**QUEEN'S
UNIVERSITY
BELFAST**

Fifteen-minute consultation: How good is this test?

Waterfield, T. (2019). Fifteen-minute consultation: How good is this test? *Archives of Disease in Childhood: Education and Practice Edition*. <https://doi.org/10.1136/archdischild-2019-317037>

Published in:

Archives of Disease in Childhood: Education and Practice Edition

Document Version:

Peer reviewed version

Queen's University Belfast - Research Portal:

[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

Copyright 2019 BMJ. This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Fifteen-minute consultation: How good is this test

1st Author:

Dr Thomas Waterfield

Queen's University Belfast

Belfast

Northern Ireland

Email: Thomas.waterfield@gmail.com

Word count: 1448

Abstract

As technology evolves and cultural attitudes towards diagnosis change there is an increasing move towards newer, faster and more accurate diagnostic testing. As new tests are developed clinicians are increasingly required to appraise data from diagnostic test accuracy (DTA) studies.

The accuracy of a test is fluid and changes depending upon the population, setting, timing and position within the diagnostic pathway. This article attempts to provide a short guide to understanding diagnostic test accuracy and a simple approach to appraising DTA studies.

Introduction

Every year that passes sees the introduction of a new test, or a new “cheaper” or “more rapid” way to test for something but how should we use these tests and should we use them at all?

In this age of evidence-based medicine it is vital that when we prescribe a treatment, we balance the risks and benefits of that treatment and consider its efficacy and cost. We rarely however, apply the same standards to the investigations we use (1). On the face of it this comparison may seem facile but consider for a moment the implications of a false positive cancer screening test and the stress and harm that could cause? Not only might the false positive test result in psychological stress it could result in physical harm through further invasive procedures. Also consider conversely using a rule out test for a potentially life changing disease only to find that test carries a high false negative rate. For any investigation there needs to be an understanding of the test accuracy, consideration of the risks and benefits and careful thought as to where in the diagnostic pathway the test could be best used (1).

Understanding test accuracy studies

Let's stick with the comparison between investigations and treatments and let's imagine for a moment that a pharmaceutical company wants to develop a new drug. Before they can bring it to market, they have to complete a series of clinical trials. This is a rigorous and structured process that ensures that

the product is safe, has a beneficial effect and has a defined treatment role compared to existing care. Now let's imagine they want to introduce a new diagnostic test. The process is far simpler with only a need to demonstrate that the testing is safe and reproducible (1). There is almost no mechanism for ensuring that the new test is accurate and useful.

Figure 1 compares the structured approach to clinical trials with the different designs of diagnostic test accuracy (DTA) studies. Unlike the clinical trials for a new therapy the DTA studies do not have to follow this stepwise approach and can be brought to market very early in the assessment process.

As shown in figure 1 a phase 1 clinical trial is the first step towards introducing a new therapy and is aimed at determining the safety and dosing of a new therapy. The closest comparison for a DTA study would be a case control study. In a case control study, the test is performed on diseased individuals and healthy controls. This type of study helps to determine if the test has any potential value in humans. A case control study however, cannot be used to provide meaningful data on the test's accuracy as it will overestimate the test's accuracy and provides no insight on how it will perform in a clinical setting.

Phase 2 clinical trials for new therapeutic agents focus on demonstrating the efficacy of the new therapy. The parallel for DTA studies would be a diagnostic accuracy study that conforms to STARD criteria. STARD criteria for reporting DTA studies help to minimise bias and provide a truer estimation of the test's accuracy in a clinical setting (1,2). Most STARD DTA studies report on the accuracy of a test in a small population of 100-300 people and provide little detail on the best testing strategy.

Larger multicentre STARD DTA studies that compare testing strategies can be likened to phase 3 clinical trials in which the new therapy is compared to existing treatments.

For any new therapies there are ongoing phase 4 trials after introduction into clinical practice to monitor the ongoing safety and efficacy. Similar studies in DTA are uncommon but would involve assessments of impact on care such as length of stay, survival, acceptability to patients. This could be performed as a randomised control trial of test implementation.

Appraising DTA studies

The lack of a rigorous assessment process for new diagnostic tests makes assessing the literature and marketing materials a little difficult but there are some key questions you can ask in addition to those you would ask of any study.

- Is this a case control study?

If it is a case control study then the purpose of the study is to simply demonstrate that the test has some use in identifying the diseased state. The reported accuracy has little relevance to routine clinical practice and this type of study should be seen as unreliable and at high risk of bias (1,2).

- Does the study adhere to STARD criteria for reporting for DTA studies?

The EQAUTOR network have produced a standardised reporting tool for DTA studies (1,2). This helps to identify bias and improve the quality of DTA studies. If a study doesn't include a STARD checklist then consider it as at high risk of bias.

- Are there significant sources of bias?

Table 1 outlines the common sources of bias in DTA studies. A high-quality study will have a clearly defined population that represents the clinical application of the test. There will be a clear and reliable reference test that was performed blinded to the index test result (1,2).

- Does the studied population reflect the intended clinical use?

The DTA study will typically report on the use of the test in a specific population at a specific point in the diagnostic pathway. The study should reflect the intended clinical use. The diagnostic accuracy of a test is not fixed but fluid and varies with the population, timing and pre-testing (1). For example, consider a new sepsis biomarker that was studied by testing all paediatric intensive care unit (ICU) admissions on arrival to the ICU. The study was performed according to the STARD criteria and the test performs exceptionally well and was able to identify all children with sepsis from the other admissions. On the back of this a paediatric emergency department (ED) invests in this test. They use the new test as a rapid test for sepsis in all febrile children attending the ED but find that the test

performs poorly. This may reflect that the test in the DTA study was performed later in the diagnostic pathway (after admission to ICU) and in a different cohort of patients (ICU admissions). It could be that the new biomarker rises late in the illness by which time the child is already very unwell. This would not have been demonstrated by the original DTA study.

Lies and statistics

There are a number of measures of diagnostic accuracy as outlined in Table 2 (3-5). The majority of these statistical tools are designed for measuring and comparing the accuracy of tests but have little to no clinical application (3-5).

The most clinically relevant and most easily understood measures of DTA are the negative (NPV) and positive predictive values (PPV). The NPV describes the probability of not having a disease in a subject with a negative test result. This means that if a test has a NPV of 99% then in an individual with a negative result there is only a 1% risk of disease. The issue with NPV and PPV is that the measure is a combination of test accuracy and disease prevalence.

For example, let's consider again the new sepsis biomarker from earlier. The manufacturers perform a follow up study in general practice and they test every child attending the practice over 12 months. They then report that the NPV is 99.5% based on testing of 2000 children and that the test can be used as a rule out test in primary care. On the face of it these results make a compelling case for the new test.

However, the study population had a very low rate of sepsis with only 10 children having sepsis during the study. This means that a coin toss (a useless test) would have the same NPV. Figure 2 demonstrate DTA summary of a coin toss for diagnosing sepsis in the same population.

It is important to be analytical of the statistics and consider the relevance of the findings to clinical practice. Unfortunately, very few DTA studies report on how the new test impacts on patient care. The most meaningful studies compare how a new testing strategy effects patient outcome.

Applying it to practice

As demonstrated, it is difficult to apply DTA study findings to our daily practice. When reviewing the literature

- Consider the level of the evidence and the potential risks of bias in the study.
- Avoid drawing an inference from case control studies and studies at high risk of bias.
- Consider if the study population reflects the intended clinical application.
- Consider where in the diagnostic pathway the test should be used.
- Ask if the analysis is transparent and meaningful.
- Consider the cost of the new test.

References

- 1) STARD Criteria available from; <http://www.equator-network.org/reporting-guidelines/stard/> - last accessed 04/03/2019.
- 2) Cohen JF, Korevaar DA, Altman DG et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016 Nov 14;6(11):e012799. doi: 10.1136/bmjopen-2016-012799.
- 3) Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004;17;329(7458):168-9.
- 4) Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem*. 2004;50(7):1118-25
- 5) Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;56(11):1129-35.

Table 1: Sources of bias in Diagnostic Test Accuracy (DTA) studies

Flawed study due to population chosen	
Spectrum bias	The study population doesn't represent the intended clinical application e.g. a case control study where only the diseased and healthy controls are tested.
Sample bias	Use of the test is not consecutive or random within the population e.g. a convenience sample where a test is only applied at clinician discretion.
Flawed study due to reference standard	
Inappropriate reference standard	The reference standard poorly diagnoses the target condition.
Verification bias	The reference standard is only applied dependent on the result of the index test
Disease progression bias	The index test and reference standard are applied at very different time points after which the disease has changed.
Flawed due to poor blinding	
Diagnostic review bias	The reference test is performed without blinding to the index test or clinical condition.
Incorporation bias	The index test is required for the diagnosis of the target condition.
Flawed due to poor analysis of results	
Indeterminate result bias	Indeterminate results are excluded from analysis
Threshold cherry-picking	A cut-off is determined as a "best fit" based on the data providing an exaggeration of DTA.

Table 2: Statistical measures of test accuracy

	Definition	Clinical Relevance	Disadvantage	Advantage
Sensitivity	Proportion of subjects with the disease with a positive test result.	A high sensitivity makes for a good rule out test.	Difficult to apply to clinical practice.	Disease prevalence has a minimal effect. Good for comparing tests in different populations.
Specificity	Proportion of subjects without the disease with a negative test result.	A high specificity makes for a good rule in test.		
Negative Predictive Value (NPV)	Probability of not having a disease in a subject with a negative test result	A high NPV makes for a good rule out test.	Accuracy dependent on disease prevalence. A rare disease will have a high NPV even with an inaccurate test.	Easy to apply clinically e.g. NPV of 99% equates to a 1% risk of disease if the test is negative.
Positive Predictive Value (PPV)	Probability of having a disease in a subject with a positive test result	A high PPV makes for a good rule in test.		
Likelihood Ratio (LR)	The ratio of test result in subjects with disease to the subjects without disease	A LR greater than 1 indicates increased likelihood of disease. A LR under 1 indicates reduced likelihood of disease.	Difficult to apply to clinical practice.	Links the test to the pre and post-test probability.
Diagnostic Odds Ratio	Ratio of the odds of positivity in subjects with disease relative to the odds in subjects without disease	None as the odds only relate to the test and not disease.	Clinically irrelevant	None
Receiver Operator Characteristics (ROC) Curve	A graphical representation of the sensitivity and specificity of a test over the full range of cut-off values.	Can help identify an optimal cut-off that balances sensitivity and specificity of a test.	Difficult to apply to clinical practice.	Useful for comparing tests.

Figure 1: Visual comparison of clinical trial and Diagnostic Test Accuracy (DTA) studies

Figure 2: Diagnostic accuracy of a coin-toss in a low prevalence disease