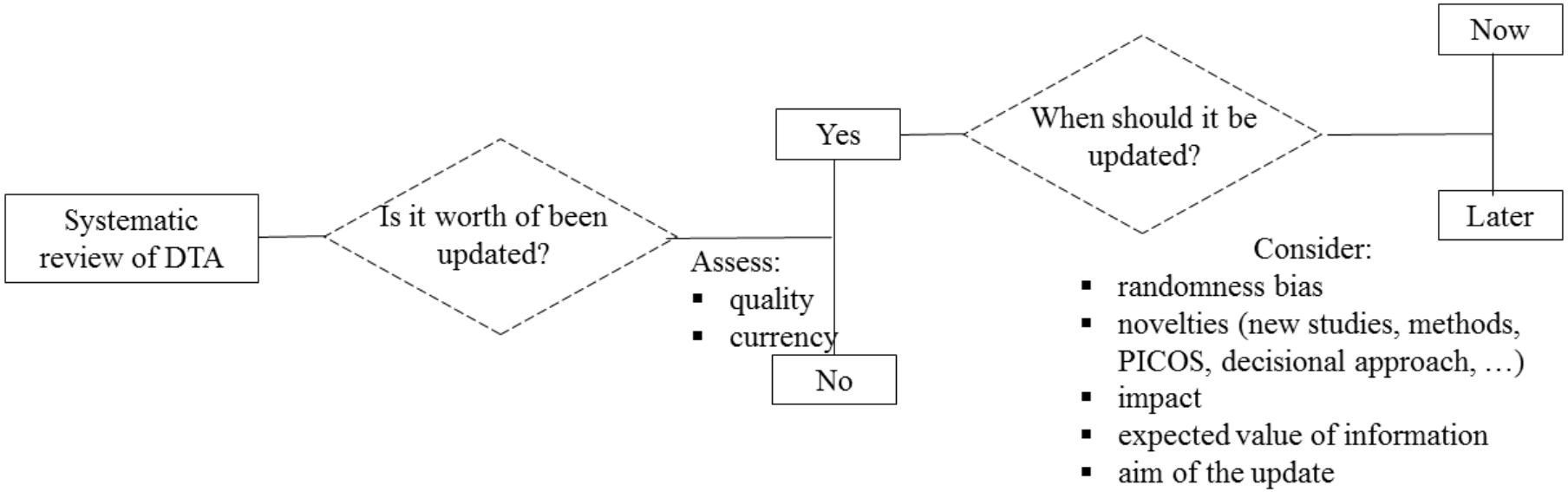

Table 1: Comparison of four SR and their corresponding updates published in the Cochrane Library.

	First version of SR	**Update of SR**
Title of first version: GALACTOMANNAN DETECTION FOR INVASIVE ASPERGILLOSIS IN IMMUNOCOMPROMIZED PATIENTS [20,21]		
YEARS	August 2005- April 2007	February 2014
N. STUDIES	42 (of whom 30 included in the meta-analysis); 6,792 subjects	54 (of whom 50 included in the meta-analysis); 8,305 subjects
OBJECTIVES	Assess the diagnostic accuracy of galactomannan detection in serum for the diagnosis of invasive aspergillosis in immunocompromised patients, at different cut-off values for test positivity.	Same
PATHWAY	There is substantial variation in the way the galactomannan ELISA is currently used in the clinic: -Some clinicians do not use it at all; -Others use the galactomannan ELISA as a screening tool, to monitor whether patients at risk develop Invasive aspergillosis (IA) or not. In those cases, serum is tested for IA once or twice every week. - Sometimes the galactomannan ELISA is used to test for IA in BAL fluid when IA is already suspected and in those situations, the test is only used in serum when there is no BAL fluid. -In most situations, the galactomannan ELISA is used as a triage test: if the ELISA is positive, patients will be referred for further diagnostic testing. -The test is also used in the definition of proven, probable or possible IA, or as final decision making tool to start antifungal therapy.	Same In addition: Further diagnostic testing may involve either laboratory testing of BAL fluid, CT scanning or radiography, or a combination of tests. Patients may also be referred for further diagnostic work-up on the basis of clinical signs and symptoms.
INDEX	Two commercially available assays for the detection of galactomannan: - the Pastorex© latex agglutination test : rarely used -the Platelia© sandwich ELISA test: mostly used for the detection of antigen in serum and in fluid that is obtained via bronchoalveolar lavage (BAL). Other specimens in which the	Same

	test can also be used are cerebrospinal fluid (CSF) or urine. The SR focused on the ELISA test in serum.	
REFERENCE	The following reference standards can be used to define the target condition: -autopsy (gold standard combined with a positive culture of *Aspergillus* species from the autopsy speci- mens, or with histopathological evidence of *Aspergillus*; however autopsy is rarely reported) -the criteria of the EORTC/MSG (reference standard in the SR) -the demonstration of hyphal invasion in biopsies, combinedwith a positive culture for *Aspergillus* species from the same specimens. The criteria of the EORTC/MSG divide the patient population into four categories: patients with proven IA, patients who probably have IA, patients who possibly have IA, and patients without IA. Clinical studies have shown that these criteria do not match autopsy results perfectly. This especially true for the possible category. For clinical trials investigating the effect of treatment, for example, it is recommended that only the proven and probable categories are used.	Same In addition: The exclusion of patients with 'possible' invasive aspergillosis, which can be regarded as group of 'difficult or atypical' patients, is likely to affect the observed diagnostic accuracy of a test. Also, the exclusion of any other of the reference standard groups may affect the accuracy of the index test. We therefore excluded studies explicitly excluding one of the four categories of patients from the review, as well as studies in which it is not clear how many patients with proven, probable, possible or no invasive aspergillosis had positive or negative index test results.
HETEROGENEITY	Three sources of heterogeneity: effect of cut-off value, effect of the reference standard and existence of clinical subgroups.	Same
CONCUSIONS (taken from the abstract)	Using the test at a cut-off value 0.5 ODI in a population with a disease prevalence of 8% (overall median prevalence): -sensitivity 78%, 22% false negatives -specificity of 81%, 19% false negatives Using the test at cut-off value 1.5 in the same population: - sensitivity 64%, 36% false negatives	Using the test at a cut-off value 0.5 ODI in a population with a disease prevalence of 9% (overall median prevalence): -sensitivity 82%, 18% false negatives -specificity 81%, 19% false negatives Using the test at cut-off value 1.5 in the same population: -sensitivity 61%, 39% false negatives

	- specificity of 95%, 5% false negatives. These numbers should however be interpreted with caution, because the results were very heterogeneous.	-specificity 93%, 7%false negatives. These numbers should, however, be interpreted with caution because the results were very heterogeneous
BIAS	QUADAS	QUADAS 2
Summary of Findings (SoF)	Yes	Yes
Title of first version: OPTICAL COHERENCE TOMOGRAPHY (OCT) FOR DETECTION OF MACULAR OEDEMA IN PATIENTS WITH DIABETIC RETINOPATHY [22,23]		
YEARS	May 2011	June 2013
N. STUDIES	9; 768 subjects; 1,325 eyes	10; 830subjects; 1,387 eyes
OBJECTIVE	To determine the diagnostic accuracy of OCT for detecting diabetic macular oedema (DMO) and clinically significant macular oedema (CSMO), defined according to ETDRS 1985.	To determine the diagnostic accuracy ofOCT for detectingDMO and clinically significant macular oedema CSMO, defined according to ETDRS 1985, in patients referred to ophthalmologists after DR is detected. In the update of this review we also aimed to assess whether OCT might be considered the new reference standard for detecting DMO.
PATHWAY & ROLE	Measurements of retinal thickness may be obtained directly from the tomograms either by manually measuring the distance between the inner and outer retinal boundaries or by using computer image processing techniques. OCT is increasingly used for detecting macular oedema in people with DR because it is an objective and reliable tool. Furthermore, OCT allows a quantitative follow up of the effects of treatment. However, purchasing an OCT machine is costly and personnel are needed to use it. OCT is unlikely to be used bu primary care professionals as a triage test to detect DMO; OCT is mainly used by secondary	In the updated version of this review, we acknowledge that the clinical pathway of patients with DMO is unclear and probably dependent on the country and setting. Thus, the applicability of the results of the review will depend on patient selection in included studies, such as inclusion criteria and results of prior testing.

	care professionals to further investigate patients who are suspected of having macular oedema. As such it would be used by an ophthal- mologist as an add-on test, to assess the need for laser treatment by recording macular thickness.	
INDEX	The index test was OCT, regardless of the generation of development of the instrument (low or high resolution, three-dimensional or spectral-domain OCTs).	The index test was OCT, regardless of the generation odevelop-ment of the instrument (low or high resolution, three-dimensional or spectral-domain OCTs).Despite the fact that retinal thickness measurements with OCT have been compared to those obtained with the Retinal Thickness Analyzer in at least one study, based on their best knowledge, authors believed that such a comparison is no longer of interest given the dominant use of OCT devices. Authors were not aware of any other instruments that can be compared to OCT.
REFERENCE	In the ETDRS study DMO was defined on the basis of stereoscopic fundus photography (ETDRS 1985). This technique is complicated and difficult to use in a clinical setting. It was replaced by contact fundus biomicroscopy, which was found to be in close agreement with stereophotography, particularly for CSMO. Non-contact fundus biomicroscopy is more commonly used, since sophisticated fundus lenses have been proposed for binocular fundus observation during the past two decades, yet it has been shown to be slightly less sensitive than contact fundus biomicroscopy. Finally, valid reference tests considered in this review were stereoscopic fundus photography and contact lens or non-contact lens biomicroscopy of the fundus.	Same In addition: In the update of this review, authors acknowledge that OCT is increasingly thought of as a new reference standard for DMO and will not update the review further. Although the American Academy of Ophthalmology's Preferred Practice Patterns (AAO PPP 2012) still considers clinical examination as the current recommendation for routine diagnosis of DMO, Schneider 2013 found that the use of OCT has greatly increased for patients with neovascular age-relatedmacular degeneration or DMO in recent years, while that of fluorescein angiography or fundus photography has decreased.
HETEROGENEITY	Heterogeneity related to retinal thickness cut-off, to index test, to reference standard, to characteristics of the study population, to methodological study quality items of the QUADAS checklist.	Same

CONCUSIONS	Central retinal thickness measured with OCT cannot be used as a stand-alone test to diagnose the central type of CSMO and decide on the use of laser photocoagulation in patients who are referred to retina clinics. In fact, there is a substantial disagreement ofOCT with the ETDRS definition of CSMO based on clinical examination. Some researchers have observed that OCT can detect macular thickening earlier than clinical examination, but also found that such cases did not necessarily progress to CSMO and need photocoagulation. Care should be taken in applying the conclusions of this review to other test-treatment pathways. In fact, OCT will become an essential tool to manage antiangiogenic therapy, an expanding therapeutic option for patients with macular oedema due to DR, because OCT is a component of the diagnostic algorithms of studies on this new treatment.	Using retinal thickness thresholds lower than 300 μm and ophthalmologist's fundus assessment as reference standard, central retinal thickness measured with OCT was not sufficiently accurate to diagnose the central type of CSMO in patients with DR referred to retina clinics. However, at least OCT false positives are generally cases of subclinical DMO that cannot be detected clinically but still suffer from increased risk of disease progression. Therefore, the increasing availability of OCT devices, together with their precision and the ability to inform on retinal layer structure, now make OCT widely recognised as the new reference standard for assessment of DMO, even in some screening settings. Thus, this review will not be updated further.
BIAS	QUADAS	QUADAS 2
SoF	Yes	Yes
Title of first version: THE DIAGNOSTIC ACCURACY OF THE GENOTYPE® MTBDRSL ASSAY FOR THE DETECTION OF RESISTANCE TO SECOND-LINE ANTI-TUBERCULOSIS DRUGS [24,25]		
YEARS	January 2014	September 2015
N. STUDIES	21	27
OBJECTIVE & ROLE	• Primary objectives: -To assess and compare the diagnostic accuracy of MTBDR*sl* for the detection of resistance to fluoroquinolones (FQs) in patient specimens (using direct testing) and culture isolates (using indirect testing) confirmed as tuberculosis (TB) positive. -To assess and compare the diagnostic accuracy of MTBDR*sl* for the detection of resistance to second-line injectable drugs (SLIDs) in patient specimens (using direct testing) and culture isolates (using indirect testing) confirmed as TB positive. -To assess and compare the diagnostic accuracy of MTBDR*sl* for the detection of extensively drug-resistantTB (XDR-TB) in patient specimens (using direct testing) and culture isolates (using indirect testing) confirmed as TB positive.	• Primary objectives: Same. In addition: The populations of interest were people with MDR-TB or rifampicin resistant TB, which is considered a proxy for MDRTB in high burden settings. • Secondary objectives: Same. In addition: Subsequent to the published protocol, we added an investigation of heterogeneity in relation to microscopy smear grade.

	• Secondary objectives: -To investigate heterogeneity in relation to the reference standard (culture-based drug susceptibility testing (DST) compared with: (1) genetic sequencing (2) culture-based DST and genetic sequencing (3) culture-based DST followed by genetic sequencing with discordant results) and individual drugs within a drug class (for example, ofloxacin and moxifloxacin within the FQ class). Authors also prespecified in the protocol investigations of heterogeneity in relation to HIV status, condition of the specimens (fresh or frozen, volume of specimen), patient population (patients suspected of having MDR-TB or XDR-TB) and whether WHO-recommended critical drug concentrations were used for culture-based reference testing.	
PATHWAY	Depending on the setting, DST is either performed on all patients with confirmed TB or only on patients who are clinically suspected of having DR-TB(for example, if the patient's symptoms have failed to improved on first-line therapy, or if they still have viable bacilli in their sputum after an extended period of treatment. As mentioned above, the manufacturer recommends that if the patient specimen (usually sputum) is smear-positive the assay be performed directly on the specimen (direct testing). If smear-negative, it is recommended that the assay be performed on the culture isolate grown from the patient specimen (indirect testing). DST for resistance to the second-line drugs is only performed if resistance to the first-line drugs is confirmed. Where routine molecular (genotypic) testing is well established, phenotypic DST is not usually performed. However,authors expected research studies evaluating the accuracy of molecular DSTs, such as the MTBDRsl assay, to almost always include phenotypic DST as a reference standard. Furthermore, authors also	Depending on the setting, DST is either performed on all patients with confirmed TB or on patients who are clinically suspected of having drug-resistant TB (for example, if the patients' symptoms have failed to improve on first-line therapy, or if they still have M. tuberculosis bacilli in their sputum after an extended period of treatment). DST for resistance to the second-line drugs is usually only performed if resistance to the first-line drugs is confirmed. Specifically, a patient with suspected drug-resistant TB provides a specimen (usually sputum), which is examined by smear microscopy. If smearpositive, MTBDR*sl* version 1.0 or version 2.0 can be performed directly on the specimen. If smear-negative, MTBDR*sl* version 1.0 should not be performed directly on the specimen, but rather on the culture isolate. MTBDR*sl* version 2.0 may be performed directly on a smear-negative specimen. A molecular test for firstline drug resistance (for example, the MTBDR*plus* assay) may be performed prior to testing withMTBDR*sl* if resistance to the firstline drugs is yet to be

	expected some studies to use genetic sequencing to resolve any discordant index test-reference standard results.	confirmed. Phenotypic DST may still be performed on culture-positive isolates.
COMMENT		This updated systematic review summarizes the current literature and includes 27 studies and integrates six new studies: five new studies forMTBDR*sl* version 1.0 identified since the original Cochrane review, and one study for MTBDR*sl* version 2.0. For MTBDR*sl* version 1.0, the findings in this updated review are consistent with those reported in the previous version of the review.
INDEX	Studies that evaluated the MTBDR*sl* assay were included. MTBDR*sl* would be used as an initial test replacing phenotypic culture-based DST as the initial test.	The index test was MTBDRsl version 1.0 or version 2.0. Comment: 1 study on version 2.0. The role of MTBDRsl would be as the initial test, replacing culture-based DST, for detecting second-line drug resistance.
REFERENCE	1. Phenotypic culture-based DST: solid culture or a commercial liquid culture system (BACTEC 460, MGIT 960 and MGIT Manual System, Becton Dickinson, USA) incorporating the drug of interest. It is the conventional reference standard, but it is considered to be imperfect and is dependent on the drug concentration threshold used to define resistance. 2. Genetic sequencing of the *gyrA* or *rrs* genes, or both. Genetic sequencing is considered to be more accurate than phenotypic culture-based DST; however, this is only if it targets all known resistance determining regions, which are not completely defined for the FQs and the SLIDs. Therefore,genetic sequencing can miss mutations that may cause drug resistance which fall outside of the targeted genes. Furthermore, genetic sequencing is usually applied only to culture isolates when results for the index test and the culture-based reference test do not agree. In this latter situation, there is potential for verification bias because the same reference standard is not being used to verify all index test results. 3. Two reference standards used together: phenotypic culturebased DST and genetic sequencing of the same samples. If a specimen was resistant according to phenotypic culture-based	1. Same 2. Sequencing of the *gyrA* or *rrs* genes (MTBDR*sl* version 1.0) or additionally the *gyrB* and *eis* promoter regions (MTBDR*sl* version 2.0). Sequencing is considered to be more accurate than culture-based DST; however, this is only if it targets all known resistance-determining regions, which are not fully known for the FQs and the SLIDs. Therefore, targeted sequencing may miss mutations that cause drug resistance. 3. Same 4. Same

	DST or had a mutation in the *gyrA* or *rrs* genes, the specimen was classified as having the target condition. If both phenotypic culture-based DST and genetic sequencing indicated susceptibility, the specimen was classified as not having the target condition. 4. Two reference standards used sequentially: phenotypic culture-based DST followed by selective testing by genetic sequencing of samples with discordant results (also referred to as discrepant analysis). Discordant results may be either index test positive/phenotypic culture-based DST negative or index test negative/phenotypic culture-based DST positive.	
HETEROGENEITY	Within each stratum (for example SLID resistance), heterogeneity was investigated through visual examination of forest plots of sensitivity and specificity. Then, if sufficient studies were available, we explored the possible influence of the following pre-specified categorical covariates: -reference standard (culture, genetic sequencing,culture and genetic sequencing, culture followed by genetic sequencing) -individual drug (amikacin, kanamycin and capreomycin).	Within each stratum (for example SLID resistance), heterogeneity was investigated through visual examination of forest plots of sensitivity and specificity. Then, if sufficient studies were available, we explored the possible influence of the following pre-specified categorical covariates: -reference standard (culture, genetic sequencing,culture and genetic sequencing, culture followed by genetic sequencing) -resistance to the following drugs: ofloxacin, moxifloxacin, levofloxacin, gatifloxacin, amikacin, kanamycin, and capreomycin -drug concentration used for culture based DST. In addition, for this updated review, authors added an investigation of heterogeneity in relation to microscopy smear grade.
CONCUSIONS (from the abstract)	- A positive MTBDR*sl* result for resistance to the fluoroquinolone drugs or the second-line injectable drugs is reliable evidence that the person has drug-resistant TB and further conventional drug-resistance testing is not required. -However, when the test reports a negative result, clinicians may still wish to carry out conventional testing.	-In people with rifampicin-resistant or multidrug-resistant tuberculosis, MTBDRsl performed on a culture isolate or smear-positive specimen may be useful in detecting second-line drug resistance. MTBDRsl (smear-positive specimen) correctly classified around six in seven people as having fluoroquinolone or SLID resistance, although the sensitivity estimates for SLID resistance varied. -However, when second-line drug resistance is not detected (MTBDRsl

		result is negative), conventional DST can still be used to evaluate patients for resistance to the fluoroquinolones or SLIDs. -Authors recommend that future work evaluate MTBDRsl version 2.0, in particular on smear-negative specimens and in different settings to account for different resistance-causing mutations that may vary by strain. -Researchers should also consider incorporating WHOrecommended critical concentrations into their culture-based reference standards.
BIAS	QUADAS 2	QUADAS 2
SoF	Yes	Yes
Title of first version: DIAGNOSTIC ACCURACY OF LAPAROSCOPY FOLLOWING COMPUTED TOMOGRAPHY (CT) SCANNING FOR ASSESSING THE RESECTABILITY WITH CURATIVE INTENT IN PANCREATIC AND PERIAMPULLARY CANCER 1) [26,27]		
YEARS	September 2012	May 2016
N. STUDIES	15; 1,015 subjects	16; 1,146 subjects
OBJECTIVES	• Primary objective: -To determine the diagnostic accuracy of diagnostic laparoscopy performed as an add-on test to CT scanning in the assessment of curative resectability in pancreatic and periampullary cancer. • Secondary objective: Authors planned to explore the following sources of heterogeneity: 1. Studies at low risk of bias versus those at unclear or high risk of bias. 2. Full text publications versus abstracts. 3. Prospective studies versus retrospective studies. 4. Proportion of patients with pancreatic cancer, ampullary cancer, and bile duct cancers. 5. Procedures performed under the same anaesthetic versus procedures performed under a different anaesthetic. 6. Different definitions for resectable cancer on laparotomy. 7. Additional pre-tests performed (besides CT scan).	Same
PATHWAY	There is no standard algorithmcurrently available for assessing the resectability of pancreatic and periampullary cancers, with	Same

	different clinicians following their own algorithms based on either their clinical experience or what they were taught. Currently, almost all algorithms include a CT scan as one of the tests. CT may be the only test performed before laparotomy. Other tests such as diagnostic laparoscopy, positron emission tomography (PET scanning), magnetic resonance imaging (MRI), or endoscopic ultrasound (EUS) may be used in addition to CT scan to assess resectability.	
INDEX	Only diagnostic laparoscopy in which histopathological confirmation of metastatic spread was obtained on a paraffin section, was included. Diagnostic laparoscopy can be considered as an add-on test to the CT scan prior to laparotomy done with the intention of performing a potentially curative resection.	Same
REFERENCE	Confirmation of liver or peritoneal involvement by histopathological examination of suspicious (liver or peritoneal) lesions obtained at diagnostic laparoscopy or laparotomy. Authors accepted only paraffin section histology as the reference standard. In clinical practice, depending on the urgency of the results, a frozen section biopsy, may be done to obtain immediate results. However, this is always confirmed by subsequent paraffin section histology (which can take several days) because frozen section biopsy is not as reliable as paraffin section histology. Authors also accepted the surgeon's judgement of unresectability at laparotomy when biopsy confirmation was not possible. For example, if the tumour has invaded the adjacent blood vessels the surgeon may not resect the tumour because of the danger posed by resecting part of a large blood vessel, and so biopsy confirmation cannot be obtained.	Same
HETEROGENEITY	Authors planned to explore heterogeneity by using the different sources of heterogeneity as covariate(s) in the regression model. However, this was not possible because the information was either not available or was the same in all the studies	Same

CONCUSIONS (taken from the abstract)	Diagnostic laparoscopy may decrease the rate of unnecessary laparotomy in patients with pancreatic and periampullary cancer found to have resectable disease on CT scan. On average, using diagnostic laparoscopy with biopsy and histopathological confirmation of suspicious lesions prior to laparotomy would avoid 23 unnecessary laparotomies in 100 patients in whom resection of cancer with curative intent is planned.	Diagnostic laparoscopy may decrease the rate of unnecessary laparotomy in people with pancreatic and periampullary cancer found to have resectable disease on CT scan. On average, using diagnostic laparoscopy with biopsy and histopathological confirmation of suspicious lesions prior to laparotomy would avoid 21 unnecessary laparotomies in 100 people in whom resection of cancer with curative intent is planned.
BIAS	QUADAS 2	QUADAS 2
SoF	Yes	Yes