Review

Andrew C. Don-Wauchope*, Karina Rodriguez-Capote, Ramy Samir Assaad, Seema Bhargava and Annalise E. Zemlin

A guide to conducting systematic reviews of clinical laboratory tests

https://doi.org/10.1515/cclm-2023-0333
Received April 1, 2023; accepted July 19, 2023; published online August 4, 2023

Abstract: Clinical laboratory professionals have an instrumental role in supporting clinical decision making with the optimal use of laboratory testing for screening, risk stratification, diagnostic, prognostic, treatment selection and monitoring of different states of health and disease. Delivering evidence-based laboratory medicine relies on review of available data and literature. The information derived, supports many national policies to improve patient care through clinical practice guidelines or best practice recommendations. The quality, validity and bias of this literature is variable. Hence, there is a need to collate similar studies and data and analyse them critically. Systematic review, thus, becomes the most important source of evidence. A systematic review, unlike a scoping or narrative review, involves a thorough understanding of the procedure involved and a stepwise methodology. There are nuances that need some consideration for laboratory medicine systematic reviews. The purpose of this article is to describe the process of performing a systematic review in the field of laboratory medicine, describing the available methodologies, tools and software packages that can be used to facilitate this process.

Keywords: evidence based laboratory medicine; GRADE; systematic review

Introduction

Clinical laboratory professionals have an instrumental role in supporting clinical decision making with the optimal use of laboratory testing for screening, risk stratification, diagnostic, prognostic, treatment selection, and monitoring of different states of health and disease. The quality and consistency of medical practice is increasingly becoming an important topic. This is reflected in many national policies to improve patient care using clinical practice guidelines or best practice recommendations. The development of these guidelines requires synthesis of the different scientific studies that build the evidence for clinical practice. A systematic review is an established part of guideline development and is an integral part of understanding the evidence that supports the clinical use of a test [1]. Implementation of laboratory test results into clinical practice requires appraisal of the available evidence. Systematic attempts to record observations in a reproducible unbiased fashion markedly increase the confidence one can have in knowledge about patient prognosis, the value of diagnostic tests and the efficacy of treatment [2]. The goal of evidence-based laboratory medicine is to improve clinical outcomes in patient care while ensuring the effective use of healthcare resources [3]. The development of the GRADE system [4] for evaluating evidence is an important part of the process for a systematic review, and the application of the GRADE scoring to evidence based laboratory medicine is a key component of clinical practice guideline development.

The scope of laboratory medicine includes a wide range of tests from different pathology disciplines and a variety of reasons for using each test [5]. Diagnostic test accuracy (DTA) should be evaluated for a question related to the specific purpose of a test [6]. For example, its efficiency to confirm or exclude a disease or to stratify prognosis or treatment response. The guidance given in this review cannot be comprehensive for all applications of laboratory medicine.
and, thus, focuses on aspects that are more common. This will enable a wider application across the spectrum of laboratory medicine disciplines.

Firstly, the use of the test can be described under a few categories. Diagnostic tests should be evaluated for diagnostic accuracy for the condition(s) where clinical utility in support of making diagnostic decisions is expected. An example is the use of natriuretic peptides for heart failure diagnosis [7, 8]. Monitoring tests can be evaluated for use in management or monitoring of disease or process. A well-known example is HbA1c in diabetes management [9]. Predictive tests can be used to predict a variety of outcomes to therapy or other interventions, e.g. companion diagnostics to predict disease free survival in colon cancer [10]. For early disease diagnosis, the concept of screening tests is applied, e.g. Human Papilloma Virus detection or cytology of cervical samples for early detection of cervical cancer [11]. Prediction of clinical outcomes tests are used on their own or in combination with other factors, e.g. the Framingham Risk Score to predict cardiovascular outcome [12].

Secondly, a comparison between two tests is often made to determine which test is better at determining the diagnosis or desired clinical outcome. One of the tests may be accepted as the standard, but there are other possibilities. For example, the test can be evaluated against a defined outcome, often considered a higher level of evidence [13]. In these situations, comparison between tests is not a requirement. Tissue tests from anatomical pathology are often the gold standard for diagnostic comparison in systematic reviews that are not specifically looking at laboratory diagnostic testing. However, the diagnostic accuracy of these gold standard tests has not always been confirmed in outcome studies.

Even though there are several reviews on the topic of how to perform a systematic review for diagnostic testing [6, 14–17], there is a gap in this form of evidence to support the use of laboratory medicine testing for many tests and diseases. Thus, laboratory professionals should be seeking opportunities to improve the evidence synthesis.

The Cochrane Collaboration formed a DTA working group in the late 1990s and started publishing on-line guides in 1996. The Cochrane Methods Group on Systematic Review of Screening and Diagnostic Tests, had provided guidance on doing systematic reviews of diagnostic tests on the Cochrane website [18]. A review of the systematic reviews of diagnostic tests published in 2000, pointed out that many systematic reviews did not meet the specified standards [19]. The quality of systematic reviews published since then has been inconsistent and many of them have not used appropriate methodology [20]; however some improvement has been noted since the publication of Standards for Reporting Diagnostic accuracy studies (STARD) guidelines [21].

A helpful development in the systematic review field is the availability of software packages that assist in the processing of systematic review. These include tools to support searching for literature, screening the literature and analysing the data [22]. An example of a commercial software package is available at the following link. https://www. evidencepartners.com/resources/guides-white-papers/buyers-guide-to-systematic-review-software [23]. There are several websites with tools that are useful in this area. University libraries may have a resource page similar to the one at University College London https://library-guides.ucl.ac.uk/systematic-reviews/software [24]. The systematic review toolbox is an on-line catalogue that provides a number of resources http://systematicreviewtools.com/index.php [25].

This review is aimed at providing the basic information required to perform a systematic review in laboratory medicine. A recent publication described the process of systematic review in 24 steps [26] and we have adapted this approach to the clinical laboratory and cross referenced the Cochrane Guide to Systematic Reviews in Diagnostic testing [27].

Setting the scenario for the systematic review

Define the research question

Ideas for research questions arise from many sources, including inquiries received by the clinical laboratory professional. Transforming the idea into a good research question is the first and most critical step when performing a systematic review. Poorly framed questions make it difficult to design the search strategy, decide on inclusion and exclusion criteria, or make appropriate conclusions from the data. To formulate a good question the use of a structured framework such as the Population, Intervention (or Index test), Comparator, Outcome, Time and Setting (PICOTS) criteria (Figure 1) should be defined [28–30]. There are several other mnemonics which can be used in place of PICOTS, e.g. SPIDER [31], SPICE [32], etc. which are better suited for research questions dealing with qualitative analysis, or ECLIPSE [33] which is suited for health policy/management information. The value of a well-structured question using a framework such as PICOTS keeps the focus on the question at hand and helps to prevent diluting the information with interesting, but only semi-related studies [34].
Before initiating a detailed literature search of the defined research question, it is pertinent to ascertain that this research question has not already been successfully dealt with. A preliminary search of the literature will provide an idea about the amount, scope and depth of the literature pertaining to the question. Critically appraising any previously conducted systematic reviews against the defined question will allow the team to establish if an update or repeat evaluation is required. As of 2002, the National Center for Biotechnology Information (PubMed) database has a systematic review search filter so that citations for this type of article can be more easily identified (filter terms meta-analysis and systematic review under type of article). Cochrane Collaboration [35] has many diagnostic accuracy systematic reviews in the database [36]. Using federated search engines such as ACCESSS (ACCESSSS | Home) [37] and Trip Medical Database (https://www.tripdatabase.com/) [38] will help identify systematic reviews related to the specific question. The PROSPERO https://www.crd.york.ac.uk/prospero/ platform is one registry of systematic reviews that are in progress [39].

If the question has not already been answered, it should now be edited with the knowledge gained from the initial literature search. The refined question should clearly identify the population, the diagnostic test investigated, the reference or comparator test, and the outcome of interest [40]. Other aspects relevant to testing (pre-analytical and post-analytical) such as timing and the setting of the test should be considered.

The question may address etiology, diagnosis, prognosis, treatment, or prevention. Some possible areas to focus a question are:
- analytical performance of a test e.g. impact of test imprecision or timing on outcome
- mode of delivery of a test e.g. use of point-of-care testing in reducing time of diagnosis
- use of a test for diagnosis e.g. rule in or rule out
- use of a test for prognosis i.e. predicting clinical outcomes
- use of a test for treatment selection i.e. personalised medicine
- use of test for treatment optimisation and compliance

Table 1 provides some examples of PICOTS criteria in a selection of systematic reviews. In one example, a pathology group was interested in the use of imprint cytology to examine the diagnostic performance of sentinel node(s) in women with breast cancer, compared to paraffin section...
examination to identify women at risk of metastatic breast cancer [41]. They set Population, Intervention (or Index test), Comparator and Outcome (PICO) criteria (see Table 1 row 8). The question formulated was: In patients undergoing breast biopsy, what is the diagnostic performance of imprint cytology compared to paraffin embedded sections for the diagnosis of metastatic disease [41].

When there is only one comparator test and one outcome of interest, the research question is simpler. However, when there are several comparator tests and several outcomes, then the question may be single, but the objectives would be several. For example, when a group wants to evaluate the diagnostic and prognostic value of procalcitonin in patients with sepsis [42], the details

Table 1: Descriptions from some examples of systematic reviews of the different population, index test, comparator test, outcome, timing and setting.

<table>
<thead>
<tr>
<th>Study</th>
<th>P</th>
<th>I</th>
<th>C</th>
<th>O</th>
<th>T</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smedemark et al. [44]</td>
<td>People presenting with symptoms of acute respiratory infections</td>
<td>CRP (POCT) guided antibiotic therapy</td>
<td>Guideline based antibiotic administration</td>
<td>1. Participant recovery in seven day follow up</td>
<td>At presentation</td>
<td>Primary care</td>
</tr>
<tr>
<td>Haraka et al. [76]</td>
<td>People being investigated for tuberculosis</td>
<td>Xpert MTB/RIF</td>
<td>Sputum smear microscopy</td>
<td>1. Successful treatment outcome</td>
<td>With sputum collection</td>
<td>Secondary or tertiary care or surveillance</td>
</tr>
<tr>
<td>Colli et al. [77]</td>
<td>Adults with chronic liver disease</td>
<td>AFP (cut-off 20 ng/mL) US abdomen AFP + US</td>
<td>Diagnosis of hepatocellular carcinoma</td>
<td>1) Mortality at up to 28 days; 2) time receiving antimicrobial therapy; 3) change in antimicrobial regimen (broad to narrow spectrum)</td>
<td>Intensive care</td>
<td></td>
</tr>
<tr>
<td>Andriolo et al. [42]</td>
<td>Adults with a diagnosis of sepsis, severe sepsis or septic shock</td>
<td>PCT (at least one measurement)</td>
<td>Three different possible groups i) standard methods to diagnose and stage sepsis; ii) serum PCT evaluation or PCT-guided therapy; iii) other biomarkers (e.g. CRP, interleukins etc.)</td>
<td>1) Mortality at up to 28 days; 2) time receiving antimicrobial therapy; 3) change in antimicrobial regimen (broad to narrow spectrum)</td>
<td>Not defined</td>
<td>Intensive care</td>
</tr>
<tr>
<td>Nagar et al. [44]</td>
<td>Infants ≥34 weeks of gestational age receiving phototherapy or in the post-phototherapy phase</td>
<td>Transcutaneous bilirubin (TcB) Total serum bilirubin (TSB)</td>
<td>Agreement statistic between TcB and TSB measurements provided either as correlation coefficient or as the mean and standard deviations of absolute differences in bilirubin values by the two methods</td>
<td>During phototherapy or in the post-phototherapy phase</td>
<td>Not defined</td>
<td>Intensive care</td>
</tr>
<tr>
<td>Sethi et al. [78]</td>
<td>Patients suspected of ACS</td>
<td>s/hsT I/T</td>
<td>Conventional troponin</td>
<td>Diagnosis of ACS or AMI</td>
<td>At the time of presentation</td>
<td>Intensive care</td>
</tr>
<tr>
<td>Matthaiou et al. [79]</td>
<td>Critically ill adults with suspected or proven sepsis</td>
<td>PCT-guided antibiotic therapy</td>
<td>Empirical or guideline based antibiotic administration</td>
<td>–</td>
<td>Duration of antibiotic therapy for first episode of infection</td>
<td>Not defined</td>
</tr>
<tr>
<td>Tew et al. [41]</td>
<td>Patients undergoing examination of sentinel nodes during breast biopsy</td>
<td>Imprint cytology Paraffin embedded sections</td>
<td>Diagnosis of metastatic disease</td>
<td>Mortality at 28 days</td>
<td>Diagnosis of metastatic disease</td>
<td>Intensive care</td>
</tr>
</tbody>
</table>

P, population; I, intervention; C, comparator; O, outcome; T, timing; S, setting; CRP, C-reactive protein; POCT, point of care test; MTB, mycobacterium tuberculosis; RIF, resistance to rifampin; AFP, alpha fetoprotein; US, ultrasound; ACS, acute coronary syndrome; s/hsT, sensitive/high sensitivity troponin; AMI, acute myocardial infarct; PCT, serum procalcitonin.
(see Table 1 row 4) demonstrate multiple possible combinations of Index test to comparator test and the review reported on five different combinations. The outcomes are also categorised into primary and secondary and the five comparator groups are assessed against the multiple outcomes [42].

**Build a team**

Systematic reviews are not solitary endeavours and a team should aim to gather the following range of skills.

a) **Content expert.** These team members know the area of medicine of interest. For a review of a diagnostic laboratory test, a mix of clinicians and laboratory physicians or scientists will provide the wide range of expertise needed.

b) **Process expert.** These team members know the process and pitfalls of systematic reviews.

c) **Librarian.** A professional librarian can be very helpful in designing and performing the initial search for candidate studies.

d) **Statistical experts.** A statistician will be valuable in analysing the data and making inferences. If pooling the data using meta-analysis is planned, then statistical advice is essential to guide the appropriate data analysis with consideration of reporting and other bias.

e) **Support.** Large reviews will require support staff to help with the acquisition of papers, extraction of data and record keeping. Smaller studies can often be managed by one or two interested investigators.

Knowing that the team, have the expertise and the ability to avoid bringing their pre-conceived bias to the project is important. Some expert reviewers to support the team may also be helpful and should be consulted at key stages of the process to ensure that the systematic review is conducted to the standards that will allow publication. These people are more likely to be content experts, but the addition of a process or method expert is also helpful.

**Define criteria for selection of studies**

The research question is now set and the selection criteria along with inclusion and exclusion criteria and timing and setting components need to be established. In the examples (Table 1), the population defined is often broad and there could be modulating factors. Hence, it is important to clearly define the inclusion and exclusion criteria to get papers with data that can be summarised and combined for potential meta-analysis. With respect to population, it may be important to define age sex, health status, presence, or absence of specific medical condition. The methodology of the index and reference tests should be defined as clearly as possible. A systematic review of the accuracy of transcutaneous bilirubin in neonates exposed to phototherapy is an example of well formulated inclusion and exclusion criteria for population, intervention and comparator test. The inclusion criteria were primary studies or case series with 20 or more human newborns (≥34 weeks); transcutaneous bilirubin measurement on all subjects; serum or whole blood total bilirubin on all subjects; and a comparison between those two methods [43]. In a systematic review looking at the use of point of care C-reactive protein, several clinical outcomes were well defined. These included the number of participants given an antibiotic prescription at the index consultation; the number of participants given an antibiotic prescription within 28 days; clinical recovery within seven days; mortality within eight days; the number of participants needing hospital admission within 28 days; and clinical recovery within 28 days. The timing is specified, and the setting is defined as primary care elsewhere [44].

**Create the data collection tool**

Planning the data collection helps in establishing the selection criteria as well as defining the search strategy. The Data collection tool can be built in several different applications such as spreadsheets, database applications or specific software such as Covidence (Veritas Health Innovation Ltd, Melbourne, Australia) or DistillerSR (Distiller SR Inc, Ottawa, Canada). Collecting all the details for the PICO components, inclusion and exclusion criteria, the study design and methods, and study quality is important. This process should be clearly defined so that it can be used by a wide range of potential data extractors. The tool should be tested out on a few of the studies that the team has used to formulate the question.

**Design the search strategy**

The components of the PICO question define the key words for an effective search strategy.

There are several approaches to locating candidate papers [45] including; electronic searching of the databases of publications [46]; searching the reference lists of relevant papers; communication with well-known authors in the field of interest. The use of the so called ‘grey literature’ includes, but is not limited to, non-peer reviewed
publications (conference proceedings, letters to the editor), product monographs, presentations, and personal communications. In the field of laboratory medicine and diagnostic testing, this ‘grey literature’ may well form a significant source of appropriate information and a decision to use or exclude this must be made early on [47].

Searching the databases of published papers to locate the candidate papers of interest usually requires the assistance of a librarian. There is no single search term that can be used to denote diagnostic studies [45] and several trial searches may be necessary to refine the search terms. Appropriate selection of the key words and use of the Boolean operators (AND and OR) will ensure that the search is wide enough to capture all relevant studies. Terms that are too narrow may exclude appropriate studies, whereas terms that are too general will yield large numbers of irrelevant papers. The most widely used databases, PUBMED, MEDLINE, EMBASE, and GoogleScholar have significant overlap but there are papers that may be missed by using only one database [35].

Getting the systematic review started

Document and register the study protocol

A well written protocol is helpful for the investigators as well as for registration with a database that records the systematic reviews that are being conducted. The PRISMA-P https://prisma-statement.org/Extensions/Protocols [48] provide a list of items that should be included in the review protocol [49]. To ensure clarity, the Cochrane training guide recommends using experts to review the protocol to make sure that it is understandable and complete. Registering the protocol with a platform such as PROSPERO https://www.crd.york.ac.uk/prospero/ [50] is important [39].

Perform the literature search and build the database

Using the criteria defined for the PICO question each of the selected databases must be searched. Defining a search strategy in each of the databases is required, as these are not identical. The search results of should be saved in a file format that is compatible with the selected software. Import the saved files into a software application such as Covidence, DistillerSR or Rayyan (Rayyan Systems, Inc, Cambridge MA, USA) or into a reference manager software such as EndNote™ (Clarivate PLC, London, UK) or Zotero (Corporation for Digital Scholarship, Vienna, VA, USA). Each package will require a specific file format, and this should be included in the study protocol. The software applications provide instructions about exporting the files in the correct format from the searchable databases for import.

Identify and remove duplicates

The use of multiple databases may result in replication of the citations imported into the software application for managing the references and a defined strategy and systematic approach to the removal of duplicates will save time [51].

Reviewing the evidence

Title and abstract review

The vast majority of the comprehensively collected studies can be excluded by applying the inclusion and exclusion criteria in a title and abstract screen. A screening questionnaire that itemises the inclusion and exclusion questions is a good tool for documenting this phase. This tool should be tested on a few articles prior to full implementation. Further screening of the full text of remaining articles will identify those papers that meet the criteria for the review. The screening phase should be conducted by two independent reviewers to increase reliability. The title and abstract should be screened simultaneously and if title only is available, then these should move forward to the next phase of screening with the exception of those clearly meeting the exclusion criteria. Software packages that facilitate this have been mentioned in the introduction and artificial intelligence (AI) assisted learning applications such as ASReview https://asreview.nl/ [52, 53] are available to assist with sorting the references into ones that are more likely to meet the selection criteria. All screened articles that are selected by both primary reviewers will proceed to the full text review. Discrepancies can be adjudicated by a third reviewer who should be an expert in the field. Alternatively, the two reviewers can meet and agree on the final decision.
Collect full texts using predefined collection criteria

All the relevant manuscripts can be collected in the reference manager or via other software applications. A librarian and library access to electronic archives of published manuscripts enables this. Some articles require the assistance of library resources to locate the full text through interlibrary loan. In some instances, authors may need to be contacted to obtain a copy of the full text. The screening questionnaire designed for title and abstract review can be enhanced with the additional details expected to be found in the full text for the inclusion and exclusion criteria. The enhanced questionnaire can be documented in the software package to facilitate the selection of articles that meet the criteria. This phase of screening is also done by two independent reviewers, with discrepancies managed as described previously. Studies that meet the criteria move forward to the data collection phase while those that are excluded by both reviewers are removed. The reasons for exclusion are documented for the report.

Identify grey literature with the expert reviewers

The expert reviewers (members of the systematic review team) may be able to identify ongoing research and grey literature that is not easily identified in the standard databases. They would also be able to review the collected set of manuscripts to ensure that no important studies are missed and that the best report is included in the final data set. Additional references should be placed into a data set and run through the process from Title and Abstract review.

Confirm the completeness of the literature collection

Once the updated list is available it is helpful to look for publications that have been either cited in or that cite the included list. Although these may have already been through the screening process this is a good check for completeness of the data set. Using a citation database can facilitate this action as many of these now include the citation data e.g. Scopus and Web of Science. Previous reviews and systematic reviews should also be checked for references to make sure that these have been put through the screening process. Any new manuscript identified needs to be fully screened from Title and Abstract review.

Identify the included literature selection

The final list of manuscripts should now be defined. Each stage of the screening process should be documented in a flow chart that describes the different inclusion and exclusion stages, the number of manuscripts reviewed at each stage and the reasons for exclusion. This is an important part of any published systematic review. Figure 2 is an example of a flow chart.

Figure 2: An example of a flow chart. Used with permission from Smedemark SA, Aabenhus R, Llor C, Fournaise A, Olsen O, Jørgensen KJ. Biomarkers as point-of-care tests to guide prescription of antibiotics in people with acute respiratory infections in primary care. Cochrane Db Syst Rev. 2022;2022(10):CD010130. Copyright © 2019 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.
Processing the evidence

Data extraction

Extraction of the data is best done using the pre-determined series of questions in the data collection tool. Easily retrievable information such as country of origin, year of publication, details of study design and recruitment can be extracted by content experts as well as non-experts (students, support personnel), whereas detailed content specific information is best extracted by either content or process experts on the team. Each item should be independently extracted by two people and then matched. Discrepancies should be corrected by consensus or a third person. Data should be collected in a consistent way which may include the need to convert units of measurement, maintain consistent definitions, abbreviations etc. The use of experts to guide the collection of data is particularly important if there are known analytical method differences. The content should be collected in a way that facilitates the analysis.

Risk of bias evaluation for individual studies

The oft quoted phrase of “garbage in-garbage out” applies equally to systematic reviews. Grouping and pooling data from several manuscripts can only be done if the quality of the underlying papers is considered in the appropriate context. The methodological quality of each paper selected for inclusion is critically appraised for potential sources of bias and variation that may have influenced their estimates of test accuracy. The applicability to the review question is concurrently evaluated. There are guidelines and tools such as the Quality Assessment tool for Diagnostic Accuracy Studies (QUADAS/QUADAS-2) [54, 55] that are useful in designing the assessment of quality.

Ideally, studies evaluating diagnostic tests must include the characteristics of the study in enough detail that readers can evaluate the quality of the data and conclusions. The STARD criteria provide a checklist of 25 items, which greatly improve the quality of studies evaluating diagnostic tests [56]. STARD was updated in 2015 with additional five criteria bringing the total to 30 with some revisions to other criteria [57]. However, poor reporting of original diagnostic test accuracy research remains a challenge [20].

The QUADAS-2 is the current version of the QUADAS tool, developed by Whiting et al. [55]. This tool consists in a checklist of questions or items structured into four key domains, each of these rated in terms of risk of bias and applicability to the research question.

Each question is answered as yes, no, or unclear. Depending on the diagnostic test under evaluation, some of the QUADAS questions will be more relevant than others in assessing study quality. It is not necessary to use all the questions, but only those applicable to the study. In a systematic review one should, however, use the same questions on all included papers. An overall or cumulative score is not recommended [58]. Table 2 is a previously published example of a QUADAS-2 table. Since 2011, QUADAS-2 has been adopted widely and applied in reviews of diagnostic accuracy. There has been some concern raised regarding QUADAS-2 by some authors. Schueler et al. indicated a limitation associated with calculating inter-rater agreement only on the domain questions [59]. Cook et al. felt that the tool was not able to discriminate between poorly and strongly designed studies [60], and that the QUADAS-2 offered no obvious advantage over the original 14-item QUADAS [54]. They further criticized the purposively qualitative nature of the QUADAS-2, which does not recommend scoring a study using a numeric value, a fundamental quality of assessment scales [60].

There are also adaptations of the tool for other types of study such as QUADAS-C for comparative diagnostic accuracy studies [61] and QUAPAS for prognostic accuracy studies [62].

Regardless of the tool used, the risk of bias and applicability to the research question should be clearly reported for each included study, summarised, and discussed in the systematic review report. The tool used should be clearly reported. Authors need to define in the protocol what quality items they will include and how these items will be used specific to their review. Assessments should be undertaken by at least two authors, and there should be an explicit procedure to resolve disagreements. Table 3 describes the different types of bias that should be considered.

Data analysis

Studies of diagnostic tests involve comparing the test result in groups of patients with and without the outcome of interest. The outcome of interest is determined by an independent reference or “gold standard” test. In this context the test results are classified into dichotomous groups based on expected performance for diagnosis. Systematic review allows for the development of more widely applicable 2 × 2
tables and the calculation of false positive (FP), true positive (TP), false negative (FN), true negative (TN), positive predictive values (PPV), negative predictive values (NPV), likelihood ratios (LR), diagnostic odds ratios (OR) and receiver operator characteristics (ROC) curves.

Sensitivity refers to the proportion of individuals with the outcome who return a positive test (TP/TP + FN).

Specificity refers to the proportion of individuals without the outcome of interest who return a negative test (TN/TN + FP).

Likelihood ratios refer to the ability of positive and negative tests to rule in or rule out the outcome of interest. The describe how many more times the test result is likely to be found in patients with the outcome than in a patient without the outcome of interest. Ratios greater than 10 or less than 0.1 provide convincing evidence, those greater than 5 or less than 0.2 provide strong evidence. Ratios less than 5 or greater than 0.2 provide only equivocal evidence.

A convenient nomogram by Fagan [63] that will convert pre-test probability and likelihood ratio into post-test probability, has been reprinted in numerous publications and textbooks.

In the context of using laboratory test for screening, prognosis, and monitoring of disease the studied outcomes are more complicated, especially those evaluating predictive health outcomes. In such studies, other calculations and metrics are used, such as overall survival (OS), disease free survival (DFS), hazard ratio (HR), odds ratio (OR) and relative risk reduction (RRR). There may be health economic data that could be captured for example cost effectiveness, admission and/or readmission rates.

### Table 2: Example of a published QUADAS-2 table.

<table>
<thead>
<tr>
<th>Study</th>
<th>Risk of bias</th>
<th>Concerns about applicability</th>
<th>Index text</th>
<th>Risk of bias</th>
<th>Concerns about applicability</th>
<th>Reference standard</th>
<th>Time and flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalona 2003</td>
<td>High</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
</tr>
<tr>
<td>Catalona 2011</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Fillea 2014</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Guazzoni 2011</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Ito 2013</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Jansen 2010</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Khan 2003</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Lazzeri 2012</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Lazzeri 2013</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Le 2010</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Maerin 2014</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Mikolajczyk</td>
<td>High</td>
<td>Unclear</td>
<td>High</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Miyakubo 2011</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
</tr>
<tr>
<td>Ng 2014</td>
<td>High</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Sokoll 2008</td>
<td>High</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Sokoll 2010</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
</tr>
<tr>
<td>Stephan 2009</td>
<td>Low</td>
<td>High</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
</tr>
</tbody>
</table>


### Table 3: Types of bias.

<table>
<thead>
<tr>
<th>Types of bias</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population selection bias</td>
<td>Recruiting online and biasing those without internet</td>
</tr>
<tr>
<td>Statistical bias</td>
<td>Poor study design</td>
</tr>
<tr>
<td>Verification/performance bias</td>
<td>No gold standard/differences in care</td>
</tr>
<tr>
<td>Citation bias</td>
<td>Positive results more likely to be cited</td>
</tr>
<tr>
<td>Publication bias</td>
<td>Positive results more likely to be published</td>
</tr>
<tr>
<td>Language bias</td>
<td>Studies more likely to be published in English/more likely to use studies published in own language</td>
</tr>
<tr>
<td>Database bias</td>
<td>Some databases more likely to index certain languages or journals</td>
</tr>
<tr>
<td>Truncation bias</td>
<td>Study published in shorter form with less details</td>
</tr>
<tr>
<td>Time-lag bias</td>
<td>Delayed publication of bias leading to outdated evidence</td>
</tr>
</tbody>
</table>

Decide about the appropriateness of meta analysis

One of the important decisions is whether to combine the data and perform a meta-analysis. Meta-analysis is a statistical technique for combining the results of independent studies that will result in a summary statistic [64]. The most important factor in deciding whether to combine studies is the heterogeneity between the studies. If the studies are very similar with respect to patient characteristics, study design, reference test, and outcomes, it makes sense to combine them. Hatala et al. provided an easy to read guide on how to approach combining studies [65]. Although the increased number of studies reflects the expanded interest for the test of concern, two studies is enough to perform a meta-analysis. There are key criteria that should be met for a meta-analysis [66]. All the individual study outcomes need to be comparable and described in a format that allows meaningful pooling. The interventions and comparators need to be similar enough and the data needs to be reported adequately for pooling. The detail for techniques of meta-analysis are beyond the scope of this manuscript. However, for example in a DTA meta-analysis, there are several types including the following [67]:

a) Univariate logistic regression without and then with random effects
b) Diagnostic odds ratio (OR) to give a combined measure of sensitivity and specificity.
c) Bivariate meta-analysis to ascertain within-study variability using the binomial likelihood model
d) Summary receiver operating characteristics (sROC) curves derived using regression models

In addition, there are ways to use modification of meta-analysis such as random-effects model or meta-regression [66, 68]. There are statistical packages and tools to perform the meta-analyses [69–71].

Thus, applying the right statistical tools and packages to the data of a systematic review can be a complicated process. This is where the team statistician will be required to guide the appropriate use of methods.

Describing the evidence

Report the findings

The report should describe the overall literature search, the populations included, interventions or investigations used, the comparators and the outcomes. The types of studies should be reported with the key characteristics and tabulated into sections that help the reader find relevant summary material. This should include any pertinent summary statistics such as OR, hazard ratios (HR), relative risk (RR), diagnostic accuracy etc. Where 2 × 2 data is available the relevant statistics should be reported. Figure 3 demonstrates a Forest Plot that is another common format for reporting summary findings.

Consider the impact of reporting bias

Reporting bias is prevalent as more studies with positive findings are published than those with negative findings. This selection bias is derived from both investigators and
journals for a range of different reasons. The Cochrane handbook and the GRADE handbook describe this in more detail. To try and avoid this, systematic review investigators need to look for grey literature and unpublished studies. The use of a funnel plot (Figure 4) is one way of assessing the opportunity for publication bias with either a visual assessment or the use of the Egger test [72].

A funnel plot assumes that studies with high precision will be plotted near the average and studies with low precision spread evenly on both sides of the average, creating an inverted funnel shape. A visually asymmetric funnel may reveal that studies with lower precision may even have greater effect (better diagnostic performance) than higher precision studies, denoting what is known as “small study effect”, discovering the reporting bias as the main reason behind it. In Figure 4 Open circles indicate smaller studies of inadequate quality with weaker evidence for effects. (a) symmetrical plot in the absence of bias; (b) asymmetrical plot in the presence of publication bias (smaller studies showing no beneficial effects are missing); (c) asymmetrical plot in the presence of bias due to low methodological quality of smaller studies whose results are biased towards larger beneficial effects.

The purpose of checking for reporting bias is to discuss the potential impact of the different factors that may lead to the bias and how it may impact the overall evaluation of the quality of the evidence reported in the systematic review.
Investigate heterogeneity in the results

Heterogeneity is the variability between results from different studies. Too much heterogeneity indicates that meta-analysis may be contra-indicated [66]. An assessment of the inconsistency of results is an important step. Understanding the reasons behind the heterogeneity of results can be explored with techniques such as sub-group analysis and/or stratification using different possible effect modifiers. There are statistical models that can be used to explore heterogeneity, and these are described in the Cochrane handbook. The heterogeneity statistic \( I^2 \) can be obtained and considered in the reporting as an indicator of heterogeneity. The GRADE handbook provides a rule of thumb interpretation [73]. All the statistical approaches have limitations and the key item here is to review the heterogeneity and try and understand the underlying factors and discuss this in the report.

Grade the evidence

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) working group https://www.gradewerkinggroup.org/ [4] has developed a handbook https://gdt.gradepro.org/app/handbook/handbook.html#h. f7lwc8w9c3nh8 [73] that provides the basis of grading recommendations made in clinical practice guidelines and systematic reviews. One of the purposes of a systematic review is to synthesise the knowledge from primary studies and this synthesised knowledge is fundamental to the recommendation that will guide clinical practice. Commenting on the factors that facilitate the GRADE system (Study design, Risk of bias, Indirectness, Inconsistency, Imprecision, and publication bias) should all be done as part of the systematic review. The GRADE handbook does point out that the methodology used for primary studies in the diagnostic testing realm is less robust, because there are less randomised controlled trials and more observational studies, but in principle the criteria are applicable. Table 4 and Figure 5 describes the GRADE evaluation and the up and down grading factors. Consistency of the results and the confidence limits around the estimate of effect are important factors.

Finalise the report and submit a manuscript for publication

Updating the contents of the systematic review prior to submission for publication is important. New publications may have appeared in the literature while the team have been working on the data extraction and analysis. Thus doing a updated search using the initial criteria is important to locate and add in any new publications.

The systematic reviews in laboratory medicine can be found in several places in the published literature and in

**Table 4:** Factors that impact the quality of evidence for diagnostic test accuracy (DTA) studies.

<table>
<thead>
<tr>
<th>Factors affecting the quality of evidence</th>
<th>Application for DTA studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>Unlike other intervention studies, diagnostic accuracy cross-sectional or cohort studies, directly comparing subjects test result with the available “gold” standard diagnostic strategy are considered high quality evidence (unless affected by other factors).</td>
</tr>
<tr>
<td>Risk of bias (limitations in study design and execution)</td>
<td>Evaluated differently than for intervention studies by examining:</td>
</tr>
<tr>
<td>Indirectness</td>
<td>This refers to lowering quality of evidence due to:</td>
</tr>
<tr>
<td>Important inconsistency in study results</td>
<td>Regarding sensitivity, specificity or likelihood ratios; especially if inconsistency is unexplained.</td>
</tr>
<tr>
<td>Imprecise evidence</td>
<td>Referring to wide confidence intervals for estimates of test accuracy.</td>
</tr>
<tr>
<td>High probability of publication bias</td>
<td>Referring to small studies effect or asymmetry in a funnel plot.</td>
</tr>
<tr>
<td>Upgrading for dose effect and large effects residual plausible bias</td>
<td>There is disagreement if and how dose effects play a role in assessing the quality of evidence in DTA studies, and methods have not been properly developed.</td>
</tr>
</tbody>
</table>

Adapted from the GRADE Handbook Table 7.2. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Updated October 2013. Holger Schünemann, Jan Brozek, Gordon Guyatt, and Andrew Oxman [71].
organization reports. Most laboratory medicine journals will consider systematic reviews for publication, as will many other journals. There is a good description of the requirements for publication in the Annals of Internal Medicine [6].

To successfully publish systematic reviews it is worth considering the guidance offered by the STARD 2015 initiative [57] in assessing the original papers the quality evaluation along the line of the QUADAS-2 tool [54, 55]. A commentary on reporting measures of accuracy is also helpful in preparing manuscripts [74]. The PRISMA checklist [49] or more specifically PRISMA-DTA [75] is an important guide to the structure of the report that should be checked prior to submission to a journal.

There are other evaluative tools available, and these are useful guides to consider in performing a systematic review and in the writing up of such a report. These include the Centre for Evidence Based Medicine (CEBM) in Oxford and the Cochrane Collaboration [30] that provide a structure for doing systematic review and the structure for reports that get published in the Cochrane database.

**Conclusions**

In conclusion, systematic reviews of laboratory tests are required for evidence based laboratory medicine. There are established techniques to follow to undertake this type of research and publication is possible. This paper has described the process and will assist investigators in planning these types of studies.

**Acknowledgments:** Dr. Owen Wiese for Figure 3. Rachel Don-Wauchope for Figures 1, 4, and 5.

**Research ethics:** Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Research funding: None declared.

References


50. PROSPERO international prospective register of systematic reviews. University of York: Centre for reviews and dissemination, https://www.crd.york.ac.uk/prospero/ [Accessed 24 Jul 2023].


