Interpreting the evidence-base for making recommendations on medical tests

Gopalakrishna, G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter Eight

Targeted Test Evaluation: Designing Diagnostic Accuracy Studies with Clear Hypotheses

Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PMM
Introduction

We take it for granted that new pharmaceuticals can only be released to the market if they have been shown to be sufficiently effective and safe. The benefits and harms of such pharmaceuticals are typically evaluated in comparative studies known as randomized controlled trials, where outcomes are compared between one group that receives the drug in question and a second group that receives a placebo or an alternative drug for the same purpose [1]. These trials have a pre-specified hypothesis: the drug in question is assumed to be effective. This hypothesis is statistically tested against the null hypothesis of no effect. The sample size of the trial is based on the desired power: the probability of rejecting the null hypothesis if the drug has a minimal clinically relevant effect [2]. Investigators report effect sizes (e.g., mean difference or odds ratio) and calculate p-values: the conditional probability that the effect size has or exceeds the calculated value, if the null hypothesis of no effect is true.

The randomized controlled trial has become the undisputed cornerstone of evidence-based healthcare [3]. Besides new pharmaceuticals, many other new interventions are now evaluated in the same way. In contrast, the situation is very different for medical tests. Comparative trials that focus on the effect of a test on patient outcomes such as those related to the test results or testing process are relatively rare [4]. Such outcomes may be (1) direct health effects (e.g., improved treatment selection or side effects), (2) emotional, social, cognitive or behavioral responses as a result of testing, (3) legal or ethical effects of testing, (4) costs as a result of testing [5]. Additionally, there is no requirement to demonstrate a reasonable harms/benefits balance for new tests before they can be introduced into clinical practice [6].

In clinical epidemiology, we are familiar with a different type of study for evaluating the clinical performance of medical tests. Known as diagnostic accuracy studies, such studies evaluate a test’s ability to correctly distinguish between patients with and without a target condition by comparing the results of the test against the results of applying the reference standard in the same patients [7].

Somewhat surprisingly, most diagnostic accuracy studies do not have a pre-specified hypothesis, and are reported without an explicit statistical test of the null hypothesis. In an analysis of 126 diagnostic accuracy studies, Ochodo and colleagues observed that only 12% reported any statistical test of a hypothesis somewhat related to the study
objectives, and no more than 11% reported a sample size calculation [8]. Others found similar results in varying clinical fields of diagnostic testing [9-11].

The absence of a pre-specified study hypothesis, and the failure to evaluate it as a statistical test has several unfortunate consequences. It often leads to underpowered studies with wide confidence intervals around accuracy estimates, in some cases barely informing the reader about the performance of the test [11]. It also opens the door to over-optimistic conclusions that may not be supported by the study findings and premature claims about the clinical value of diagnostic tests. In the same evaluation, Ochodo and colleagues noticed that 31% of the study reports contained one or more signs of generous interpretation of the study findings, such as overly optimistic abstracts, discrepancies between study aims and conclusions, or conclusions based on selected subgroups [8].

We see no reason why the logic of having clear study hypotheses, defined apriori, would only apply to randomized controlled trials of pharmaceuticals, and not extend to diagnostic accuracy studies. Study quality is likely to benefit from this, as explicitly defining a study hypothesis forces researchers to express what accuracy values would make a test acceptable before initiating a study, and invites them to perform proper statistical tests of study hypotheses. Readers would also benefit since having hypotheses limits the room for ‘spin’ in the reporting and interpretation and facilitates the drawing of meaningful conclusions. A clearly defined study hypothesis also enables an informed judgment of the appropriateness of the study’s design, analysis and conclusions [12]. We believe that defining clear study hypotheses and testing them in diagnostic accuracy studies is doable. In this article, we give the reader an introduction to how such hypotheses can be meaningfully derived. The accompanying statistical hypothesis tests have all been defined before in the literature; the same applies for the methods for sample size calculations [13]. We focus our exposition on single test evaluations.
A diagnostic accuracy study

In diagnostic accuracy studies, a series of patients suspected of having a target disease undergo both the index test, the test that is being evaluated, and the clinical reference standard, which is the best available method for establishing if a patient does or does not have the target condition [7].

Assuming that the results of the index test are dichotomous – either positive or negative – and that the target condition is either present or absent, we can present the results of the comparison of the index test against the reference standard in a 2x2 table, which shows the extent to which the results of both tests are in agreement (Figure 1).

Discrepancies between the index test results and the classification based on the reference standard are considered to be false positive index test results and false negative index test results.

Although it is possible to generate a single estimate of the index test’s accuracy, such as the diagnostic odds ratio [14], it is more meaningful to report two statistics: one for patients with the target condition, and one for patients without the target condition. One reason is that downstream actions will usually differ for positive test results compared to negative test results, while the consequences of misclassifications related to the harm, cost or burden of false positive test results differ from those related to false negative test results.
The two measures used most frequently in diagnostic accuracy studies are the test’s sensitivity and specificity, which we will focus on. With the number of true positive and false negative results, an estimate of the index test’s sensitivity can be calculated, based on the proportion of patients with the target condition who have a positive index test result (Figure 1). Similarly, the number of true negatives and false positives generate an estimate of the test’s specificity: the proportion of patients without the target condition who have a negative index test result.

As a visual aid, we can picture a test’s sensitivity and specificity in the receiver operating curve (ROC) space, which has these two dimensions: sensitivity (y-axis) and specificity (x-axis) (Figure 2).

Legend: ROC space has two dimensions: sensitivity (y-axis) and specificity (x-axis). A worthless test generates, proportionally, as many true positives as false positives. In that case sensitivity and one minus specificity add up to one. Informative tests have a higher proportion of true positives than false positives: the test’s accuracy will then be in the upper triangle.

Figure 2. Receiver operating curve (ROC) space.
Hypotheses about a test’s diagnostic accuracy

Since we have two measures of accuracy, we need a joint hypothesis; one about sensitivity and a second about specificity. Having a hypothesis about sensitivity or specificity only would be pointless, since one can always arbitrarily set the positivity rate for a test to exceed a pre-specified value; should one, for example, categorize all patients as positive with the index test, the sensitivity would automatically be 100%, at a specificity of 0%.

A worthless test generates, proportionally, as many true positives as false positives. In that case sensitivity and one minus specificity add up to one. Informative tests have a higher proportion of true positives than false positives, and sensitivity and specificity will typically be higher than 0.50 for each of these measures. In ROC space, the test’s accuracy will then be in the upper triangle (Figure 2).

A very simple hypothesis would be then both sensitivity and specificity exceed 0.50:

\[ H1: \text{Sensitivity} > 0.50 \text{ and } \text{Specificity} > 0.50. \]

This could be evaluated against the following null hypothesis:

\[ H0: \text{Sensitivity} \leq 0.50 \text{ and/or } \text{Specificity} \leq 0.50. \]

However, when a diagnostic accuracy study is designed, it is typically already known that the index test is somewhat informative, but this is not necessarily enough for the test to be clinically useful. For example, if a test is meant to rule out disease with a high level of certainty, then sensitivity should be very high (i.e., low proportion of false negatives), and a value of just above 0.50 will not be enough, as the test will still misclassify almost half of those with the target condition. Similarly, if one wants a new, expensive test to replace an existing, inexpensive test, then the accuracy of that new test has to at least exceed that of the existing test. So the required levels of sensitivity and specificity will depend on the context in which the new test is going to be used in clinical practice.

What would therefore be useful to know is whether the test meets minimally acceptable criteria (MAC) for a predefined test performance:

\[ H1: \text{Sensitivity} > \text{MAC(Sensitivity)} \text{ and } \text{Specificity} > \text{MAC(Specificity)}. \]

This could be evaluated against the following null hypothesis:

\[ H0: \text{Sensitivity} \leq \text{MAC(Sensitivity)} \text{ and/or } \text{Specificity} \leq \text{MAC(Specificity)}. \]
In ROC space, this can be defined as the rectangle in the upper triangle that corresponds to the MAC (Figure 3). The test will be considered acceptable if both the sensitivity and specificity are in the rectangle, which we will refer to as the ‘target region’ in ROC space.

The target region will always touch the upper left corner of ROC space, which is the point for perfect tests, where both sensitivity and specificity are 1.0. From there, the rectangle extends down, to the MAC for sensitivity, and extend to the right, towards the MAC for specificity.

A diagnostic accuracy study will produce point estimates of sensitivity and specificity, along with 95% confidence intervals around it. If we position these in ROC space, then both the point estimates and the 95% confidence intervals should be completely positioned in the target region. So if the MAC for sensitivity is set at 0.85 and the MAC for specificity at 0.90, the lower limit of the 95% confidence interval should be ≥0.85 for sensitivity and ≥0.90 for specificity. Sample size calculations can be performed to make sure that a sufficient number of patients are included to test this hypothesis with sufficient power. The test statistics to evaluate such a null hypothesis have been previously defined [13].

Legend: The target region of a diagnostic accuracy study will always touch the upper left corner of ROC space, which is the point for perfect tests, where both sensitivity and specificity are 1.0. From there, the rectangle extends down, to the MAC for sensitivity, and extend to the right, towards the MAC for specificity. The red square represents the target region of a diagnostic accuracy study with a MAC(sensitivity) of 0.60, and a MAC(specificity) of 0.60. MAC = minimally acceptable criteria.

Figure 3. Receiver operating curve (ROC) space.
Defining minimally acceptable criteria (MAC) for diagnostic accuracy

We suggest the following six steps in defining the MAC of sensitivity and specificity for any given test being evaluated in a diagnostic accuracy study (Figure 4).

**Step 1: Identify the existing test-treatment pathway in which the index test will be used**

The accuracy of tests typically varies depending on the clinical setting in which they are applied, and on how the tests are used. So the sensitivity and specificity of a single test is likely to differ across different settings and applications. In addition, consequences of testing may vary if applied in different settings. Tests, therefore, should be evaluated based on the clinical context in which they might be used. This can be done by first defining the existing pathway for a specific target condition in mind.

We propose a number of criteria that authors can use to help guide them in defining the test-treatment pathway:

*Target condition:* This is defined as a particular disease or disease stage that the proposed index test is intended to identify.
Patients: These can be those fulfilling certain characteristics such as gender, age, those presenting with certain signs and symptoms, or those having undergone other tests.

Setting: This is where patients will be recruited, and the index test will be applied. This can, for example, be in primary or secondary care; at an emergency department, an outpatient clinic, community and so on.

Tests in the test-treatment pathway: These are all the tests that a specific type of patients that present in a specific clinical setting, as defined above, undergo before the diagnostic work-up is closed and a decision on further management is taken. Patient history and physical examination are commonly the first tests to be applied when a patient seeks medical care. Based on this, the physician forms an idea of the probability that the patient has the target condition. Because a certain level of uncertainty generally remains in this phase of the clinical pathway, the physician may then want to perform additional tests which should be defined in the pathway.

Step 1: Example
In patients with non-small cell lung cancer (NSCLC), mediastinal lymph node staging is performed to select them for optimal treatment. Annema and colleagues aimed to evaluate the optimal strategy for detecting mediastinal lymph node metastases in patients with potentially resectable NSCLC, presented in university hospitals, who need to undergo such staging according to criteria defined in existing clinical guidelines [15]. So here the authors have defined the following: target condition as mediastinal lymph node metastases, type of patients as those with potentially resectable NSCLC, and the setting as patients presenting at university hospitals.

The authors sketched the existing test-treatment pathway as follows: “To detect mediastinal metastases, patients are routinely investigated with CT and fluorodeoxyglucose PET, followed by mediastinal tissue staging for enlarged or PET-positive intrathoracic nodes, as imaging alone is inaccurate. Mediastinal tissue staging is classically performed by mediastinoscopy, a surgical diagnostic procedure with a sensitivity of approximately 78%.”

So the first test in this pathway is imaging. However, imaging has low specificity, hence it produces many false positives. Guidelines therefore recommend to perform additional testing in patients with a positive imaging test, or in those that are at high risk for me-
diastinal lymph node metastases, according to certain other criteria. The authors of this study selected such patients for inclusion. The next test in the existing test-treatment pathway is mediastinoscopy, the results of which generally determine how patients are subsequently treated.

Step 2: Define the specifications of the index test in the test-treatment pathway
This step involves defining where in the existing test-treatment pathway the index test under evaluation will be positioned. There are a number of possible roles for diagnostic tests relative to an existing test such as triage, add-on and/or replacement tests [16].

A triage test is used before the existing test(s), and its results determine which patients will undergo the existing test. An add-on test is used after an existing test to improve the diagnostic accuracy of the testing strategy. A replacement test aims to replace an existing test, either because it is expected to have higher diagnostic accuracy, is less invasive, less costly or easier to use than the existing test.

It is also possible that the test is used in a completely new strategy, without prior tests. The latter would be the case with a new population screening strategy, where the alternative is no screening. With screening, only screen test positives will be invited for additional procedures, such as confirmatory testing and, possibly, treatment of screen-detected disease.

Defining the role of the index test in the existing test-treatment pathway helps ensure the appropriate study population is selected. For instance, if the index test is considered as an add-on test, then only patients that underwent the existing test would be selected for inclusion.

Step 2: Example
In the example for step 1, the existing pathway for detecting mediastinal lymph node metastases in patients with NSCLC was defined. As the final test in the existing pathway, mediastinoscopy, has imperfect specificity, the authors of the study aimed to add a new test in the existing pathway. Patients with a positive imaging test or other criteria, defined in clinical guidelines, would first undergo a combined approach of transesophageal and transbronchial ultrasound-guided needle aspiration (i.e. the index test which is an add-on test) [15]. This test combination has a specificity close to 100%, so if positive, mediastinal lymph node metastases were considered to be present, and patients were
treated as such. The sensitivity of the combination is less perfect, so in case of a negative result, patients would still undergo mediastinoscopy, which is more invasive than the transesophageal and transbronchial ultrasound-guided needle.

**Step 3: Define the downstream consequences of the index test versus the existing test**

Bearing in mind the positioning of the index test in the test-treatment pathway, the downstream consequences of the index test/ testing strategy versus the existing test/ testing strategy need to be defined. To do so, the consequences of the index test/strategy compared to the existing test/strategy need to be considered. This can be done by considering the impact of the index test versus the existing test on two aspects. First we have to define what effect the introduction of the new test may have on the overall clinical management decisions, in terms of what will change compared to the existing testing strategy. Clinical management decisions refer to actions leading to further testing, wait and see, follow-up intervals, or treatment decisions. Secondly, we have to define the difference in downstream patient important outcomes between the index test/strategy and the existing test/strategy.

**Step 3: Example**

The study by Rutten and colleagues [17] examined the use of laparascopy as an add-on test to the standard diagnostic work up of women suspected of advanced ovarian cancer. In that study, the change in clinical management decision would involve women with positive results at the standard diagnostic work up, who undergo laparoscopy (the index test). When positive, these women will directly receive chemotherapy first, instead of undergoing the surgically invasive and costly laparotomy. This means that women with extensive tumors, identified through laparoscopy, will not undergo an invasive and costly laparotomy, but will first receive neoadjuvant chemotherapy, consequently improving the overall rate of successfully operated women in the clinical pathway.

**Step 4: Weigh the consequences of true and false positive and negative index test results**

Defining MAC of sensitivity and specificity comes down to weighing the consequences of being misclassified as false positive or false negative, considering the clinical management decisions and downstream consequences, as previously outlined in Step 3. The following questions can help guide authors when weighing the consequences of test results: How serious is the target disease? Are there effective treatment options? Do these treatment options have side effects? How expensive and burdensome are they?
How serious would a delay in diagnosing the target disease be? Does the test result lead to additional (more accurate) testing to safeguard false positive and false negative results? Do these additional tests have side effects? How expensive are they?

Additional, the role of the index test in the pathway (as defined in Step 2) can influence the weight given to the consequences of being correctly classified or being misclassified. For instance, triage tests tend to have large numbers of false positives, which means that a positive triage test result typically does not confirm a disease. This may be acceptable as long as the number of diseased patients that are classified by the test as negative is minimal; there must be few or no false negatives. Only those patients with a positive test would be sent for further confirmatory testing. Further down the test-treatment pathway however, it is crucial to decrease the number of false positives to a minimum, since positive test results may lead to radical treatment decisions with potentially serious side effects. Therefore, add-on tests generally require lower proportions of false positives than triage tests. For replacement tests, the accuracy measures should, commonly, be at least as good as that of the existing test.

An approach on how to weight these consequences is described by Pepe and colleagues in their recent study [18]. They propose a method for defining what they refer to as an action threshold for medical procedures among test positives, as compared to controls or test negatives. It provides a means of making a “transparent, educated guess” by weighing the costs and benefits of being detected with a condition versus not.

**Step 4: Example**

In the review by Schmidt et al. [19] on the use of PET-CT for the detection of resectable non-small cell lung cancer (NSCLC), true positives are patients with no spread to the mediastinum who may therefore be candidates for resection, as well as patients with either distant or mediastinal metastases, or both, who may need to be biopsied before a treatment plan can be developed. True negatives are those patients who are found to have unresectable NSCLC and who will then have their cancer stage pathologically confirmed by a number of other tests that are considered suitable for the location of the affected lymph node(s). False negative patients would be those with nodes that are not significantly enlarged on (a prior) CT and may therefore miss treatment. False positives are of a lesser concern, since they should always be followed by a further confirmatory test in the pathway.
Step 5: Define the expected prevalence

The expected prevalence of the target condition in the population of interest should also be taken into account in this process, because absolute numbers of false positive and negative results will vary with disease prevalence. If 100 patients are tested by a test with a sensitivity and specificity of 90%, a disease prevalence of 50% will lead to 5 false positives and 5 false negatives. When disease prevalence is only 10%, there will only be 1 false positive but 9 false negatives. So the relative proportion of false positives and false negatives, as well as the relative proportion of negative downstream consequences of such tests results, depends on the prevalence of the target condition.

Multiple strategies can be used for defining the expected prevalence of the target condition in the specific clinical setting:

Previous literature: Previously performed diagnostic accuracy studies on the index test for detecting the target condition can be used. Disease prevalence may highly vary across studies, for example due to heterogeneity in the included population, or due to chance. Ideally, a systematic review of such studies is identified or performed, so that variation in prevalence across different settings can be mapped, and form the basis for defining the expected prevalence.

Pilot study: If no (reliable) previous literature is available on the prevalence of the target condition in the setting under investigation, a pilot study can be performed to estimate this. Such a study can be done retrospectively, on existing patient data. It may be more reliable, however, to do it prospectively, applying the exact same inclusion criteria and participant selection mechanisms as those that will be used in the diagnostic accuracy study.

Clinical experts: In the absence of previous literature and resources to perform a pilot study, clinical experts could be consulted to guess the expected prevalence of the target condition in the setting under investigation.

Step 5: Examples

In the systematic review by Korevaar and colleagues, the diagnostic accuracy of several minimally-invasive markers for detecting airway eosinophilia in patients with asthma was investigated [20]. The authors reported that “the prevalence of eosinophilia ranged from 20% to 88%” across included studies. A table in the appendix specifies the prevalence for each individual study.
In the study by Annema et al, referred to in Step 1 and 2, the authors reported in their description of the sample size calculation that they assumed “a prevalence of mediastinal nodal metastases of 70% [...] However, during a pre-specified interim monitoring of the study, the prevalence of mediastinal nodal metastases was found to be 55%, and therefore the sample size was increased [...]” [15].

**Step 6: Define minimally acceptable criteria (MAC) for sensitivity and specificity**

Based on the weighted consequences of true and false positives, and true and false negative test results, minimally acceptable criteria for sensitivity and specificity can be defined. The target zone in the ROC space can be drawn (Figure 3).

**Step 6: Example**

Despite substantial technological advances, diagnosing acute coronary syndromes (ACS) in patients admitted to emergency departments remains a major challenge. Missed cases of ACS rank high in medical malpractice lists. Most testing protocols currently rely on troponin I or T, used alone or in combination with creatine kinase MB or myoglobin.

Mitchell and colleagues wanted to evaluate the performance of a multimarker panel in ruling out ACS in the emergency department [21]. The markers were monocyte chemoattractant protein-1 (MCP), myeloperoxidase (MPO), C-reactive protein (CRP), and brain natriuretic peptide (BNP).

They first focused on the MAC for sensitivity. The sensitivity should not be lower than that of other protocols for the workup of patients with suspected ACS at chest pain units. They therefore defined the MAC for sensitivity at 96%. Then they focused on the specificity, which they required to be at least 10%, while anticipating that it would be at least 15%. With that specificity, 10% of patients without ACS would not need to go undergo conventional testing for suspected ACS. So the MAC for specificity is a very modest 10%.

This could lead to the following hypotheses:

H1: Sensitivity $> 0.95$ and Specificity $> 0.10$.

This could be evaluated against the following null hypothesis:

H0: Sensitivity $\leq 0.95$ and/or Specificity $\leq 0.10$. 
The authors then collected usable data from 414 patients admitted to their chest pain unit, of which only 7 were confirmed to have ACS. The sensitivity of the combination of markers was estimated at 86% (95% CI: 42% to 100%) with a specificity of 17% (95% CI: 13% to 21%). As the point estimate and confidence intervals for sensitivity are not in the target region, the conclusion must be that this multimarker panel does not have the desired test performance (figure 5).

Who defines minimally acceptable performance criteria?

The strategy that we have outlined to arrive at the MAC for diagnostic accuracy studies requires the expertise of multiple professionals. There should be clinical expertise to identify the management actions that will result from positive or negative test results. Clinical expertise is also needed for weighing the downstream consequences of test results, and in some cases it may be desirable or inevitable to include patients or their advocates in this process. There should be methodological expertise, to avoid mistakes and
asymmetries in drawing the clinical pathway, and to promote consistency in the process. The proposed strategy can be qualitative, semi-quantitative or fully quantitative. A qualitative approach could involve a Delphi based approach involving expert opinion only, without performing any form of evidence synthesis.

A semi-quantitative approach would include a synthesis of the evidence for the effects of downstream management actions that are guided by test results, and the consequences of misclassifications. For example, treatment for a target condition that is not present will expose patients to the harms of that treatment, without any benefit. Researchers could do a systematic review of these harms and risks.

A fully quantitative approach would require the building of a decision model, where all the steps are explicitly linked, resulting in a final estimate of the selected downstream outcomes that may be affected by the test, for example in terms of life expectancy, or quality-adjusted life expectancy. From the desired group-based difference in patient-relevant outcomes between the two strategies – the one with and the one without the index test – one could then use the model to calculate the combinations of sensitivity and specificity that would lead to that difference.

Discussion

We have proposed the concept of targeted test evaluation for diagnostic accuracy studies, in which study hypotheses are based on MAC of sensitivity and specificity, which can be represented as a target region in ROC space. Defining a study’s hypotheses and performing corresponding sample size calculations at the planning stages of a diagnostic accuracy study may provide reassurance that the study will have sufficient precision to be informative; it facilitates drawing meaningful conclusions, and helps to prevent overoptimistic interpretation of study findings leading to ‘spin’.

There is a growing recognition that explicitly specifying study hypotheses in the protocol-development phase of the study is crucial in test accuracy research. The updated STARD (Standards for Reporting of Diagnostic Accuracy Studies) statement for reporting diagnostic accuracy studies, published in 2015, now requires authors to report “study hypotheses” (item 4) and the “intended sample size and how it was determined” (item 18) [22]. Similar methods for focusing on the MAC of test performance are also increasingly
being implemented among systematic reviews and clinical guidelines. The Cochrane Handbook for Diagnostic Test Accuracy Reviews, for example, now encourages authors to describe the clinical pathway in which the test under evaluation will be implemented, including prior tests, the role of the index test and alternative tests, if applicable [23]. A similar practice is advised by the recently established GRADE (Grading of Recommendations Assessment, Development and Evaluation) quality assessment criteria for diagnostic accuracy studies, which encourages guideline developers to focus on and weigh consequences of testing [24].

The process described here is not that different from hypothesis formulation and sample size calculations in randomized controlled trials. Even though randomized controlled trials generally have a simple null hypothesis – no effect – the calculation of the required number of study participants depends on the definition of a “minimum important difference”: the smallest difference between trial groups that the study will be able to detect. The DELTA (Difference ELicitation in TriAls) group recently provided a systematic overview of methods for specifying the target difference in randomized controlled trials [25].

In our exposition, we focused on the MAC for the sensitivity and specificity of a new test, and null hypotheses based on these criteria, to be used in the evaluation of single tests. In some cases, diagnostic accuracy studies that compare two or more tests or testing strategies can be more informative than single test evaluations. Hayen and colleagues have described how one can use meaningful measures and statistics in such studies, such as the difference in true and false positives, the likelihood ratios or relative likelihood ratios [26].

Mainly due to technological innovations, the field of diagnostic testing has been evolving quickly over the past decades. Premature incorporation of new diagnostic tests into clinical practice based on overoptimistic conclusions of diagnostic accuracy studies may lead to unnecessary testing, waste of resources, and faulty clinical decision making. Considering the requirements needed for a test to be clinically useful before initiating new diagnostic accuracy studies will lead improve methodological study quality and hopefully better and more meaningful evidence synthesis of such studies.
References


