Colombian Journal of Anesthesiology

OPEN

Received: 1 November, 2023 ▸ Accepted: 26 February, 2024▸ Online first: 2 Abril, 2024

DOI: https://doi.org/10.5554/22562087.e1108

Innovative perspectives on the value of diagnostic tests in clinical practice

Perspectivas innovadoras sobre el valor de las pruebas diagnósticas en la práctica clínica

Kelly Estrada-Orozco[a-c], Juliana Cuervo[a]

[a] Clinical Epidemiology doctoral program, Clinical epidemiology and Biostatistics Department, Pontificia Universidad Javeriana. Bogotá, Colombia.
[b] Evidence Synthesis and Technology Management Unit, Health Technology Evaluation Institute (IETS). Bogotá, Colombia.
[c] Clinical Research Institute, Facultad de Medicina, Fundación Universitaria Sanitas. Bogotá, Colombia.

Correspondence: Cochrane, Facultad de Medicina, Universidad Nacional de Colombia, Cra. 30 calle 45, Campus Universitario. Bogotá, Colombia.

How to cite this article: Estrada-Orozco K, Cuervo J. Innovative perspectives on the value of diagnostic tests in clinical practice. Colombian Journal of Anesthesiology. 2024;52:e1108.

Email: kpestradao@unal.edu.co

Abstract

Diagnostic tests have intrinsic characteristics such as sensitivity, specificity, overall accuracy and likelihood ratios which define their operational performance. It is not uncommon to find in the literature that test value and clinical utility are defined based exclusively on those characteristics. This paper introduces several arguments aimed at prompting a reflection regarding the characteristics that define the true value of diagnostic tests in clinical practice. It concludes with the view that the value of each diagnostic test needs to be established in accordance with the circumstances in which it is used, taking into account extrinsic characteristics such as in whom it is used, when, where and by who.

Key words

Diagnosis; Diagnostic test; Clinical value; Sensitivity and specificity; Clinical practice.

Resumen

Las pruebas diagnósticas tienen características intrínsecas, como la sensibilidad, especificidad, exactitud global y las razones de verosimilitud, que definen su desempeño operacional. No es infrecuente encontrar en la literatura que se valore la prueba y se defina su utilidad clínica exclusivamente de acuerdo con estas características. En este documento se presentan varios argumentos que permiten reflexionar sobre las características que verdaderamente definen el valor de las pruebas diagnósticas en la práctica clínica. Se concluye con una perspectiva en la que el valor de cada prueba diagnóstica se establece de acuerdo con las circunstancias de uso de la misma: de quién, cuándo, dónde y en quién se use la prueba, y todas estas son características extrínsecas de una prueba diagnóstica.

Palabras clave

Pruebas diagnósticas; Diagnóstico; Valor clínico; Sensibilidad y especificidad; Práctica clínica.

Lea la versión en español de este artículo en www.revcolanest.com.co

INTRODUCTION

A diagnostic test is any means capable of modifying the diagnostic probability of a condition. Specifically in clinical practice, diagnostic tests are approaches used to identify a patient's disease with high accuracy in order to provide early and adequate treatment (1).

Tests can be used for several purposes, including detection, risk assessment, diagnosis, prognostic characterization, staging, monitoring or surveillance (1). On the other hand, as part of the diagnostic process, a test can be introduced as: 1. Replacement (i.e., tests associated with a lower burden, invasiveness, cost, or superior accuracy); 2. Triage (i.e., tests that define continuation of a diagnostic process and, therefore, minimize the use of an invasive or costly test); 3. Addition (i.e., to improve accuracy within the existing diagnostic process); or 4. Parallel or combined tests (widely used in clinical practice, these are tests for the same or different health conditions which allow to rule out differential diagnosis within the syndromic approach) (2).

It is not uncommon to find in the literature that a diagnostic test is rated as excellent when it is accurate (the measured value is as close as possible to the actual value) and precise (the measured value is repeatable and reproducible) (1,3,4). Also, a diagnostic test may be considered to be "ideal," "the perfect test" or "suitable" when it correctly identifies the subjects with and without the disease condition with 100% accuracy (5,6). Although accuracy and precision are the minimum required characteristics to rate a diagnostic test as ideal, they are not enough to define the test's value and utility. Besides, a test's true value does not depend only on its intrinsic operational characteristics such as sensitivity, specificity, positive and negative predictive values or overall accuracy, but on how much the test can be used in a specific context and to what extent it helps the user in terms of the clinical decision and the ability to provide adequate and timely treatment that results in benefit for the patient, that is to say, how useful the test is. Moreover, there are also extrinsic peculiarities such as in whom the test is performed, when, where and by who (7,8).

This paper aims to present evidence-based arguments as to why the intrinsic operational characteristics which characterize the technical validity of the test, including its sensitivity, specificity and diagnostic accuracy, among others, are only the starting point to assess the value of a diagnostic test. In practice, extrinsic factors that characterize the clinical context where the test is applied determine its operational performance. Consequently, they need to be considered in order to guide decisions regarding its use and in order to define its true value or utility.

The discussion that follows covers: 1. The role played by the test's intrinsic characteristics in the diagnostic process; 2. The role played by the certainty of the test's intrinsic characteristics in the diagnostic process; 3. The variability of the test's operational performance as a function of the user and the setting in which it is used; and 4. Other factors influencing the use of the tests and which are involved in defining their value.

Role played by the intrinsic characteristics of diagnostic tests

The term intrinsic comes from the Latin intrinsĕcus and is used to qualify that which belongs to something (9). In the setting of diagnostic tests, intrinsic characteristics are those that define their diagnostic "performance," that is to say, their ability to correctly classify individuals with or without the condition of interest. These include, primarily, standard measurements such as sensitivity (Sen), specificity (Sp), positive and negative predictive values (PPV and NPV), overall accuracy (OA), positive and negative likelihood ratios (LR+ and LR-), diagnostic odds ratio (diagnostic OR), and Youden index (J). Other less well known measures have also been proposed as a summary of test "yield" (test performance in specific clinical scenarios), such as the number needed to diagnose (NND), the number needed to misdiagnose (NNM), and even an index to measure the clinical utility of a positive or negative result based on the corresponding predictive values and the sensitivity and specificity, respectively, with thresholds which define the degree to which a test is "useful" in clinical practice (Table 1) (1,10-12).

No matter how elaborate the measurements may appear, assessing a test's utility based only on its basic operational characteristics without taking

Table 1. Description of the use of the Clinical Utility index according to A. Mitchell.

Clinical utility index (CUI)	Utility interpretation
CUI > 0.81	Excellent
0.64 ≤ CUI < 0.81	Good
0.49 ≤ CUI < 0.64	Fair
0.36 ≤ CUI < 0.49	Little
CUI < 0.36	Very little

Source: Authors, from (11).

into account the context and how its results are actually interpreted and applied may be arbitrary and inadequate. For example, how much value do tests with higher sensitivity, specificity or accuracy add to the clinical decision? or To what extent are tests with an excellent "utility" index really useful?

Let us take the HIV self-test as an example. This test has a 100% sensitivity — 100% of the people with HIV infection test positive — and a specificity of 99.8% — 99.8% without HIV infection test negative. Moreover, this is a highly reliable test and a study which examined the feasibility of use by non-professionals showed that more than 99.2% of the participants obtained an interpretable result and more than 98.1% interpreted the result correctly. Positive results were interpreted correctly in 100% of cases (13).

Despite being a test that would have a CUI that classifies it as an excellent diagnostic test in a context of high prevalence of infection - and, therefore, a high PPV - it does not provide a definitive diagnosis and, according to the management guidelines, a confirmatory test is required in all positive cases (14). A false positive result would have implications in terms of initiation of anti-retroviral therapy, the impact on the mental health of the individual, and other social consequences, requiring the use of a second test in order to obtain a definitive diagnosis, thus giving the self-test a screening role.

The role of this test is not defined merely on the basis of its operational characteristics: it works, it is accurate and reliable, but insufficient as a single diagnostic tool, given that any judgement of its performance requires looking into the consequences of misdiagnosis, even if it is unlikely. On the other hand, the self-administered test offers benefits in terms of access to diagnosis and timely care because, should it be positive, it prompts the individual to seek medical care and benefit from treatment once a laboratory test confirms the result. When anti-retroviral treatment is initiated early on, the life expectancy of individuals with HIV can be similar to that of the general population.

In another example, the American Pregnancy Association (APA) recommends the use of home pregnancy tests, stating that their accuracy ranges between 97% and 99% when done correctly, and that they are a rapid, low-cost alternative that guarantees the user's privacy. Despite their high diagnostic accuracy, these tests are not sufficient when it comes to confirming or ruling out pregnancy, and the reason is simple: a false positive or a false negative result has huge effects. For example, a false negative result would delay timely enrollment in prenatal care programs, with its implications for maternal and fetal health. Therefore, although a home pregnancy test has an excellent utility index, high diagnostic accuracy, and sensitivity and specificity values greater than 95%, it would no qualify as a test for definitive diagnosis.

In 2014, Josephson et al. published a meta analysis describing the combined estimated sensitivity and specificity for CT angiography (CTA) as well as for MR angiography (MRA) in the detection of vascular malformations in patients with intracranial bleeding (15). In CTA studies, the combined estimate for sensitivity was 95% (95% confidence interval [CI]: 90 to 97%) and 99% for specificity (95% CI: 95 to 100%). In MRA studies, the combined estimate for sensitivity was 98% (95% CI: 80 to 100%) and 99% for specificity (95% CI: 97 to 100%). The answer to the question on which of the two tests to use in order to make a surgical decision for a patient with intracranial bleeding can be as simple as "use whichever is available or is less expensive, or is preferred by the clinician, because they are both highly accurate and have an excellent clinical utility index." However, other considerations might tilt the balance towards CTA over MRA, at least according to the data derived from this study. These include the consequences of the decision in terms of the frequency of false negative results when using MRA (Sen 95% CI: 80 to 100%). Even clinical characteristics and patient history, such as trauma or other comorbidities, may tilt the balance, indicating again that a set of conditions that are external to the test determine its use and clinical utility.

It has been believed that the more stable the intrinsic characteristics in relation to the prevalence of the condition of diagnostic interest (16), the better the test is for clinical decision making, hence the positioning of high sensitivity and specificity as desirable characteristics in a test. In truth, however, a sensitive or specific test selected in accordance with its objective, does not solve the issues faced by its users and, contrary to held belief, tests can offer different degrees of information depending on the prevalence of the condition among the population in which they are used (17-22).

The same is true for other intrinsic characteristics of diagnostic tests, as is the case with positive and negative likelihood ratios. For example, liver and biliary ultrasound is considered the gold standard for acute cholecystitis, partly due to the excellent operational characteristics of the test. In emergency care, the positive and negative likelihood ratios of the ultrasound finding of free fluid surrounding the gall bladder are 10.7 and 0.8, respectively (23); however, the post-test probability realized with its positive result in a patient with acute abdominal pain is only 20%, and remains unchanged (~ 2%) when its result is negative (2% retest likelihood, based on the 5-10% prevalence of cholelithiasis in the general population, and only 20% of patients with cholelithiasis develop cholecystitis) (24,25). Its true value is observed in settings with pretest probability greater than 10%, that is to say, in clinical populations selected on the basis of other diagnostic tests and the review of clinical signs. Therefore, it is flawed to think that liver and biliary ultrasound has an excellent clinical utility overall, because its utility depends on the situation in which it is applied, i.e., it is context-dependent.

Given the above, although the intrinsic characteristics of the tests are necessary,

they are not sufficient to determine their value. High sensitivity or specificity or accuracy alone do not determine the test's value for clinical decision-making, as there are other context or setting-related characteristics that define it.

The role of certainty regarding the intrinsic characteristics of the diagnostic test

Performance measurements of diagnostic tests are estimated with a certain degree of uncertainty. The determination of a test's intrinsic characteristics requires a comparator with unsurpassable operational characteristics in the context in which it is applied and for the condition of interest, such comparator being the gold standard. The gold standard can be defined as the best available method to determine the presence or absence of the condition of interest (26); its characteristics are not solely operational considering that its use is the result of a process of consensus, proof of additional benefit, and acceptance (2).

Although the importance of having a reference test with the characteristics of a gold standard is recognized, in daily practice, verifying true diagnoses, that is to say, confirming that the subjects actually have the condition of interest using the gold standard, may not be very feasible, either because of risk to the patient, the cost in terms of human and institutional resources, low practicality, or ethical conflicts derived from its use. In other situations, such as some psychiatric diseases — including anxiety, depression(27) or schizophrenia (28,29)— the gold standard is not even available. Although the lack of a perfect gold standard is frequent in research practice, there is no consensus regarding the best option to avoid introducing biases when comparing the new test against the gold standard and assessing its intrinsic characteristics (30).

The term reference standard or criterion is preferred in the absence of a gold standard. The difference is that these two are strategies or tests consistent with the best current and accepted approach for diagnosis and which allow comparison with the test of interest to be assessed, even if their performance is not perfect. In other cases, even if the gold standard is available, there are ethical or feasibility risks that limit its use — e.g. brain biopsy as the gold standard for the diagnosis of Alzheimer's disease — and therefore, another test with lower operational performance is preferred as the reference standard (31). Consequently, uncertainty is made evident to the extent to which the characteristics of the study test are determined against a reference standard which is considered the best available option but not necessarily the test with the best operational performance. This might mean that the new test may actually have better operational characteristics for diagnosis than the reference standard, even if it is still less good when compared to the gold standard. For example, biomarkers have been recently proposed for prostate cancer as more accurate substitutes for prostate specific antigen, even though biopsy is the gold standard (32).

Other methodological considerations of studies designed to determine the intrinsic characteristics of diagnostic tests can also affect the certainty of those measurements (33). The first step in assessing the value of a medical test before undertaking comparative impact studies is accuracy assessment (34). This assessment is done by means of cross-sectional studies nested in longitudinal designs such as cohort studies, clinical trials or case-control studies (34), the former having the advantage of a lower risk of artificial increase in accuracy as a result of biased prevalence values (30). However, design type is not the sole source of concern in relation to diagnostic accuracy studies; other recognized sources of uncertainty of the obtained results include the risk of selection bias, the application and interpretation of the study test (index test), and the reference pattern, among others (33,35,36).

Therefore, understanding the value of a test also requires understanding the degree of uncertainty surrounding measurements of its ability to discriminate and of its reliability, as well as the possibility of measurements being biased or under/overestimating actual accuracy.

Behavior of intrinsic characteristics depending on the test setting and user

Test reliability refers to the variation between test measurements of a unit of analysis, which is explained by measurement error (37) due either to repeatability or reproducibility. In measurement theory, repeatability refers to variation in measurements performed at different time points of the same unit of analysis in identical conditions which, should it occur, is attributable to errors in the measurement process. To determine whether repeatability exists, measurements must be made using the same tool or method, the same observer or reviewer, and in a time period during which no variation is expected to occur in the record of interest (37).

On the other hand, reproducibility refers to variation in measurements performed on a unit of analysis in conditions which are not identical, either because changes are expected to occur in the measured unit of analysis or because of the use of varying methods, tools or observers (37).

A diagnostic test can have excellent intrinsic characteristics, including good reproducibility and repeatability, but its true utility will depend on how it is used. For example, serologic tests detect antibodies or immunoglobulins produced as an immune response to infection in humans. When immunoglobulin M (IgM) antibodies are present, they may indicate active or recent infection, while immunoglobulin G (IgG) antibodies appear later in the infection process and often indicate past infection but do not rule out recent infection (38).

Serologic tests can play an important role in early infection detection. These

tests are easy to operate and provide fast antibody screening in 10-15 minutes. Moreover, due to their low cost and fast and easy processing, they are used as detection tools for the general population (39).

Antibody tests have been developed to detect not only IgG, but also IgM and total antibodies for the detection of SARS-CoV-2 infection; however, the operational characteristics of these tests vary significantly depending on the clinical stage in which they are applied, as well as the characteristics of the individual patients. Antibody tests carried out one week after the initial symptoms detect only 30% of people with COVID-19, with this figure increasing to 70% in the second week and to more than 90% in the third week (40). On the other hand, in asymptomatic patients, the combined sensitivity for IgM is 28.6% (95% CI: 23.8-33.7%). In symptomatic patients tested 8-11 days or less since the onset of symptoms, the combined sensitivity for IgM is 33% (95% CI: 23-43%), and in symptomatic patients after more than 11 days since the onset of symptoms, sensitivity for IgM is 66% (95% CI: 61-70%) (39).

As observed in the example, the test's sensitivity varies according to the characteristics of the subject (symptomatic or asymptomatic) and the time elapsed since exposure or onset of symptoms. Again, it is clear that the test's intrinsic characteristics cannot define its utility in absolute terms. For this particular case, its performance varies according to the time point along the course of the disease at which it is used, highlighting the need to know when to use a diagnostic test, recognize its role in the diagnostic process, and understand how it works and why it is used. This undoubtedly means that the user of the test needs to have a certain minimum experience.

Other implications of the use of diagnostic tests

Thinking about implications brings us back to the test's extrinsic characteristics. Although some progress has been made by way of considering the consequences derived from using false negative or false positive results, such as treating more or not treating the patient, the implications regarding the use of the test require reflections that go beyond what is derived from the intrinsic characteristics, to include considerations of the financial and human resources needed to apply the test. It also requires reflecting on the risk-benefit of the results from a social and ethical perspective.

Such is the relevance of these considerations that, in some settings, the test with the greatest value is not the most accurate but the one that is available to allow timely decision-making that can help change the clinical course of a patients when there are no other options available. Such test could even be as simple as a clear, well directed and semiologically rich clinical history.

In conditions of very limited resources or staff with insufficient training, very accurate tests which are difficult to implement or interpret can be of little use or value, while tests with good but lower accuracy which are low-cost, fast, easy to implement and interpret with minimum training can be of great usefulness and value for a population.

On the other hand, it might not be ethical to diagnose patients with conditions for which it is not possible to carry out an intervention to cure or modify the clinical course. Conducting a test in such a situation can potentially infringe any of the four ethical principles, namely, beneficence, non-maleficence, justice and autonomy. Genetic tests in Alzheimer's disease (AD) are examples of diagnostic tests whose excellent intrinsic characteristics (100% sensitivity and 98.9% specificity) (41) can be at odds with their utility and value when factoring in extrinsic factors.

Late onset Alzheimer's disease is the most common form of this condition and is generally sporadic. However, some alleles that increase the risk of AD have been identified. APOE ε4 is a well established risk factor for AD and is associated with a four-fold increase in the risk of developing the disease (42,43). Although genetic tests can readily identify the presence or absence of these susceptibility genes, this is of little clinical or diagnostic benefit because of the lack of a risk modifying treatment. Moreover, the diagnostic uncertainty remains given that a patient may be a carrier of the APOE ε4 allele and not develop AD, or develop the disease in the absence of the APOE ε4 allele (42). Consequently, what clinical utility could the test have if no early or adequate treatment can be offered? Furthermore, knowledge of the carrier status could impose a huge emotional burden given the uncertainty and the current inability to provide effective interventions.

Another example in which the utility of the test is defined by its extrinsic characteristics, despite excellent intrinsic characteristics, is the COVID-19 diagnosis. The gold standard for diagnosis is RT-PCR (reverse transcription polymerase chain reaction) with a sensitivity of 85.7% (95% CI: 81.5-89.1%) in hospitalized patients, 95.5% (95% CI: 92.2-97.5%) in outpatients, and 89.9% (95% CI: 88.2-92.1%) in all patients (44). However, test availability in some regions is low and turnaround time is long; moreover, flaws at the time of taking the sample or problems with transport and processing, as well as cost, mean that it is not a test with the highest clinical utility. In contrast, rapid antigen tests (Ag-RDT) with a sensitivity of 84 to 97% and specificity of 97 to 100% compared to RT-PCR (45), are done very quickly and are easier to use and interpret. The turnaround time for Ag-RDT tests is less than 30 minutes, contributing to diagnosis, tracking and study of contacts, thus slowing SARS-CoV-2 transmission in a community (46).

CONCLUSIONS

Based on the arguments presented in this document, it is possible to conclude that, in both clinical practice as well as in public health, the utility and value of a test are

not defined exclusively by its intrinsic characteristics. The value of each diagnostic test is determined in accordance with the circumstances in which it is used: who, when, where and in whom, all of which are extrinsic characteristics. Therefore, a reflective and systematic exercise is needed in order to make decisions about the use or introduction of a test based not only on its intrinsic characteristics and certainty of its performance, but also and in particular, based on the circumstances that prompt its use and the context in which it is used. This includes retest likelihood, the consequences of missing a diagnosis or overdiagnosing, the risks associated with the use of the test, the feasibility of its correct application, its acceptability and interpretability, availability, costs, and other resources, and the ethical consequences of its use. In conclusion, it is the view of the authors of this article that there is no ideal or better diagnostic test for a given condition but only tests that add value to the clinical decision depending on each setting and context in which they are used.

Conflicts of interest

KEO is a member of the GRADE group and the GRADE Diagnosis Group. The authors have no other conflicts of interest to disclose.

Founding

None declared by the authors.

REFERENCES

1. Bolboacă SD. Medical Diagnostic Tests: A review of test anatomy, phases, and statistical treatment of data. Comput Math Methods Med. 2019;1891569. doi: https://doi.org/10.1155/2019/1891569

2. Schünemann HJ, Mustafá RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-41. doi: https://doi.org/10.1016/j.jclinepi.2019.12.020

3. Šimundić AM. Measures of diagnostic accuracy: Basic definitions. EJIFCC. 2009;19(4):203-11.

4. Shreffler JHM. Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 [cited 16 Jan 2024]. Available at: https://www.ncbi.nlm.nih.gov/books/NBK557491/

5. Wong HB. Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. Proceed Singapore Healthc. 2001;20(4):316-8. doi: https://doi.org/10.1177/201010581102000411

6. Pluddemann A BA, O'Sullivan J. Spectrum bias: Sackett Catalogue Of Bias [internet]. 2019 [cited 16 Jan 2024]. Available at: https://catalogofbias.org/biases/spectrum-bias/

7. Buehler AM, Ascef BdO, Oliveira HAd, Ferri CP, Fernandes JG. Rational use of diagnostic tests for clinical decision making. Revista da Associação Médica Brasileira. 2019;65. doi: https://doi.org/10.1590/1806-9282.65.3.452

8. Schünemann HJ, Mustafá RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82. doi: https://doi.org/10.1016/j.jclinepi.2019.02.003

9. Definición de intrínseco [internet]. 2024 [cited 16 Jan 2024]. Available at: https://definicion.de/intrinseco/

10. Habibzadeh F, Yadollahie M. Number needed to misdiagnose: a measure of diagnostic test effectiveness. Epidemiology. 2013;24(1):170. doi: https://doi.org/10.1097/EDE.0b013e31827825f2

11. Mitchell AJ. The clinical significance of subjective memory complaints in the diagnosis of mild cognitive impairment and dementia: a meta-analysis. Int J Geriatr Psychiatry. 2008;23(11):1191-202. doi: https://doi.org/10.1002/gps.2053

12. Mitchell AJ. Sensitivity × PPV is a recognized test called the clinical utility index (CUI+). Eur J Epidemiol. 2011;26(3):251-2. doi: https://doi.org/10.1007/s10654-011-9561-x

13. Santé. Autotest VIH France [internet]. 2024 [cited 16 Jan 2024]. Available at: https://www.autotest-sante.com/en/autotest-VIH-par-AAZ-139.html

14. Ministerio de Salud y de Protección Social. Guía de práctica clínica basada en la evidencia para la atención de la infección por VIH/SIDA en personas adultas, gestantes y adolescentes. Colombia: Minsalud; 2022.

15. Josephson CB, White PM, Krishan A, Al-Shahi Salman R. Computed tomography angiography or magnetic resonance angiography for detection of intracranial vascular malformations in patients with intracerebral haemorrhage. Cochrane Database Syst Rev. 2014;2014(9):Cd009372. doi: https://doi.org/10.1002/14651858.CD009372.pub2

16. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract. 2006;12(2):132-9. doi: https://doi.org/10.1111/j.1365-2753.2005.00598.x

17. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ. 2013;185(11):E537-44. doi: https://doi.org/10.1503/cmaj.121286

18. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137(7):598-602. doi: https://doi.org/10.7326/0003-4819-137-7-200210010-00011

19. Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests". J Epidemiol Community Health. 2002;56(5):330-2. doi: https://doi.org/10.1136/jech.56.5.330

20. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16(9):981-91. doi: https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<981::aid-sim510>3.0.co;2-n.

21. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62(1):5-12. doi: https://doi.org/10.1016/j.jclinepi.2008.04.007

22. Hultcrantz M, Mustafá RA, Leeflang MMG, Lavergne V, Estrada-Orozco K, Ansari MT, et al. Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper. J Clin Epidemiol. 2020;117:138-48. doi: https://doi.org/10.1016/j.jclinepi.2019.05.002

23. Jang TB, Ruggeri W, Kaji AH. The predictive value of specific emergency sonographic signs for cholecystitis. J Med Ultras. 2013;21(1):29-31. doi: https://doi.org/10.1016/j.jmu.2013.01.006

24. Zarate AJ. ÁM, King, I, Torrealba. A. Colecistitis aguda. Universidad Finis Terrae: Escuela de Medicina 2016;7.

25. Halpin V. Acute cholecystitis. BMJ Clin Evid. 2014;2014.

26. Gelaye B, Tadesse MG, Williams MA, Fann JR, Vander Stoep A, Andrew Zhou X-H. Assessing validity of a depression screening instrument in the absence of a gold standard. Ann Epidemiol. 2014;24(7):527-31. doi: https://doi.org/10.1016/j.annepidem.2014.04.009

27. Davison TE, McCabe MP, Mellor D. An examination of the "gold standard" diagnosis of major depression in aged-care settings. Am J Geriatr Psychiatry. 2009;17(5):359-67. doi: https://doi.org/10.1097/JGP.0b013e318190b901

28. Wood SJ, Yung AR. Diagnostic markers for schizophrenia: do we actually know what we're looking for? World Psychiatry. 2011;10(1):33-4. doi: https://doi.org/10.1002/j.2051-5545.2011.tb00006.x

29. van Os J, Tamminga C. Deconstructing psychosis. Schizophr Bull. 2007;33(4):861-2. doi: https://doi.org/10.1093/schbul/sbm066

30. Estrada-Orozco K. Diseño de una prueba para diagnóstico de trastorno cognitivo y validación en una cohorte de sujetos mayores de 50 años en Colombia en el 2016-2017. Bogotá, D.C: Universidad Nacional de Colombia; 2018.

31. Pietrzak K, Czarnecka K, Mikiciuk-Olasik E, Szymanski P. New perspectives of alzheimer disease diagnosis - the most popular and future methods. Med Chem. 2018;14(1):34-43. doi: https://doi.org/10.2174/1573406413666171002120847

32. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ (Clinical research ed). 2006;332(7549):1089-92. doi: https://doi.org/10.1136/bmj.332.7549.1089

33. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174(4):469-76. doi: https://doi.org/10.1503/cmaj.050090

34. Mathes T, Pieper D. An algorithm for the classification of study designs to assess diagnostic, prognostic and predictive test accuracy in systematic reviews. Systematic Reviews. 2019;8(1):226. doi: https://doi.org/10.1186/s13643-019-1131-4.

35. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013;137(4):558-65. doi: https://doi.org/10.5858/arpa.2012-0198-RA

36. Whiting P, Rutjes A, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36. doi: https://doi.org/10.7326/0003-4819-155-8-201110180-00009

37. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultr Obstet Gynecol. 2008;31(4):466-75. doi: https://doi.org/10.1002/uog.5256

38. Mahajan A, Manchikanti L. Value and validity of coronavirus antibody testing. Pain Physician. 2020;23(4s):S381-s90. doi: https://doi.org/10.36076/ppj.2020/23/S381

39. Mercado M, Malagón-Rojas J, Delgado G, Rubio VV, Muñoz Galindo L, Parra Barrera EL, et al. Evaluation of nine serological rapid tests for the detection of SARS-CoV-2. Rev Panam Salud Pública. 2020;44:e149. doi: https://doi.org/10.26633/RPSP.2020.149

40. Deeks JJ, Dinnes J, Takwoingi Y, Davenport C, Spijker R, Taylor-Phillips S, et al. Antibody tests for identification of current and past infection with SARS-CoV-2. Cochrane Database of Systematic Reviews. 2020(6). doi: https://doi.org/10.1002/14651858.CD013652.pub2

41. Veiga S, Rodríguez-Martín A, García-Ribas G, Arribas I, Menacho-Román M, Calero M. Validation of a novel and accurate ApoE4 assay for automated chemistry analyzers. Scientific Reports. 2020;10(1):2138. doi: https://doi.org/10.1038/s41598-020-58841-7

42. Atkins ER, Panegyres PK. The clinical utility of gene testing for Alzheimer's disease. Neurol Int. 2011;3(1):e1-e. doi: https://doi.org/10.4081/ni.2011.e1

43. Bertram L, Tanzi RE. Genome-wide association studies in Alzheimer's disease. Hum Mol Genet. 2009;18(R2):R137-45. doi: https://doi.org/10.1093/hmg/ddp406

44. Kortela E, Kirjavainen V, Ahava MJ, Jokiranta ST, But A, Lindahl A, et al. Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients. PloS One. 2021;16(5):e0251661-e. doi: https://doi.org/10.1371/journal.pone.0251661

45. Peeling RW, Olliaro PL, Boeras DI, Fongwen N. Scaling up COVID-19 rapid antigen tests: promises and challenges. The Lancet Infectious Diseases. 2021;21(9):E290-5. doi: https://doi.org/10.1016/S1473-3099(21)00048-7

46. World Health Organization. WHO provides one million antigen-detecting rapid diagnostic test kits to accelerate COVID-19 testing in Indonesia. World Health Organization [internet]. 2021 [cited 16 Jan 2024]. Available at: https://www.who.int/indonesia/news/detail/17-03-2021-who-provides-one-million-antigen-detecting-rapid-diagnostic-test-kits-to-accelerate-covid-19-testing-in-indonesia.