Received: 25 July 2024 | Accepted: 4 September 2024

DOI: 10.1111/jebm.12641

LETTER

WILEY

Evaluating clinical utility in diagnostic tests: Likelihood ratios confidence intervals and proposal of a simple index

Diagnostic test accuracy (DTA) studies are pivotal to evaluate the performance of diagnostic tests in medical research. Sensitivity and specificity values calculated by these studies derive further positive and negative likelihood ratio (LR+ and LR−). The fundamental role of LR is to inform the shift in probability of a disease before and after a test is conducted. If LR+ and LR− have a value of 1.0 or very close to that, this might indicate lack of clinical utility of the test.[1] However, the likelihood ratio is not always mentioned in DTA studies and even when mentioned, its confidence interval is not presented. Consequently, many clinically irrelevant tests may be mistakenly considered to be useful in DTA studies that are published without proper recognition of their limitations. To address these challenges, we propose a dual-method approach to evaluating the utility of diagnostic tests based on the thorough analysis of likelihood ratios:

Direct Method: When authors provide 95% confidence intervals (CIs) for LR+, and LR−, interpreting these is crucial. Various methods for this interpretation have been proposed.[2] If CIs include or cross the value of 1.0, the test enters the 'diagnostic null zone', suggesting minimal clinical impact.

Indirect Method: In cases where CIs are not provided, we introduce a simple, practical index calculated from the lower bounds of the sensitivity and specificity CIs. If the sum of these values is 1.0 or less, it similarly suggests that the diagnostic test is within the diagnostic null zone, providing a quick assessment tool when detailed statistical data are lacking.

To illustrate the importance of those methods, we detail an extreme yet potentially common scenario where the sum of sensitivity and specificity approaches or equals 1.0. Researchers designed a study to assess urine cultures for detecting acute coronary syndrome (ACS) in a coronary care unit (CCU) for "cases" and a general Intensive Care Unit (ICU) for "controls." Since urine cultures lack intrinsic capability to confirm or rule out ACS, the prevalence of positivity likely mirrors each other across both groups, influenced by external factors such as the prevalence of urinary infections in hospitalized patients. The resulting contingency table shows that the sensitivity and specificity are 5% and 95%, respectively. It is critical to understand that in this hypothetical scenario, urine cultures returned positive in 5% of hospitalized ICU patients—regardless of whether they were in the coronary or general sections—and this prevalence was mirrored among both cases and controls. Consequently, the test does not effectively confirm or exclude the condition and thus appears by chance, acting as a "coincidental factor." This factor is the actual basis for the results in the contingency table. Succinctly, the decision to use this index test was poorly conceived. Without critical analysis, both researchers and readers might remain oblivious to this fact, despite the apparent high specificity of 5% sensitivity and 95% specificity.[3] This test could misleadingly be deemed highly specific.

In this described case, and indeed in any situation where the prevalence of an inadvertently coincidental factor is equal between cases and controls, sensitivity and specificity values are produced as complementary values (complementary values here refer to the sum equaling 1.0) because the positivity index will be the same in cases and controls groups. In addition, the LR+ and the LR− will also be equal to 1.0. To understand why this phenomenon occurs, we revisit the formulas for LR+ and LR−. The positive likelihood ratio (LR+) is calculated as follows[4]:

$$LR+ = \frac{Sensitivity}{(1 - Specificity)}.$$

And the negative likelihood ratio (LR−) is:

$$LR- = \frac{(1 - Sensitivity)}{Specificity}.$$

Complementary values imply that when one metric (e.g., sensitivity) is subtracted from 1, it yields the exact other metric (e.g., specificity). If we define x as the sensitivity and recognize that sensitivity and specificity sum to 1 (or 100%), it follows that $y = 1-x$ for specificity. Inserting these variables into the formulas for LR+ and LR−, we get:

$$LR+ = \frac{x}{(1 - y)}.$$

Given that $y = 1-x$, we substitute for y in the denominator to find:

$$LR+ = \frac{x}{1 - (1 - x)} = \frac{x}{x} = 1.0.$$

Similarly, for LR−, the formula becomes:

$$LR- = \frac{1 - x}{y}.$$

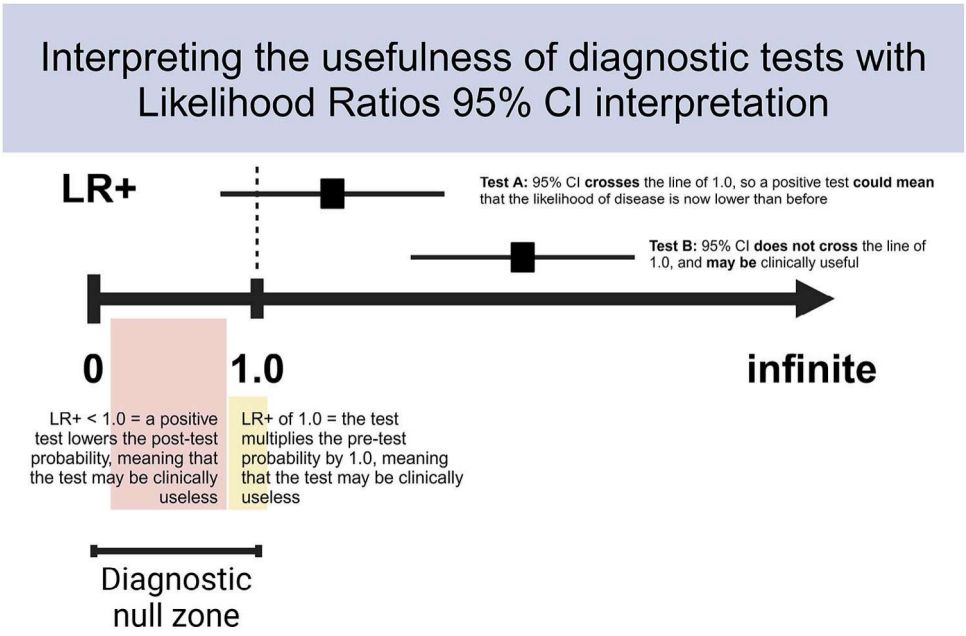

FIGURE 1 Visual representation of the interpretation of 95% confidence intervals (CI) for likelihood ratios (LR+ and LR−) in assessing the clinical relevance of diagnostic tests.

Since $y = 1 - x$, substituting for y gives:

$$LR- = \frac{1-x}{1-x} = \frac{y}{y} = 1.0.$$

This mathematical fact shows that, in situations where sensitivity and specificity are complementary, the test in question may not effectively alter the probability of the presence or absence of the disease. Thus, the high specificity observed in this scenario is misleading because it is not due to the diagnostic significance of urine culture for ACS, but due to the prevalence of unrelated conditions within the studied populations.

To add another layer to the issue at hand, we suggest that readers consider the natural variation in the prevalence of "coincidental factors" between groups in case-control or cross-sectional studies. This inherent variability can have a significant impact on the interpretation of diagnostic test results. For example, using Fisher's exact test, the probability that an event expected to occur in 5% of the "case" population occurs in 8% of the "case" population is approximately 56.8%.[5–7] In this scenario, a higher prevalence of positive urine cultures in the "case" group compared to the expected rate could reasonably occur by chance. In this case, the LR+ would be 1.6 and the LR− would be 0.97; neither LR+ nor LR− would be exactly equal to 1.0. This discrepancy illustrates how a clinically irrelevant test could be regarded as useful simply by chance.

Therefore, we suggest that DTA studies calculate or derive confidence intervals (CI) for LR+ and LR−. This interpretation is similar to the interpretation of odds ratios[8]: when CI crosses the line of 1.0, it is not possible to confirm that a positive test genuinely increases the probability of the disease and that a negative test genuinely decreases the probability of the disease. Therefore, the test should be considered potentially "clinically irrelevant" or "within the diagnostic null zone" (Figure 1).

When CI for LR are not provided in DTA studies, a reader may apply a straightforward index calculated from the lower bounds of the confidence intervals for sensitivity and specificity. This index, when summing to 1.0 or less, suggests the test falls within the "diagnostic null zone," indicating minimal clinical utility. This new index is designed to be a practical tool for clinicians and researchers to quickly assess the potential utility of diagnostic tests, both in binary and continuous outcomes.

To illustrate the practical application of this index, we examine its use in both binary outcome scenarios and in screening potential cutoff points for continuous outcomes within DTA studies. A study investigated the accuracy of detecting a new or presumably new left bundle branch block (index test) for myocardial infarction (reference test). A sensitivity of 42% (95% CI: 24%–62%) and specificity of 65% (95% CI: 57%–72%) were reported.[9] The sum of its sensitivity and specificity is 107% or 1.07. The likelihood ratios derived from these figures are an LR+ of 1.2 and an LR− of 0.89. These likelihood ratios indicate that the test has a weak ability to modify the probability of disease, but is still potentially useful. However, by applying our proposed index to these data, the summed lower bounds of sensitivity and specificity (24% + 57%) total 81%, placing this test within the diagnostic null zone. This example demonstrates how this index can effectively identify tests that may not significantly alter clinical decisions. This means that, within the confidence interval for the results of this test, a positive test could reduce the likelihood of developing the disease, whereas a negative test could increase it. This paradoxical result should be sufficient to reject this test or strategy in clinical practice.

We emphasize and recognize that when selecting optimal cutoff points for ROC curves, researchers already engage in a similar evaluation. The Youden index,[10] presented by the formula $J = Sensitivity + Specificity - 1$, demonstrates that when sensitivity and specificity are complementary, this point will be on the diagonal of the curve.

Limitations should be acknowledged: first, the generalizability of the method used to calculate likelihood ratios from the lower bounds of the 95% confidence interval of sensitivity and specificity may not apply to all diagnostic tests. In particular, this approach may fail in cases where diagnostic tests show a highly skewed confidence interval.[11] Additionally, our strategy did not take into account the variability in test performance that could occur across different populations or settings.[12] Lastly, our proposed emphasis on the statistical boundary of 1.0 for likelihood ratios had the potential for misinterpreting. Clinicians and researchers might mistakenly consider this threshold to be an absolute criterion rather than a flexible guideline, which could lead to premature dismissal of diagnostic tests that may have clinical utility in specific situations.

This strategy may be a significant advance in the field of medical research, specifically in the study of DTA. By proposing a new index that simplifies the assessment of diagnostic tests' clinical utility, our approach can prevent the use of tests that lack real clinical value. We advocate for the STARD and PRISMA-DTA guidelines to require authors of DTA research to include this index in future studies.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

José Nunes de Alencar[1]
Gabriel Gonçalves da Costa[2]
Vitor Borin Pardo de Souza[3]
Felipe Nogueira Barbara[4]
Yung Gonzaga[5]
Arn Migowski[6,7]

[1]*Department of Cardiology, Instituto Dante Pazzanese de Cardiologia, São Paulo, Brazil*
[2]*Department of Medical Biochemistry, Institute of Medical Biochemistry Leopoldo de Meis, Rio de Janeiro, Brazil*
[3]*Department of Internal Medicine, Botucatu Medical School, São Paulo State University (UNESP), Botucatu, Brazil*
[4]*Postgraduate Program in Medical Sciences, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil*
[5]*Department of Hematology, Instituto Nacional de Cancer, Rio de Janeiro, Brazil*
[6]*Clinical Research and Technological Development Divison, Research and Innovation Coordination, National Cancer Institute (INCA), Ministry of Health, Rio de Janeiro, Brazil*
[7]*Professional Master's Program in Health Technology Assessment, Teaching and Research Coordination, Instituto Nacional de Cardiologia (INC), Rio de Janeiro, Brazil*

Correspondence
José Nunes de Alencar, Department of Cardiology, Instituto Dante Pazzanese de Cardiologia, Avenida Dr. Dante Pazzanese, São Paulo, Brazil.
Email: Jose.alencar@dantepazzanese.org.br

ORCID

José Nunes de Alencar https://orcid.org/0000-0002-3835-6067

REFERENCES

1. De Alencar Neto JN, Santos-Neto L. The post hoc pitfall: rethinking sensitivity and specificity in clinical practice. *J Gen Intern Med*. 2024;39(8), 1506–1510.
2. Montero-Alonso MA, Roldán-Nofuentes JA. Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification. *J Biopharm Stat*. 2019;29(1), 56–81.
3. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. *J Clin Epidemiol*. 2010;63(8), 883–891.
4. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329(7458), 168–169.
5. Sprent P. Fisher exact test. In: Lovric M, ed. *International Encyclopedia of Statistical Science*. Springer; 2011:524–525.
6. Fisher RA. Statistical methods for research workers. In: Kotz S, Johnson NL, eds. *Breakthroughs in Statistics: Methodology and Distribution*. Springer; 1992:66–70. Springer Series in Statistics.
7. Fisher SRA. The Design of Experiments. Oliver and Boyd; 1951: 272.
8. Bland JM. Statistics notes. The odds ratio. *BMJ*. 2000;320(7247), 1468.
9. Kontos MC, McQueen RH, Jesse RL, Tatum JL, Ornato JP. Can myocardial infarction be rapidly identified in emergency department patients who have left bundle-branch block? *An Emerg Med*. 2001;37(5), 431–438.
10. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1), 32–35.
11. Bland JM, Altman DG. Statistics notes: transformations, means, and confidence intervals. *BMJ*. 1996;312(7038), 1079.
12. Hall MK, Kea B, Wang R. Recognising bias in studies of diagnostic tests part 1: patient selection. *Emerg Med J*. 2019;36(7), 431–434.