Study Designs: Diagnostic Studies

NITIN DHOCHAK, RAKESH LODHA

From Department of Pediatrics, All India Institute of Medical Sciences, New Delhi, India.
Correspondence to: Dr Rakesh Lodha, Department of Pediatrics, All India Institute of Medical Sciences, New Delhi 110 029.
rakesh_lodha@hotmail.com

Diagnostic tests are evolving with betterment of technology, quest for patient safety with less invasive medicine, and evolution of new diseases. It is important to assess diagnostic accuracy of a new test, and clinical impact of introduction of new test on outcomes and cost. A diagnostic study is planned for the index test based on place of new test in diagnostic pathway (screening, triage, diagnostic or add-on test) and established information of the test. A reference standard is used to classify population into diseased and healthy, and the discriminating ability of index test is measured. A sample size is calculated for expected sensitivity/specificity, margin of error and prevalence of disease in population. For dichotomous outcomes, sensitivity, specificity, predictive values and likelihood ratio are used to describe diagnostic accuracy. Efforts should be made to avoid common forms of bias including spectrum bias and partial verification bias, and blinding of observers should preferably be done.

Keywords: *Diagnostic accuracy, Index test, PPV, ROC, Sensitivity, Specificity.*

Published online: April 20, 2021; ***PII:*** S097475591600317

Diagnostic tests evolve with development of newer medical technologies and refinements of older technologies. With focus on patient safety, there is a trend towards increasing use of non-invasive tests (doppler monitoring cardiac output *vs.* conventional invasive catheterization) and radiation free imaging modalities like ultrasound and magnetic resonance imaging (MRI) of chest vs. chest *X*-ray and computed tomography (CT). There has been a recent interest in various biomarkers for diagnosis and prognosis. Though these new diagnostic tests appear attractive for clinicians, it is equally important to ascertain that diagnostic accuracy is not significantly compromised over conventional reference standard tests. Studies evaluating diagnostic tests utilize unique methodology and statistical methods. We, hereby, review the methodology, statistics and pitfalls while performing clinical studies to evaluate diagnostic tests.

Indications of Diagnostic Studies

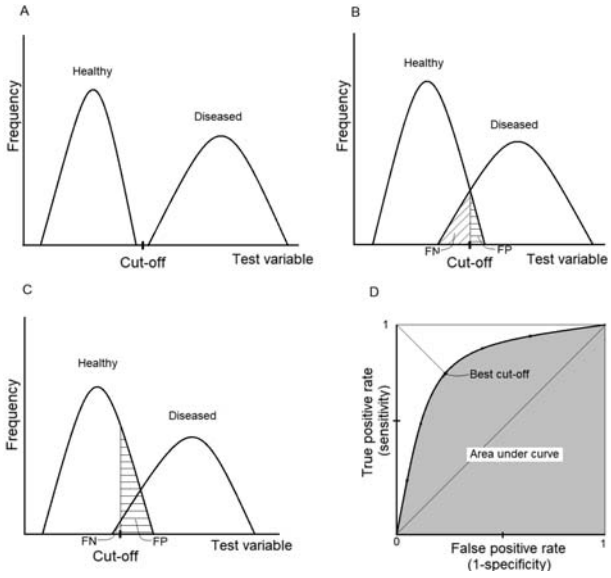
A diagnostic test is based on differential expression of certain characteristic among diseased, affected at-risk and healthy population. It could be a molecule of metabolic pathways (e.g. lactate for shock, creatinine for renal failure), or as a combination with clinical features (e.g. eschar for scrub typhus, PICADAR score for primary ciliary dyskinesia). An ideal diagnostic feature should not overlap between diseased and general population (**Fig. 1A**). However, for a continuous variable (e.g. lactate), a cut off is decided to differentiate diseased and healthy population with minimum overlap (**Fig. 1B**).

An algorithmic approach in the diagnostic pathway guides the characteristics of the test (**Fig. 2**). The various types of tests are:

a. *Screening test*: A screening test is used to identify individuals who are diseased/at high risk among asymptomatic population. Screening test should be highly sensitive to identify most of the diseased population, while they might also be positive in healthy individuals (lower specificity, often a trade-off for high sensitivity); e.g. immunoreactive trypsinogen (IRT) for cystic fibrosis (CF) in neonates [1]. Patients positive on screening test should undergo confirmatory test to corroborate the diagnosis.

b. *Triage test*: Triage test are utilized for screening positive population to further decrease number of individuals requiring confirmatory diagnostic test. This approach is useful when confirmatory test is expensive, inaccessible or invasive. For example women with positive screening on pap-smear are traditionally subjected to invasive tests including colposcopy. Introduction of triage test (human papilloma virus (HPV) test) reduces the number of patients needing colposcopy without additional risk of missing cervical malignancy [2]. Triage test should be highly sensitive and reasonably specific.

c. *Diagnostic test*: Diagnostic test confirms presence of a disease in screen positive population or individuals coming to clinics with symptomatic diseases. Diagnostic tests are desired to have high sensitivity as



FN: false negative, FP: false positive.

Fig. 1 A. Ideal diagnostic test with no overlap of measurements between diseased and healthy population. **B.** Diagnostic test demonstrating overlap of measurements with cutoff for best diagnostic accuracy. **C.** A screening test with lower diagnostic cutoff. **D.** Receiver operating characteristic curve for a diagnostic curve. Cut-off for best diagnostic accuracy corresponds to the point nearest to left upper corner of the graph.

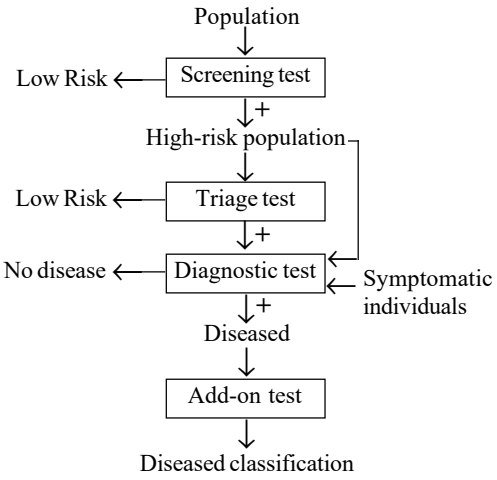


Fig. 2 Diagnostic pathway demonstrating place for various tests.

well as high specificity, e.g. sweat chloride assay for confirming diagnosis of CF in symptomatic neonates with elevated IRT or in a child with recurrent pneumonia.

d. Add-on diagnostic test: Add-on tests are used to increase sensitivity or specificity of current established diagnostic tests. These tests can be used with established test as either positive (to decrease false negative) or both positive (to decrease false positive) approach for starting treatment. These tests are usually costly, or invasive, but might be useful in subset of population where diagnostic test have limitations. For example, the use of positron emission tomography for distant metastasis where conventional imaging (CT or MRI) is inconclusive [3].

The diagnostic accuracy of a new test or new indication of an old test may be evaluated in any of these situations [3]:

1. *Replacement*: New screening or diagnostic test may have superior diagnostic accuracy over conventional diagnostic algorithms. For example, *c*omparison of GeneXpert with sputum smear for diagnosis of tuberculosis. It may instead have similar efficacy but can be less-expensive, faster, non-invasive, less radiation exposure. For example, MRI chest instead of CT for follow-up of mediastinal pathologies.
2. *Triage*: Addition of new test in diagnostic pathway like HPV test in cervical cancer screening in population with positive pap smear, decreases need for invasive testing without additional risk of missing malignancy.
3. *Add-on*: To test benefit of an add-on test on existing diagnostic pathways.

Study Designs for Evaluation of Diagnostic Tests

Similar to clinical trials for new interventions, diagnostic studies can also be classified as four phases [4]. *Phase 1* studies focus on establishing a normal range for the new test. It involves cross-sectional observational studies with random sampling of healthy subjects from the population.

Phase 2 studies focus on establishing diagnostic accuracy of the new test. These include case control, or cohort studies with healthy subjects and diseased patients, aimed at establishing cut-offs, sensitivity, specificity, predictive values, and likelihood ratios for the new test. These studies also include comparison of diagnostic accuracy of a new test with a reference test, like comparison of sweat chloride estimation and sweat conductivity for diagnosis of CF [5], or diagnostic accuracy of QuantiFERON-TB gold test and tuberculin skin test for diagnosis of tuberculosis [6]. These studies are paired and have advantage of smaller sample size, and less bias due to heterogeneity of population. Randomized trial study design is preferred in situations where paired study cannot be performed because of interference of one test with another or invasive nature of tests. It is desirable to evaluate a diagnostic test in a diseased population similar to the final population where it is likely to be used.

For example, a rapid diagnostic test for enteric fever should be tested in children with fever, all of whom would undergo testing for enteric fever with blood culture. This approach will be preferable to a study recruiting patients with culture confirmed enteric fever and healthy individuals.

Phase 3 studies establish clinical impact of new diagnostic test in diagnostic pathway with respect to patient benefit and harm. These involve randomized trials where individuals undergo new test or comparator test, and outcomes and further treatment depends on the results of these tests. Outcome parameters include change in diagnosis, change in treatment choices, patient outcomes, and cost-effectiveness. A non-inferiority randomized trial of procalcitonin guided antibiotic administration to adults with acute respiratory infection is an example of addition of a triage test [7]. The potential benefits of procalcitonin guided regimen are decrease in antibiotics administration while concerns/potential harm are adverse clinical outcome such as treatment failure, or increased hospital stay.

Phase 4 studies are follow-up studies to determine clinical impact in different settings. These studies are aimed at establishing diagnostic accuracy of a new test and clinical impact of introduction of new test (triage/add-on) in clinical pathway, like efficacy of clinical scores in predicting mortality or guiding hospitalization.

Measurement of Diagnostic Accuracy

The aim of diagnostic studies is estimation of ability of the test to discriminate diseased from healthy individuals. The discriminative ability of index test (test being evaluated) is compared with a reference standard test. For tests with dichotomous outcome (positive or negative), a 2 X 2 contingency table can be prepared (**Table I**). Parameters assessed include sensitivity, specificity, predictive values, and likelihood ratio.

Sensitivity and Specificity

Sensitivity is the ability of the test to detect individuals who have disease (or a condition), while specificity is the ability to detect individuals who do not have disease (or a condition). These can be calculated as below:

$$Sensitivity = \frac{True\ positive}{All\ with\ disease} = \frac{a}{a+c}$$

Table I Contingency Table for Tests with Dichotomous Outcomes

		Reference standard	
		Diseased	*Healthy*
Index test	Positive	a (true positive)	b (false positive)
	Negative	c (false negative)	d (true negative)

$$Specificity = \frac{True\ negative}{All\ without\ disease} = \frac{d}{b+d}$$

Sensitivity and specificity depend on distribution of measurement parameter between diseased and healthy individuals and ability (accuracy and precision) of the index test to measure the parameter. These do not depend on prevalence of the disease. However, they are mutually dependent according to the cut-off of the test. As in **Fig.1B** (best diagnostic accuracy) and **Fig. 1C** (lower cut-off), more diseased patients are detected if a lower cut-off is used (sensitivity increases) but simultaneously more healthy individuals are classified as diseased (specificity decreases).

Predictive Values

While sensitivity and specificity describe discriminating characteristics of the test, it is hard for a clinician to understand the significance of an individual positive or negative test based on these parameters. Positive predictive value (PPV) is the proportion of a true positive tests among all positive tests. Similarly, negative predictive value (NPV) is the proportion of true negative tests among all negative tests.

$$PPV = \frac{True\ positive}{All\ positive} = \frac{a}{a+b}$$

$$NPV = \frac{True\ negative}{All\ negative} = \frac{d}{c+d}$$

PPV and NPV depend on test characteristics (sensitivity and specificity) as well as prevalence of the disease. For example, a test kit for dengue IgM with known sensitivity (0.9) and specificity (0.9) disease may be used for 1000 febrile patients in region A (50% of febrile patients have dengue) and B (10% febrile patients have dengue) each (**Table II**). In region A, PPV = 450/(450+50) = 0.9 while in region B, PPV = 90/(90+90) = 0.5. In region A, NPV = 450/

Table II Contingency Table for IgM Dengue Tests for Two Regions With Different Prevalence (Hypothetical Data)

IgM dengue	*Region A n=1000*		*Region B n=1000*	
	Dengue	*other febrile illnesses*	*Dengue*	*other febrile illnesses*
Positive	450	50	90	90
Negative	50	450	10	810
Total	500	500	100	900

(50+450) = 0.9 while in region B, NPV = 810/(810+10) = 0.99. PPV for a test increases with increase in prevalence/ pre-test probability while NPV decreases with increase in prevalence/ pre-test probability.

Likelihood Ratios

Likelihood ratio (LR) represents the ratio of post-test odds to pre-test odds.

Post - test odds = Likelihood ratio X Pre-test odds

Positive LR (LR+) is ratio of likelihood of positive result in a diseased individual to likelihood of positive result in healthy individual. Negative LR (LR-) is the ratio of likelihood of negative test result in a diseased individual to likelihood of negative result in healthy individual. Higher LR+ and lower LR- are desired. LR+ of 10, 6, 2, and 1, and LR- of 0.1, 0.2, 0.5 and 1 are classified as excellent, very good, fair and useless test. LR can be calculated as:

$$LR+ = \frac{a/(a+c)}{b/(b+d)} = \frac{sensitivity}{1 - specificity}$$

$$LR- = \frac{c/(a+c)}{d/(b+d)} = \frac{1 - sensitivity}{specificity}$$

LR is the ratio of post-test and pre-test odds and not probability. For using LR for estimation of post-test probability, odds can be converted into probability by the following equations:

Probability = odds / (odds + 1)

Or,

$$P_1 = P_0 \text{ X } LR/(1 - P_0 + P_0 \text{ X } LR)$$

where P_1 is post-test probability and P_0 is pre-test probability.

More commonly, Fagan nomogram is used for post-test probability estimation from pre-test probability and LR [8].

Diagnostic accuracy of a test can be calculated as proportion of true positive and true negative results among all tests:

$$Accuracy = \frac{a+d}{a+b+c+d}$$

Diagnostic accuracy/discriminatory power for tests measuring continuous variable (for *e.g.* creatinine, blood glucose) with dichotomous outcomes (*e.g.* acute kidney injury: yes/no, diabetes: yes/no), can also be represented as area under (AU) receiver operating characteristic (ROC) curve. ROC curve is plotted with true positive rate (sensitivity) on y-axis and false positive rate (1-specificity) on x-axis for different cut-offs of the test (**Fig. 1D**). AUC of 0.5 to 0.6 is almost useless, 0.6 to 0.7 is poor, 0.7 to 0.8 is fair, 0.8 to 0.9 is very good and >0.9 is excellent.

Designing Diagnostic Studies

First step in any diagnostic study is identification of existing clinical pathway which will include the index test. Role of index test as screening, triage, diagnostic or add-on test has to be clearly defined. Expected proportion of patients with target disease among the general population is estimated based on prevalence studies or meta-analysis. Most diagnostic studies are conducted on population cohort where a proportion of individuals have a target condition but are not diagnosed. Case-control approach is more appropriate in conditions with low prevalence. Impact of the index test on the study population is ascertained, and minimally acceptable criteria (MAC) for sensitivity and specificity are decided and study hypothesis is established [9].

Sample Size Estimation

Sample size of the study is related to expected sensitivity and specificity, maximum margin of error (usually set as 0.05 or 0.02, lower limit of confidence interval should not cross MAC), α- and β-error [9]. Sample size is estimated separately for sensitivity and specificity for required individuals with target condition and without target condition respectively (true for case-control studies). In cohort studies, where diagnosis is not established in beginning, sample size is adjusted for prevalence of the target condition in population. Formula for calculating sample size for diagnostic studies is given in **Table III** [10]. Similarly, sample size can also be calculated for studies for estimating diagnostic accuracy of new test or comparison between tests on basis of predicted AU-ROC.

Statistical Analysis: A reference standard is required which could be a diagnostic test, or combined classification based on clinical tests and diagnostic test, to identify individuals with target condition/ disease amongst the enrolled population. The index test is applied to the same sample and the ability to correctly categorise into patients with or without target condition is compared with the reference standard.

Testing diagnostic accuracy of a new test: Minimally acceptable criteria (MAC) for the index test are pre-defined based on place of diagnostic test in clinical pathway. For a screening test, MAC for sensitivity will be kept at high level of greater than 0.85-0.9 while for a diagnostic test, specificity is equally important. The diagnostic accuracy parameters such as sensitivity and specificity are described with 95% confidence interval (CI), lower limit of which should not cross MAC. For example, the diagnostic accuracy of chest X-ray to differentiate bacterial and viral

pneumonia in children was based on combination of tests including viral culture and antigen testing from nasopharyngeal aspirate, and IgM and paired IgG serology for acute and convalescent samples for bacterial and viral antigens (reference standard). Sensitivity and specificity of alveolar infiltrates on chest *X*-ray for identification of bacterial pneumonia was 0.72 and 0.51, respectively. No pre-determined MAC were reported in this study [11].

Establishing cut-off: A diagnostic cut-off needs to be established for a new diagnostic test measuring a continuous variable. A lower cut-off (targeting high sensitivity) will be advised for a screening test or test identifying highly infectious and lethal illness requiring isolation. If there is no preference for sensitivity or specificity, cut-off for best diagnostic accuracy can be identified by various methods like Youden's index, point of minimal distance from top left corner of ROC curve (**Fig. 1D**), using Bayesian approach or analytical methods (numbers needed to misdiagnose) [12]. Cut-off with maximum Youden's index (sensitivity + specificity -1) is chosen.

Comparing Diagnostic Accuracy of Two Tests: For a new test aimed at replacing older test, diagnostic accuracy of both the tests is compared by AU-ROC for of both tests [13].

Comparison when there is no gold standard: There may be no accurate reference test or it may not have been performed on all individuals included in study because it is expensive or invasive. Alternatives for this situation such as imputation and bias correction methods, and differential verification (when reference standard is missing), correction methods, and use of multiple imperfect reference standard (when reference standard is imperfect are described). Other methods are study of agreement, true positivity rate or analytical validation for new test and imperfect reference standard instead of usual diagnostic accuracy tests [14].

Pitfalls

Inadequate sample size: Sample size estimation is frequently omitted in diagnostic studies. In a survey of diagnostic studies, only 2 out of 40 (5%) reported sample size calculation [15]. Inadequate sample size leads to loss of power of study while large sample size adds to cost and complexity of diagnostic studies. *A priori* sample size calculation should be done in all diagnostic studies.

Intra- and Inter-observer variability: Tests involving complex procedures, multiple steps, and subjective parameters can have significant variability when performed repeated by same observer (intra-observer variability) and different observers (inter-observer variability). For dichotomous outcomes, agreement between two observers can be simply calculated as proportions of test results agreed by both the observers (positive as well as negative). This method doesn't account for inter-observer agreement due to knowledge of prevalence of disease, which is adjusted while estimating a better parameter as kappa statistics. Kappa ranges from -1 (perfect disagreement) to 1 (perfect agreement) and values above 0.8 are considered very good and 0.6-0.8 are considered good [16]. For continuous outcomes, inter-observer variability can be expressed as coefficient of variation (=standard deviation/mean). Mean difference between paired measurements by two observers can also be assessed by Bland-Altman analysis [17].

New test is more sensitive than gold standard: Newer test especially molecular tests such as polymerase chain reaction (PCR) for detection of infectious agents are at-times more sensitive than existing gold standard/ reference test. When diagnostic accuracy is calculated for such test, both sensitivity and specificity are underestimated. For example, in a study comparing efficacy of different tests for tuberculosis, culture identified 50/125 (40%) samples as positive, while GeneXpert ultra was positive in 73/120

Table III Sample Size Estimation for Various Diagnostic Studies [10]

Study design	*Sample size*	
Diagnostic accuracy of a new test with dichotomous outcome		
a. Case-control	$cases = Z_{\frac{\alpha}{2}}{}^2 \frac{Se\ (1-Se)}{d^2}$	$controls = Z_{\frac{\alpha}{2}}{}^2 \frac{Sp\ (1-Sp)}{d^2}$
b. Cohort study (Use larger of the samples derived from sensitivity and specificity formulas)	$n = Z_{\frac{\alpha}{2}}{}^2 \frac{Se\ (1-Se)}{prevalence \times d^2}$	$n = Z_{\frac{\alpha}{2}}{}^2 \frac{Sp\ (1-Sp)}{(1-prevalence) \times d^2}$
Sample size for comparing the sensitivity (or specificity) of two diagnostic tests	$n = \frac{[Z_{\frac{\alpha}{2}}\sqrt{2P(1-P)} + Z_{\beta}\sqrt{P_1(1-P_1)+P_2(1-P_2)}]}{(P_1-P_2)^2}$	

P_1: Sensitivity or specificity of test 1; P_2: Sensitivity or specificity of test 2, P: average of P_1 and P_2; n: sample size; Se: Sensitivity; Sp: Specificity; $Z_{\frac{\alpha}{2}} = 1.96$; $Z_{\beta} = 0.84$.

Key messages

- Study design of a diagnostic study should be planned based on place of new diagnostic test in diagnostic pathway.
- *A priori* sample size estimation should be conducted in all diagnostic studies.
- Reference standard should be carefully chosen especially in cases where new diagnostic test could potentially have better diagnostic accuracy than established gold-standard.
- Blinding of assessors should be performed to avoid bias.

(60.8%) samples [18]. The reported sensitivity and specificity of GeneXpert ultra was 88% and 58.6%. Large number of patients who were detected on GeneXpert ultra were labelled as false positive which led to significant underestimation of sensitivity and specificity. In these situations, it is better to consider alternative method of reference (clinico-radiological diagnosis of tuberculosis as reference standard in above example) and compare diagnostic accuracy of new test and established gold standard.

Bias

Source of bias in a diagnostic study can arise from patient selection, index test method, reference test or study flow and outcomes. QUADAS-2 tool is used for assessing risk of bias in diagnostic studied included in systematic review and meta-analysis [19]. Common sources of bias are described below [20]:

Patient selection: It is easier for a diagnostic test to differentiate a severely ill patient from healthy individual. Studies which include only severely ill patients are prone to overestimate diagnostic accuracy of the index test. This is called spectrum bias. Spectrum bias is also likely to be higher in case-control study design where cases are typical disease phenotypes. If possible, cohort study design should be utilized for diagnostic accuracy studies. The severity of illness of the study population should be reported.

Similarly, if the center conducting the study is a referral center, patients who clearly have the target condition or do not have the target condition, get diagnosed at the referring center. So, the referral center gets mostly patients with overlapping features and applying index test in such situation is likely to underestimate the diagnostic accuracy of the test.

Index test: Methodological differences can make significant differences in performance of the test. Difference in yield of a fine needle aspirate (FNA) could vary between studies due to differences in staining methods, use of rapid on-site evaluation, experience of physician performing aspirate, use of small or larger needle or use of ultrasound guidance. Hence, it is very important to describe methodology of index test in great detail and use same method for all procedures during the study.

Reference test: An imperfect reference test can lead to misclassification of the population. This is likely to underestimate sensitivity and specificity of index test.

Patient flow: If only a fraction of patients is undergoing reference test (if too invasive or costly), it is possible that patients negative in index test receive more intensive reference standard testing. Or if reference test is performed more frequently in patients positive on index test (*e.g.* invasive biopsy following a positive FNA). These could introduce partial verification bias.

If index test and reference test are done in sequence and the observer is aware of index test results, his interpretation of reference test can be biased. For example, the interpretation of a CXR or CT of patients with interstitial lung disease may be biased if biopsy results are known before. Similarly, assessment could be biased if observer assessing clinical outcomes knows about of diagnostic algorithm used. Observers estimating index test should be blinded from result of reference test and vice-versa, and observer assessing clinical outcomes and adverse effects should be blinded from both index and reference test results.

Reporting

Standard reporting guidelines for diagnostic studies are standards for reporting of diagnostic accuracy update 2015 (STARD-2015) and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [21,22].

CONCLUSION

With evolution of technology and trend towards medical safety, increasing number of new and safer tests are being available. Appropriate study design based on place of test in diagnostic pathway and calculated sample size will help in developing reliable evidence for their use.

*Contribut*ors: ND: was involved in review of literature, and writing manuscript; RL: was involved in review of literature, and writing and reviewing the manuscript.
Funding: None; *Competing interest*: None stated.

REFERENCES

1. Paracchini V, Seia M, Raimondi S, et al. Cystic fibrosis newborn screening: Distribution of blood immunoreactive trypsinogen concentrations in hypertrypsinemic neonates. JIMD Rep. 2012;4:17-23.
2. Macedo ACL, Gonçalves JCN, Bavaresco DV, Grande et al. Accuracy of mRNA HPV tests for triage of precursor lesions and cervical cancer: A systematic review and meta-analysis. J Oncol. 2019;2019:6935030.
3. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: Assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089-92.
4. Momeni A, Pincus M, Libien J. Designing diagnostic studies. In: Momeni A, Pincus M, Libien J, editors. Introduction to Statistical Methods in Pathology. Springer International Publishing; 2018. p. 279-92.
5. Rueegg CS, Kuehni CE, Gallati S, et al. Comparison of two sweat test systems for the diagnosis of cystic fibrosis in newborns. Pediatr Pulmonol. 2019;54:264-72.
6. Lodha R, Mukherjee A, Saini D, et al. Role of the QuantiFERON®-TB Gold In-Tube test in the diagnosis of intrathoracic childhood tuberculosis. Int J Tuberc Lung Dis. 2013;17:1383-8.
7. Briel M, Schuetz P, Mueller B, et al. Procalcitonin-guided antibiotic use *vs* a standard approach for acute respiratory tract infections in primary care. Arch Intern Med. 2008;168:2000-7.
8. Caraguel CGB, Vanderstichel R. The two-step Fagan's nomogram: Ad hoc interpretation of a diagnostic test result without calculation. Evid Based Med. 2013;18:125-8.
9. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: A framework for designing diagnostic accuracy studies with clear study hypotheses. Diagn Progn Res. 2019;3:22.
10. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014;48:193-204.
11. Virkki R, Juven T, Rikalainen H, Svedström E, Mertsola J, Ruuskanen O. Differentiation of bacterial and viral pneumonia in children. Thorax. 2002;57:438-41.
12. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. Biochem Med. 2016;26:297-307.
13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A non-parametric approach. Biometrics. 1988;44:837-45.
14. Chikere CMU, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update. PLoS One. 2019;14:e0223832.
15. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: Literature survey. BMJ. 2006;332:1127-9.
16. Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med. 2005;37:360-3.
17. Giavarina D. Understanding Bland Altman analysis. Biochem Med. 2015;25:141-51.
18. Osei Sekyere J, Maphalala N, Malinga LA, Mbelle NM, Maningi NE. A comparative evaluation of the new Genexpert MTB/RIF ultra and other rapid diagnostic assays for detecting tuberculosis in pulmonary and extra pulmonary specimens. Sci Rep. 2019;9:16587.
19. Bristol U of. QUADAS-2. Available from: *https://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/*. Accessed July 30, 2020.
20. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013;137:558-65.
21. Korevaar DA, Cohen JF, Reitsma JB, et al. Updating standards for reporting diagnostic accuracy: the development of STARD 2015. Res Integr Peer Rev. 2016;1:7.
22. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: The example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open. 2019;9:e025611.