

TB Evidence Glossary

Dr. David Sackett defined evidence-based health care as the integration of 3 principles: the best research evidence, clinical expertise, and patient values and preferences. **Evidence-Based Tuberculosis Diagnosis** is based on these same principles. **Evidence** is information used to form a conclusion. In the context of a systematic review of diagnostic test **accuracy**, **quality of evidence** is the extent to which one can be confident that an estimate of effect or test characteristic (such as **sensitivity or specificity**) is correct. Quality may be higher or lower depending on study methods, applicability, and other factors.

Accuracy is a measure of the proportion of test results that are correct and encompasses an assessment of **sensitivity** and **specificity**.

Confidence Interval (also called CI) is a measure of the uncertainty around the main finding of a statistical analysis. Estimates of diagnostic test accuracy are usually expressed as a point estimate of a test characteristic, for example **sensitivity**, and a 95% confidence interval. This means that if someone were to keep repeating a study in other samples from the same population, 95% of the confidence intervals from those studies would contain the true **sensitivity**. Wider intervals indicate lower precision; narrow intervals, greater precision. See **Imprecision**.

The **GRADE** (Grading of Recommendations Assessment, Development and Evaluation) approach to developing health care recommendations has been adopted by a large number of organizations worldwide. The GRADE approach specifies 4 categories of quality of evidence: high, moderate, low, and very low. Grading the quality of evidence about the use of diagnostic methods begins with a consideration of study design and may be compromised by 5 factors: **risk of bias**, **indirectness of the evidence**, **inconsistency of the results**, **imprecision of the estimate**, and **risk of publication bias**.

- **Bias** is a systematic deviation from the true results. Flaws in the design and execution of a diagnostic accuracy study can produce incorrect results and may lead to **risk of bias** in the study. Examples are the following: use of an inappropriate **reference standard**; unsuitable manner of selecting study participants; and interpretation of test results without blinding. A standardized tool to evaluate bias in studies of diagnostic accuracy is the QUADAS tool. You may want to know: *Was the best reference standard used? Was the right group of patients recruited and were they selected consecutively? Was the result of the test under investigation interpreted without knowledge of the result of the reference standard?*

- **Indirectness** encompasses 2 concerns occurs: 1) a comparison of test A versus B is not available, but A was compared with the reference standard and B was compared with the reference standard. Such studies allow indirect comparisons of the results of A versus B. Such evidence is of lower quality than head-to-head comparisons of A and B would provide. 2) Indirectness related to population, intervention, comparator, or outcome – the question being addressed by the guideline panel or by the authors of a systematic review is different from the available evidence regarding the population, intervention (diagnostic test or strategy), comparator, or outcome. In the context of diagnostic tests, the indirectness of the outcomes (diagnostic accuracy is often used as a proxy for patient-important outcomes) is of particular importance. You may want to know: *Was there a direct comparison of Test A and Test B? Can we confidently apply these results to different groups such as children or people with HIV? Did the studies only report on accuracy or also patient-important outcomes, such as mortality, time to diagnosis, and number of people without TB mistakenly treated?*

- **Inconsistency** (also called unexplained heterogeneity) refers to differences among the results found in research studies. In **systematic reviews** of diagnostic test accuracy, large differences in results are commonly noted among studies and may be due to differences in test properties (such as the cutoff point used to define a positive test result), populations, methods, and study quality. Specialized statistics are used to account for these differences. You may want to know: *Did the studies have similar or widely varying results?*

- **Imprecision** depends on the size of the sample and may be judged by the width of the **confidence interval** (CI). Consider the statement - the pooled **sensitivity** estimate is 76% (95% CI 45-92). The point estimate (76%) is the best estimate of the true sensitivity. The CI describes the uncertainty underlying this value and provides a range of values (in this example, between 45 and 92) within which the true value likely lies (with 95% confidence). If the

95% CI is relatively narrow (e.g. 70-80), the true value is known with *precision*. If the interval is wider (45-92), there is less certainty about the true value, raising concerns about *imprecision*. You may want to know: *How many studies were combined to get these estimates? How many patients did they include; and how wide were the confidence intervals around the point estimates?*

- **Publication bias** is a systematic under or overestimate of study results due to selective publication of studies. Publication bias arises when investigators fail to report all studies they have undertaken, for example, those studies that found a diagnostic test to be inaccurate. There are no adequate methods to detect the possibility of publication bias in systematic reviews of diagnostic test accuracy studies and authors of the reviews must often guess about the likelihood of publication bias. A situation that should increase suspicion of publication bias occurs when published evidence is limited to a small number of studies, most of which were funded by industry. You may want to know: *Is it possible that more studies evaluating this diagnostic test were conducted, but some of these studies were difficult for the systematic reviewers to obtain?* (Note: publication bias is distinct from the bias related to study design and execution described on the previous page).

Meta-analysis is the use of statistical techniques in a systematic review to summarize (pool) the results of included studies. Not all **systematic reviews** include a meta-analysis.

Null hypothesis is a statistical hypothesis of there being no association between a factor of interest (e.g. the test under evaluation) and the outcome (e.g. sensitivity). For example, a null hypothesis could be stated as, "There is no difference between the average sensitivity of fluorescence microscopy and the average sensitivity of conventional microscopy for the diagnosis of active TB".

P-value indicates the probability (ranging from zero to one) that chance would produce the result found in the study (or more extreme results) if the **null hypothesis** were true. The smaller the P-value, the less likely the null hypothesis is true. For example, fluorescence microscopy was found to be on average 10% more sensitive than conventional microscopy, $P < 0.001$ (Lancet Infect Dis 2006; 6:570–81).

Reference standard is the best available test or combination of tests for establishing the presence or absence of a disease in patients. The reference standard does not need to be absolutely perfect. In a systematic review of TB diagnostic studies of pulmonary TB in adults, the reference standard is often solid or liquid culture. For TB in children, the reference standard may include clinical and radiographic criteria in addition to culture, as culture performs poorly for diagnosing TB in children. For latent TB infection, which lacks a "gold standard" to confirm the diagnosis, culture-confirmed active TB may serve as an (imperfect) reference standard.

Sensitivity of a test is defined as the probability that the result of the test will be positive in a diseased case (according to the **reference standard**). For example, if a patient has TB, how likely is she to have a positive test? Sensitivity is expressed either as a proportion (ranging from zero to one) or a percentage. $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$, where TP is the number of true positive test results and FN is the number of false negative test results.

Specificity of a test is defined as the probability that the result of the test will be negative in a non-diseased case (according to the **reference standard**). For example, if a patient does not have TB, how likely is she to have a negative test? Specificity is expressed either as a proportion (ranging from zero to one) or a percentage. $\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$, where TN is the number of true negative test results and FP is the number of false positive test results.

Systematic review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review. A systematic review of diagnostic test accuracy studies concerns measures of test performance such as **sensitivity** and **specificity**.