How good is the evidence base for test selection in clinical guidelines?

Shivani Misra a,⁎, Julian H. Barth b

a Clinical Biochemistry and Metabolic Medicine, Imperial Healthcare NHS Trust, London, UK
b Clinical Biochemistry, Leeds Teaching Hospitals NHS Trust, Leeds, UK

Abstract

Clinical guidelines are ubiquitous, manifold and form an integral component of evidence-based clinical practice. Guidelines on test selection are often considered a useful adjunct to aid clinical decision-making, as test selection is a complex process that is influenced by many patient, clinician and laboratory factors. However, it is important to carefully evaluate several aspects of these guidelines, which include the context of the test in the guideline, the quality of the studies underpinning recommendations, the extent of the evaluation of effectiveness (or performance) of the specific test and in the clinical pathway, its applicability and ease of implementation. A robust evaluation of a diagnostic test should incorporate several stages including evaluation in healthy, symptomatic but unaffected and affected populations, and importantly a measurement of impact on patient outcomes. Few diagnostic studies meet these criteria, and therefore crucial aspects of test evaluation are overlooked prior to incorporation into clinical guidelines. Whilst efforts are made to standardise reporting of studies, strength of evidence and quality of guidelines, further work is required to improve the quality of the diagnostic studies that formulate these guidelines. It is important that clinicians using guidelines for test selection appreciate the limitations of the diagnostic test, and the guidelines themselves.

1. Introduction

The term evidence-based is defined as “the conscientious, explicit and judicious use of current best evidence, in making decisions about the care of patients” [1]. In laboratory medicine, this concept can be applied at several levels of decision-making, including for test selection. Using evidence to guide test selection is an extremely important process that helps ensure laboratory tests are appropriately selected. Appropriate selection of tests in turn will be guided not only by the evidence base supporting use of the test, but also by the individual patient context, the clinical scenario, availability of the test, user preference and local economical considerations. Guidelines on test selection are an important facet of practicing evidence-based medicine. They should help to harmonise practice, reduce inappropriate test selection and reduce treatment variations secondary to inappropriate interpretation of test results due to analytical variations.

However, it is important to appreciate the type of evidence underpinning recommendations on test use in guidelines. In general terms diagnostic performance may be evaluated against a comparator that is deemed to be the reference standard and will thus evaluate diagnostic accuracy. Increasingly however, users of laboratory tests require some measure of the clinical utility, that is how well the test affects or changes health outcomes [2]. Therefore an evaluation of clinical utility in the form of randomised controlled trials is desirable, albeit not always achievable.

Evidence-based guidelines for test selection should ideally encompass many aspects of the process, but in practice may be limited in their scope. Whilst this does not negate the utility as they help standardise practice, users should understand the limitations of such an approach at an individual patient level.

One further aspect is also important to consider: the process by which guidelines are produced. Guidelines are produced by disparate groups. In the past this has led to inconsistency in quality and robustness of guidelines. Current initiatives such as AGREE, STARD and GRADE have provided a rigorous framework through which clinical guidelines can be constructed in a standardised way. Nevertheless the process is not infallible and the same body of evidence as interpreted by guideline developers may result in different recommendations.

This article reviews some of the salient issues and challenges to reflect on, when considering guidelines on test selection.

2. How are tests selected?

Test selection is a process undertaken by health-care professionals that can be influenced by laboratories, clinical guidelines and patients themselves. From a disease-perspective, tests may be selected for diagnosis, screening, monitoring, treatment selection, risk stratification and prognosis. But in practice, test selection is far more complex, with
non-disease related factors such as, patient choice, physician choice, test availability, cost and demand management strategies all impinging on the process.

Selecting a test is seldom an isolated process and laboratory tests are a component of a clinical pathway, relevant to a particular clinical scenario in a given patient [3]. The test results will usually be integrated with other clinical and laboratory findings as an investigative strategy related to the clinical scenario [4]. This is important to consider as studies evaluating tests often do so in the context of a clinical pathway, which may be affected by downstream events. These subsequent actions may affect the clinical course and thus the perceived usefulness of the test (see later).

The use of tests and their diagnostic utility alter at different time points during patient management. At the beginning of the pathway a diagnostic test requires sensitivity whereas confirmation of the test will require specificity, although it is expected that a single test will have both. Finally when a test is used for disease monitoring, analytical precision becomes the key property. However, it must be appreciated that the vast majority of laboratory tests are requested for routine assessment of unwell patients, during which the process of test selection becomes somewhat arbitrary and under these circumstances evidence is largely lacking.

### 3. Consequences of inappropriate test selection

The aim of appropriate test selection is to improve patient care, however, as discussed measuring clinical utility of testing can be challenging. Conceptually it is simpler to think about the consequences of inappropriate test selection, and to therefore consider guidelines on test selection as a means to preventing these.

Bossuyt and McCaffery published a framework to help define possible patient outcomes of a testing intervention [5]. These can be loosely categorised into clinical pathway effects, direct health effects and secondary non-clinical measures. It is therefore useful to think of these domains as those that can be affected by inappropriate test selection.

- **Clinical pathway effects**: defined as the clinical response to the test result, which may be to select, start, alter, stop or modify treatment. To order further tests, or to monitor with no further intervention.
- **Direct health effects**: defined as the physical effect to the patient in having the test. Venous blood tests are generally well-tolerated but other tests such as arterial puncture or cerebrospinal fluid sampling carry risks and complications to the patient.
- **Secondary non-clinical effects**: refers to the way in which the patient reacts to the result of the test, which can be further categorised as emotional, social, cognitive and behavioural.

Consequences at a patient level are therefore manifold and often not fully appreciated. Guidelines on appropriate test selection often focus on the clinical pathway, but it is worth considering these other potential outcomes when selecting tests.

Investigative tests are not perfect and carry the risks of false positive and negative diagnoses. Increasing numbers of investigations leads to increased interventions with their attendant risks [6]. This is particularly likely when a test is used outside its normal context. We have highlighted this in a patient with ascites due to hypothyroidism in whom multiple investigations including an exploratory laparotomy were undertaken due to the misinterpretation that an elevated CA125 was caused by ovarian cancer rather than by ascitic fluid [7].

Part of the problem lies in the definition of what constitutes an appropriate test, which has led to wide variations in estimates of inappropriate testing [8,9] Therefore guidelines on test selection are a positive step in helping to prevent any of these potential consequences through standardisation of practice.

### 4. Evidence base for laboratory tests

It is important to consider what evidence underpins guidelines on test selection. Many schemes for the evaluation of medical tests have been proposed. In the last few decades the introduction of new medical tests into practice, has been compared to the introduction of a new drug. The latter has a well-defined hierarchical model of phased evaluation and proposals to adopt a similar approach for the introduction of a new test have also been made [10–12]. This is keeping with a paradigm shift in the way undertaking medical tests are rightly regarded as equal to any other medical intervention.

#### 4.1. Ideal evaluation of a medical test

Lijmer et al. undertook a systematic review of 19 different published models for evaluation of medical tests [10]. The models reviewed had striking similarities and in general included:

- an early phase where tests were developed;
- a diagnostic accuracy phase where the test was evaluated in a variety of settings, including healthy volunteers, diseased population, population similar to the intended use and against a reference comparator;
- a clinical effectiveness phase, which included some measure of diagnostic thinking efficiency (the degree to which decision making was altered as a result of the test) and randomised controlled trials to assess effect on health outcomes;
- some models additionally assessed cost effectiveness and other outcomes e.g. secondary changes in practice.

In a departure from drug evaluations, the proposal for phased evaluations of medical tests describes a cyclical rather than a linear unidirectional course. Such a cyclical process has been described by the Centre for Disease Control for evaluation of genetic tests, further developed by the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) working group [11]. The approach asks 44 targeted questions which comprehensively review the new genetic test. These are broadly separated into 4 domains:

- **Analytical validity**: how accurately the test measures the genotype of interest
- **Clinical validity**: how consistently and accurately the test detects or predicts the intermediate or final outcomes of interest
- **Clinical utility**: how likely the test is to significantly improve patient outcomes
- **Ethical, legal and social implications**: identification of such issues and any safeguards in place.

The Global Harmonization Task Force (GHTF) for the scientific validity, determination and performance evaluation of in vitro diagnostic medical devices also published a cyclical phased approach with similar domains [12].

#### 4.2. Points to consider in existing studies evaluating medical tests

The above approaches highlight that a focused evaluation of medical tests on diagnostic accuracy studies alone is not sufficiently robust, but despite this, it may not be feasible or appropriate to undertake a thorough evaluation of all the domains. However, some evidence of clinical utility, either directly through randomised studies (see below) or indirectly through the use of decision analysis models, is imperative.

- Tests form part of a clinical/diagnostic pathway or testing strategy: In reality the evaluation of medical tests is even more complex as tests are seldom undertaken in isolation and the vast majority do not usually have a direct uninterrupted effect on patient outcome [5]. The decision to utilise a test is usually part of a clinical pathway and interpretation of the test result will be combined with downstream outcomes. The evaluation of a test will be pathway specific and care
should be taken before transferring conclusions to other scenarios [13]. For example a poorly performing test in a clinical pathway where downstream interventions are good may be deemed to have clinical utility, whereas an excellently performing test with poor downstream interventions may be erroneously thought to have limited utility [5]. Therefore simply focusing on diagnostic accuracy or test efficiency does not necessarily improve patient outcome.

• Comparing multiple testing strategies is of more value: In practice, with diagnosis in mind, a clinical user is more likely to be contemplating several testing strategies and trying to decide between them. Therefore studies comparing several clinical pathways are of more value than purely assessing diagnostic accuracy of the individual test or clinical utility of one pathway [13].

• Randomised control trials: provide the ideal platform on which to evaluate the clinical utility of a testing strategy against an existing reference, but the correct question needs to be addressed. As described by Lord et al., if a new test strategy results in additional patients being identified with disease, then the randomised clinical utility trial should focus on treatment versus no treatment on this newly identified subgroup which would have been missed by the prior strategy, rather than a comparison of people randomised to receiving the test versus no test [14,15]. Defining the question in this way, it is apparent that many test evaluation studies are lacking.

5. Source of bias to consider in diagnostic accuracy studies

By now it is apparent that few test evaluations will meet the ideal criteria. If we take the focus back to the majority of studies in which test accuracy alone has been tested, then physicians, reviewers and policy-makers face a difficult choice. Do they exclude substandard studies and risk a smaller pool of data, or include them and risk bias? A study by Lijmer et al. attempted to empirically evaluate the effect of shortcomings in study design (design bias) on estimates of diagnostic accuracy. They discussed many potential sources of bias (see Table 1) and quantified the extent to which they affected estimations of diagnostic accuracy by studying 218 diagnostic evaluations [16]. The study reported considerable increase in over-estimation of diagnostic accuracy due to spectrum bias and differential reference standard bias.

Brain natriuretic peptide is a good example of an evidence-based evaluation of a diagnostic test since it has been tested sequentially in large-scale studies targeting variable populations and in various clinical contexts [17]. The BASEL study evaluating a BNP testing strategy in the Emergency Department is one of the rare examples where a randomised control trial has been used [18].

A meta-analysis studying the use of faecal calprotectin in screening for inflammatory bowel disease, illustrates the importance of spectrum bias as 20 of 33 studies meeting inclusion criteria were excluded because the study design compared individuals with known disease to healthy controls, rather than between individuals with appropriate symptoms [19].

6. Guidelines on diagnostic tests

In the context of a diagnostic test, a guideline may address the use of a test in either a single or a variety of clinical settings, or it may be included within a clinical pathway with other interventions. Similarly the studies that form the evidence contributing to guidelines on diagnostic tests may assess the test in a variety of ways.

Most current guidelines are designed to aid patient management and are therefore based around a clinical problem in which a test can dichotomise between decisions. Examples might be troponin in chest pain to direct patients to angiography [20] or CA125 in women over 50 years with abdominal bloating to be directed to imaging [21]. The studies that underpin these guidelines usually evaluate the diagnostic tests performance as part of a pathway, but it is unusual for test to have been subjected to formal randomised studies to assess clinical performance. Moreover, they often lack the necessary detail required for the practical introduction of a diagnostic test. For example in the NICE guidance on chest pain, though troponin elevation is discussed, there is no mention of non-ischaemic causes of a raised troponin, which may be of particular relevance when considering the patient groups in whom troponin is commonly requested, nor is there any discussion regarding differences between assays [20].

Therefore when considering the evidence-base for test-selection and the guidelines themselves, it is important to ask the following questions:

• Quality of evidence: Have the diagnostic studies underpinning the guidance on a particular test been designed, conducted and reported correctly?
• Aspect of test evaluation: What aspect of the test has been evaluated in any individual study, for example has the diagnostic accuracy been assessed? Has the clinical utility been assessed?
• Test context in the guideline: Does the guidance relate to the use of the test in generic circumstances, or is it limited to a single clinical scenario?
• Test description in guideline: Does the guideline contain adequate information about the diagnostic test? Has it made readers aware of test limitations?

7. What are the key issues regarding evidence for guidelines?

7.1. Current position of diagnostic tests in guidelines

A common finding in guidelines is the woefully inadequate critique of evidence provided for laboratory tests. This will have less impact if the guidelines only state the need for a raised or lowered value or a qualitative result; however, where specific values are used, between-method bias can lead to clinical errors and a full understanding of the limitations of each test is imperative. For example, a study comparing the accuracy of LDL-cholesterol methods of various manufacturers in classifying patients into cardiovascular risk categories showed poor concordance with the LDL Reference Method. This may lead to risk misclassification and inadequate choice of therapeutic intervention which is proportional to the method bias [22].

Real-life clinical scenarios are seldom as straightforward as the diagnostic pathways outlined particularly when the potential effects of co-morbidities and medications on laboratory tests are included. The 2003 British Thoracic Societies Guideline on the investigation of unilateral pleural effusions is a typical example of the brevity with which diagnostic tests are often described [23]. In a robust critique of these guidelines, the lack of information provided on pre-analytical requirements, assay limitations and perhaps most importantly the lack of evidence-base for many recommendations relating to diagnostic tests on pleural fluid, was highlighted [24]. Furthermore, the interpretation of the biochemical tests mentioned made no attempt to highlight the equivocal nature of certain results.

The European Federation of Clinical Chemistry and Laboratory Medicine Guidelines working group has recently developed a checklist of recommendations for laboratory diagnostics in clinical guidelines, to try and improve the quality of guidelines. This checklist was used to audit 12 published clinical guidelines and a mean of only 30% compliance was found. The authors suggested that future guidelines should include laboratory medicine specialists to help formulate the recommendations on laboratory tests [25].

7.2. Attempts to standardise diagnostic accuracy studies and guidelines

Several initiatives have been launched to standardise or provide standards on reporting and conducting high quality diagnostic accuracy studies, including quality assessment of diagnostic accuracy studies (QUADAS) and standards for the reporting of diagnostic accuracy...
studies (STARD). QUADAS is a framework through which to assess the quality of previously published studies, for the purposes of a systematic review, by answering ‘yes’, ‘no’ or ‘unclear’ to 14 questions pertaining to the study methodology [26,27]. STARD attempts to improve and standardise the reporting of diagnostic accuracy studies and is primarily aimed at authors and journal editors to help guide quality reporting. It addresses the basic information about the test being investigated that should be included in scientific reports and provides a 25 item checklist to facilitate this [28].

STARD guidelines provide openness rather than assessment of methodological quality. Although STARD has been widely adopted by over 200 peer-reviewed publications, a small 2008 study assessing the reporting of BNP diagnostic accuracy studies, found that only 2 studies met a combined checklist of 22 items from QUADAS and STARD [29]. Other discipline-specific studies have also shown poor adherence to STARD criteria or only limited improvements in reporting following publication of STARD [30–32].

Similarly, oncologists have attempted to standardise reporting of tumour marker prognostic studies through REMARK (REporting recommendations for tumour MARKer prognostic studies) [33].

AGREE (Appraisal of Guidelines for REsearch and Evaluation) is another instrument aimed at producing and constructing standardised guidelines by assessment of 6 quality domains [34]: scope and purpose, stakeholder involvement, rigour of development, clarity and presentation, applicability, and editorial independence. However, it does not assess the validity of studies contributing to guidance [35,36]. A recently published study appraised all 11 published laboratory clinical practice guidelines from the National Academy of Clinical Biochemistry against the AGREE criteria for quality guidelines [37] and found a generally poor performance (median score 42%), except for the very recent diabetes guidelines [38].

Grading of recommendations assessment, development and evaluation (GRADE) is one of the few attempting to assess quality and grading medical evidence contributing to clinical guidelines [39]. It is unique as the quality of evidence is translated into a strength of recommendation, giving clinicians the ability to weight the guidance provided. GRADE has also been advocated to assess quality of evidence in guidelines on diagnostic tests, uniquely focusing on patient outcomes rather than just diagnostic accuracy [40]. GRADE advocates taking into consideration key points such as study design, use of consecutive patients, comparison with a gold standard test and use of a population with suspected disease and not healthy volunteers. GRADE is therefore a key tool to use when producing clinical guidelines, however in relation to diagnostic accuracy studies invariably evidence is graded as weak because of the indirect link to patient outcomes.

The emergence of these many initiatives clearly highlights the heterogeneity with which studies – be they diagnostic or otherwise – are reported. Standardised reporting is, however only one half of the challenge, as the design and quality of the studies themselves also need to be addressed. However, despite the many grading systems available, there does not seem to have been an improvement in the overall quality of the reporting of clinical recommendations [41–44]. Furthermore, the same body of evidence can give rise to quite divergent interpretations [45,46].

## 8. Other challenges in guidelines on test selection

### 8.1. Responsibilities of the laboratory

What is the role of the laboratory in this process? Primarily the laboratory must ensure that prescribed tests are available, appropriately standardised and that their quality is maintained to permit clinical decisions to be made at defined values. However, laboratories also have a responsibility to ensure issues related to testing methods are communicated and/or highlighted to the clinicians. Should the laboratory also provide appraisal of existing guidelines on diagnostic testing? This might be particularly important where the guidance lacks appropriate minimal information on the test. Certainly there is a role for laboratory involvement in the production and implementation of guidelines on diagnostic tests at a local level.

### 8.2. Ensuring guideline adherence

It is also not clear who is responsible for ensuring that guidelines are followed. Clinical audit activities do contribute to the assessment of practice against an established standard (that could be a guideline), but there is no impetus to audit adherence to all guidelines on diagnostic tests. In the UK, primary care physicians are rewarded for regular testing diabetic patients’ glycated haemoglobin, however, in secondary care, there is no such obligation/reward for following guidelines, except in striving to provide a high standard of care. We have found, for example, that the frequency of requests for tumour markers exhibits a 7 day cycle corresponding to clinic appointments rather than at the 3 monthly intervals as stipulated in the guidance [47] (see Fig. 1).

In the UK, national agencies such as NICE, produce robust evidence-based guidelines, which clinicians are encouraged to follow but which are not mandatory. Implementation of these guidelines might be monitored nationally by the Care Quality Commission, through local self-reporting of adherence to NICE procedures, but this is not currently strictly enforced. The French Parliament in line with this thinking has recently decided that laboratories have a legal obligation to follow recommendations and guidelines produced by the Haute Autorité de Santé [48].

Whilst Clinical Pathology Accreditation (CPA) and other accreditation bodies stipulate the requirement of laboratory practices to be audited regularly, it does not currently require clinical audits, or audits of guideline adherence to be regularly undertaken. One of the hurdles preventing this, is of course, consensus on which guidelines to follow.

The issue of monitoring adherence to guidelines is actually only one facet of a larger laboratory concern, which is assessing total quality. This should not only focus on the analytical or operational processes governed by accreditation standards, but should follow the entire process from the pre-pre-analytical phase to the post-post-analytical phase or the ‘end-to-end’ clinical service. It has been suggested that laboratory quality indicators should include 1) removal of out-of-date tests and tests with no clinical utility and 2) that tests offered should match (inter)national guidance and these could be monitored by professional bodies or accreditation visits [49].

### Table 1

<table>
<thead>
<tr>
<th>Potential sources of bias in diagnostic accuracy studies.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias type</strong></td>
</tr>
<tr>
<td>Spectrum bias</td>
</tr>
<tr>
<td>Selection bias</td>
</tr>
<tr>
<td>Verification bias</td>
</tr>
<tr>
<td>Blinding</td>
</tr>
<tr>
<td>Publication bias</td>
</tr>
</tbody>
</table>
9. Summary

There is no shortage of available guidance on diagnostic test selection, but clinicians must pay attention to the context in which the test has been assessed in the studies contributing to such guidance. There is a need and an opportunity to adopt high quality assessments of diagnostic tests through robust, cyclical and where necessary randomised studies, which can then contribute a high quality evidence basis for guidelines. The many initiatives attempting to standardise and raise the quality of diagnostic guidelines, are a key step forward but poor study design remains a hurdle to overcome.

Ultimately a guideline is only an advisory set of procedures to follow based on the best available evidence. In diagnostic laboratory medicine perhaps more than other fields, test selection therefore requires the clinician to understand the test itself, the patient and the clinical scenario. In this context, the laboratory has a responsibility to respond to and reflect on available guidance, in an effort to ensure that pathology-users are able to undertake evidence-based test selection.

Acknowledgements

The authors are both members of the European Federation of Clinical Chemistry and Laboratory Medicine and European Union of Medical Specialists Joint Working Group on Guidelines and would like to thank other members of the working group, Wytze Oosterhuis, Kristen Moberg Aakre, Michel Langlois, Joseph Watine and Sverre Sandberg for their helpful comments on drafts of this article.

References
