THE JOURNAL OF
MEDICINE, LAW & PUBLIC HEALTH

Vol 6, No 4. 2026

Commentary

The Impossibility of 100% Specificity in Rapid Diagnostic Test Reporting

José A. Martínez[1*]

[1] *Department of Business Management. Technical University of Cartagena - Member of European University of Technology. Cartagena, Spain*

Summary

Background: In commercial materials for rapid diagnostic tests, it is still common to find claims of "100% specificity". From a statistical and epistemological point of view, this kind of statement is deeply problematic.

Aims: This paper examines how justified those claims really are and what they mean in concrete clinical and public health settings.

Methods: We analyse the statistical foundations of specificity estimation, apply the Mayo-Spanos severity framework, and review three commercial rapid antigen test listings to illustrate reporting practices.

Results: Claims of 100% specificity are either logically falsified by observed false positives or unsupported by finite validation samples. The rule of three shows that zero false positives in n negatives only rule out a false-positive rate above 3/n. Severity analysis reveals that even with 210 negative samples, the claim that specificity >99.5% is supported with only 65% severity. In low-prevalence settings, assuming 100% specificity inflates positive predictive value (PPV) from 66% to 100%.

Conclusion: Reporting 100% specificity is scientifically untenable. Regulatory and editorial standards could realistically require confidence intervals and, at the very least, discourage unqualified point estimates.

Keywords: Baye's Theorem, Confidence Intervals, Point-of-Care Testing, Predictive Value of Tests, Sensitivity and Specificity, Statistics as Topic.

Received: 2026-03-23 | Accepted: 2026-08-13 | Published: 2026-08-28

DOI: 10.52609/jmlph.v6i4.318

*Corresponding author: josean.martinez@upct.es

INTRODUCTION

How a diagnostic test performs in practice is not a technical detail; it is central to both clinical care and public health decision-making. While STARD guidelines establish standards for reporting diagnostic accuracy in academic literature [1], and the European Centre for Disease Prevention and Control (ECDC) provides evaluation criteria [2], both emphasize that sensitivity (*Se*) and specificity (*Sp*) must be reported alongside 95% confidence intervals and sample sizes. The reason is straightforward: any estimate based on a finite validation sample is necessarily noisy, and confidence intervals make that uncertainty explicit.

In the marketing of rapid antigen tests (RATs), a small but telling distortion has become routine, especially since COVID-19: the repeated promise of "100% specificity". In practice, this promise tends to appear in two rather different ways. In some cases, false positives were observed in the validation study, but the Wald confidence interval has been pushed up to 100%, thereby restoring the very certainty that the data has contradicted. In others, a limited validation finds zero false positives and the label simply reads "specificity: 100%", with no further comment. The proportion in that sample may indeed be 1.00, but as a statement about the underlying population, it is, at best, an overclaim.

The ECDC's November 2020 technical report found specificities ranging from 80.2% (95% CI: 71.1–86.7) and 100% (95% CI: 98.8–100) [2]. A systematic review and meta-analysis of 135 studies encompassing 166,943 samples reported a pooled specificity of 1.00 (95% CI: 1.00–1.00), a statistical impossibility that reflects inadequate methodology rather than a genuine biological property [3]. These observations motivate the present commentary. Specifically, this study aims to evaluate the theoretical and empirical validity of claiming 100% specificity in rapid diagnostic test reporting, quantify the evidential warrant of such claims using the severity framework of Mayo and Spanos, and analyse the clinical consequences of unqualified specificity claims on Bayesian predictive values.

METHODS

To evaluate the statistical and epistemological soundness of 100% specificity claims in rapid antigen testing, we implemented a three-part analytical framework combining mathematical modelling of binomial bounds, severity testing, and observational evaluation of commercial listings.

Epistemological foundations

From a Popperian perspective, the assertion that a population parameter equals 1 is not a modest claim. It is a statement of deterministic certainty: under that model, the test simply cannot produce a false positive. Once a single false positive appears in any finite study, that statement is no longer "almost true"; it is just false. The familiar "black swan" example captures the point: one black swan is enough to overturn "all swans are white"; by the same logic, one false positive is enough to invalidate "this test never misclassifies a healthy individual."

Let $X \sim B(n, p)$ represent the number of true negatives when testing n healthy individuals, where p is the true population specificity. If $x < n$ (at least one false positive observed), the observed proportion $\hat{p} = \frac{x}{n} < 1$. Any statistical method that generates a confidence interval (p_L, p_U) with upper bound $p_U = 1$ implicitly assigns non-zero probability to $p = 1$, a value that has been logically falsified by the data. This is not a matter of degree or convention; it is a category error.

Statistical foundations

The persistence of the 100% upper bound in medical literature stems largely from the inappropriate use of the standard Wald confidence interval (CI). The Wald interval is:

DOI: 10.52609/jmlph.v6i4.318 | E-mail: josean.martinez@upct.es

$$\mathrm{CI}_{\mathrm{Wald}} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Near the boundaries of the parameter space ($\hat{p} \approx 1$), the symmetric normal approximation causes the upper bound to exceed 1.0. For example, in a study of $n = 500$ negative samples with 1 false positive ($\hat{p} = 0.998$), the standard error is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.00199$ and the Wald 95% CI is (0.9941, 1.0019). Because probabilities cannot exceed 1, software routinely truncates this to (99.41%, 100%). This truncation is mathematically invalid: it reintroduces $p = 1$ into the plausible parameter space despite its empirical falsification.

Appropriate inference requires methods that naturally respect the bounds of the binomial parameter. The Clopper-Pearson exact method solves: $\sum_{k=0}^{x}\binom{n}{k}{p_U}^k(1-p_U)^{n-k} = \frac{\alpha}{2}$ for the upper bound p_U. When $x < n$, this equation can only be satisfied by $p_U < 1$: the exact method inherently prohibits a 100% upper bound in the presence of any empirical failure. Similarly, the Wilson score interval, strongly recommended for extreme proportions [4,5], yields for the same example a 95% CI of (98.87%, 99.96%).

To resolve common ambiguities in diagnostic reporting, it is essential to distinguish among four distinct statistical and conceptual entities: (1) the observed sample proportion ($\hat{p} = x/n = 1$), a descriptive sample statistic; (2) the true population specificity parameter (p), an unknown biological property governing the test; (3) interval estimation (e.g., 95% CI), which models sampling uncertainty and, when calculated via bounded methods like Clopper–Pearson or Wilson, explicitly prohibits $p = 1$ when failures occur or bounds uncertainty when $x = n$; and (4) commercial claims of "100% specificity", which erroneously conflate sample observations ($\hat{p} = 1$) with deterministic population parameters ($p = 1$). Recognizing these distinctions confirms that observing zero false positives in finite samples provides no empirical support for absolute population perfection.

The foregoing argument applies when false positives have been observed. A distinct, and arguably more consequential, argument applies to the case where zero false positives are found: a numerically perfect validation result. This is the scenario most encountered in the rapid antigen test literature, and it is here that the claim of "100% specificity" is most aggressively deployed.

When zero false positives are observed in n negative samples, the maximum likelihood estimate of the false-positive rate is zero, and consequently, the point estimate of specificity is 100%. But this tells us about the sample, not the population. Hanley and Lippman-Hand [6] formalized the statistical problem succinctly: if nothing goes wrong, is everything all right? Their answer, embedded in what has become known as the "rule of three", is unambiguous: observing zero events in n trials imply an approximate 95% CI upper bound for the true event rate of $3/n$. Applied to specificity, if a manufacturer tests 100 negative samples and observes zero false positives, the 95% CI for the true specificity is not (100%, 100%) but approximately (97%, 100%). The study cannot, by construction, distinguish between a test with a true false-positive rate of 0.01% and one with a true false-positive rate of 2.9%.

The appropriate response to zero observed failures is not to report "specificity: 100%" but rather to report "specificity: 100% (95% CI: 97%–100%)" and to acknowledge explicitly that larger studies might identify failures. While academic reporting standards such as STARD 2015 [1] require this level of statistical transparency in scientific publications, commercial product labelling currently operates outside these mandates, frequently omitting confidence intervals entirely.

Mayo and Spanos severity framework

The concept of severity [7–9] provides a principled means of assessing how strongly the

data probe a given claim. The severity with which data $\mathbf{x_0}$ pass a hypothesis H is, informally, the probability that the test T would have yielded a result less consistent with H than $\mathbf{x_0}$ does, were H false, that is, were some discrepant alternative true. High severity means the test would almost certainly have uncovered a departure if one existed; low severity means the test was insensitive to that departure.

Let p_{fp} denote the true false-positive rate $(1 - Sp)$. The severity SEV of the claim $Sp > Sp_0$ is

$$\text{SEV}(Sp > Sp_0) = 1 - \mathbb{P}\left(X = 0 \middle| n, p_{fp} = 1 - Sp_0\right) = 1 - \left[\binom{n}{0} p^0 (1-p)^n\right] = 1 - (Sp_0)^n$$

where $X \sim B(n, 1 - Sp_0)$ is the number of false positives under the worst-case scenario that the true specificity equals exactly Sp_0.

Selection of cases

Three commercially available rapid antigen tests marketed online in 2026 were identified through a non-systematic search of digital pharmacy platforms. Three cases were selected to represent a qualitative gradient in reporting transparency: (1) claim of 100% specificity with no supporting data; (2) contingency table provided but no confidence intervals; (3) contingency table and confidence intervals reported. The selected cases were, respectively, Fluorecare 4-in-1 nasal rapid antigen test (COVID-19 + Influenza A + Influenza B + RSV)[1], SARS-CoV-2 rapid antigen test sold by Sekureco[2], and Green Spring SARS-CoV-2 rapid antigen test[3].

Clinical consequence analysis

There are both clinical and epidemiological consequences to accepting a 100% specificity claim. According to Bayes' theorem, the positive predictive value (PPV), or the probability that a positive test result reflects true infection, is where

$$\text{PPV} = \mathbb{P}(D|positive) = \frac{\mathbb{P}(positive|D)\mathbb{P}(D)}{\mathbb{P}(positive)} = \frac{Se \cdot \mathbb{P}(D)}{Se \cdot \mathbb{P}(D) + [(1 - Sp)(1 - \mathbb{P}(D))]}$$

$\mathbb{P}(positive)$ is the probability of being positive in a test, and $\mathbb{P}(D)$ is the prevalence of the disease D. When $Sp = 1$ the denominator reduces to $Se \cdot \mathbb{P}(D)$, making PPV = 1 regardless of prevalence. This mathematical collapse means that a test claiming 100% specificity appears to guarantee that every positive result is a true positive—a claim with profound implications, both for individual clinical decisions and for population-level screening strategies.

RESULTS

The product page for the Fluorecare 4-in-1 nasal rapid antigen test (COVID-19 + Influenza A + Influenza B + RSV), declares, in a bullet list: "Specificity: COVID-19 100%. Influenza A 100%. Influenza B 100%. RSV 100%." No confidence interval, no validation sample size, and no contingency table accompany these claims. The only performance data provided are the four sensitivity estimates (COVID-19: 99.05%,

[1] Farma Feroles. Test nasal rápido antígenos COVID-19, gripe A y B, VRS/bronquiolitis, 1 unidad [Internet]. Farma Feroles; [cited 2026 Mar 23]. Available from: https://farmaferoles.com/producto/test-nasal-rapido-antigenos-covid-19-gripe-a-y-b-vrs-bronquiolitis-1ud/

[2] Sekureco. Rapid test of COVID-19 diagnostic antigens (SARS-CoV-2 antigens) [Internet]. Sekureco; [cited 2026 Mar 23]. Available from: https://www.sekureco.eu/gb/coronavirus-covid-19/15395-rapid-test-of-covid-19-diagnostic-antigens-sars-cov-2-antigens.html

[3] Shenzhen Lvshiyuan Biotechnology Co., Ltd. Green Spring® SARS-CoV-2-Antigen-Schnelltest-Set (kolloidales Gold): 4 in 1 (Nase-Rachen, Nasal, Rachen, Lolli-Test) [image on the Internet]. Shenzhen (China): Shenzhen Lvshiyuan Biotechnology Co., Ltd.; [cited 2026 Mar 23]. Available from: https://d7wawvd1dp3nh.cloudfront.net/img/g55_xQIdb79iPYcR0iQQpg

Influenza A: 95.33%, Influenza B: 96.29%, RSV: 95.99%). A consumer purchasing this product is provided with no information that would allow them to estimate a reliable PPV for any virus at any prevalence.

The product page for the SARS-CoV-2 rapid antigen test sold by Sekureco represents a meaningfully superior—though still deficient—level of disclosure. The web page provides a full 2×2 contingency table from a clinical validation study against RT-PCR (Table 1) and declares the sample sizes explicitly ($n = 161$ total, 45 PCR-positive, 116 PCR-negative). The reported performance metrics are relative sensitivity 86.7%, relative specificity 100%, and accuracy 96.3%.

Table 1. Clinical validation data reported by Sekureco (Case 2)

	PCR positive	**PCR negative**	**Total**
Test positive	39	0	39
Test negative	6	116	122
Total	45	116	161
Sensitivity	86.7%		
Specificity	100%	(no CI reported)	
Accuracy	96.3%		

This case illustrates a genuine improvement: the consumer can view the contingency table, count the zero in the false-positive cell, and identify the sample sizes. Critically, however, no confidence interval accompanies the claim of 100% specificity. The consumer cannot determine what the true false-positive rate plausibly is. Applying the rule of three to $n = 116$ negatives, the 95% CI lower bound for specificity is approximately 97.4%, meaning the data are entirely consistent with a test that would generate one false positive in every 40 negative samples, a fact invisible in the product listing.

The product page for the Green Spring SARS-CoV-2 rapid antigen test provides the most complete statistical disclosure of the three cases in our sample: two full contingency tables (one for nasopharyngeal swabs, $n = 310$; one for anterior nasal swabs, $n = 263$), and, uniquely, 95% confidence intervals for both sensitivity and specificity under both methods (Table 2). This constitutes near-compliance with the STARD 2015 framework [1].

Table 2. Clinical validation data for the Green Spring test (Case 3)

Nasopharyngeal swabs	**PCR positive**	**PCR negative**	**Total**
Test positive	98	0	98
Test negative	2	210	212
Total	100	210	310
Sensitivity	98%	95% CI: 92.96%–99.76%	
Specificity	100%	95% CI: 98.26%–100.00%	
Accuracy	99.35%	95% CI: 97.69%–99.92%	
Anterior nasal swabs	**PCR positive**	**PCR negative**	**Total**
Test positive	121	0	121

Nasopharyngeal swabs	PCR positive	PCR negative	Total
Test negative	4	138	142
Total	125	138	263
Sensitivity	96.8%	95% CI: 92.01%–99.12%	
Specificity	100%	95% CI: 97.36%–100.00%	
Accuracy	98.48%	95% CI: 96.15%–99.58%	

Still, even in this best-practice case, the heading in the product title reads "Specificity 100%", and the reported specificity for the anterior nasal method is "100% (95% CI: 97.36%–100%)". The claim of 100% specificity remains, in our argument, scientifically indefensible, not because the calculation is wrong, but because the data cannot, by construction, falsify values of Sp below 100%.

Applied to this latter case, we ask: with what severity do the nasal validation data support various claims about the true population specificity Sp? Table 3 reports these severities for both the nasopharyngeal ($n = 210$ negatives) and anterior nasal ($n = 138$ negatives) validation panels.

Table 3. Severity of specificity applied to the Green Spring test (Case 3)

Hypothesis H	n = 210 neg. (NP)	n = 138 neg. (AN)
Sp >97% (WHO minimum)	99.8%	98.5%
Sp >99%	87.9%	75.0%
Sp >99.5%	65.1%	50.0%
Sp >99.9%	19.0%	12.9%
$Sp = 100\%$ % Exactly	0%	0%

From the nasopharyngeal data, the claim "Sp >97%" (the WHO minimum for acceptability) is supported with 99.8% severity: the data would have been extremely unlikely to yield zero false positives had the true specificity been only 97%. The claim "Sp >99%" is supported with 87.9% severity, a meaningful, if not overwhelming, evidential warrant. But the claim "Sp >99.5%" achieves only 65% severity, and "Sp >99.9%" achieves barely 19%. The severity of the literal claim "Sp =100%" is, by definition, zero: observing zero false positives in any finite sample provides no evidence against the possibility that the true false-positive rate is, say, 0.1% or 0.5%, because those values are entirely compatible with the observed data.

To achieve 95% severity for the claim "Sp >99%" would require approximately $n = 300$ negative validation samples, a figure far exceeding the sample sizes of all three cases examined here, and of most rapid antigen test validation studies reported in the literature. This calculation is never presented in any consumer-facing product communication: it is precisely this absence of severity-informed inference that allows the "100% specificity" label to circulate as though it were scientifically warranted.

The contrast between the three cases is thus not merely one of quantity of information: Case 3 discloses more than Cases 1 or 2, and for that, it deserves recognition. But the severity analysis reveals that even Case 3, the nearest-best-practice example we identified, does not provide the

consumer with the tools to distinguish between a test that truly cannot produce false positives and one whose true false-positive rate is 0.5%, and can produce a positive predictive value (PPV) error of over 30 percentage points in low-prevalence settings.

Consider, for example, a test with $Se = 97\%$ and a true $Sp = 99.5\%$ (plausible values, and better than ECDC minimum recommendations) used in a community screening program at a prevalence of 1% (a typical inter-peak respiratory season prevalence). The true PPV is: PPV $\approx 0.66\ (66\%)$. A consumer who has been told that the test has "100% specificity" will instead believe PPV = 100%, meaning they will be certain they are infected when in fact, one third of positive results in this scenario are false positives. The error is not inconsequential: it represents the difference between mandatory self-isolation, unnecessary clinical consultation, and the downstream anxiety of a false diagnosis, versus the realization that a confirmatory test may be warranted.

The positive and negative predictive values of a test depend on the epidemiological situation (i.e., prevalence) as well as on the test's performance characteristics. Package labeling should therefore report the expected risk of false negative and false positive results as a function of prevalence [2]. The Cochrane review by Dinnes et al. [10] demonstrated that at a 0.5% prevalence, PPVs of tests ranged from 38% to 52%, meaning that a considerable number of results may be false positives—a finding entirely obscured by the specificity claims in commercial packaging.

DISCUSSION

Taken together, the arguments and examples in this paper point to a simple conclusion: the current use of "100% specificity" in rapid diagnostic testing is very hard to defend, both statistically and epistemologically. The problem appears in at least two common scenarios. In one, false positives have been observed, but the confidence intervals are clipped at 100%, as if the fact could be undone arithmetically. In the other, a finite validation produces no false positives, and the result is treated as definitive evidence about the population. In both cases, the uncertainty inherent in test validation is blurred, and that blurring is not harmless when clinical decisions depend on the numbers.

The persistence of 100% specificity claims in commercial product communications reflects a convergence of statistical naivety, inappropriate methodological choices, and commercial incentives. The asymptotic Wald interval, despite well-documented limitations near parameter boundaries [4,5], remains widely used in diagnostic accuracy studies. Its truncation to 100% when the upper bound exceeds unity is a mathematical convenience that transforms a falsified hypothesis into a plausible one. Regulatory standards should explicitly prohibit this practice and mandate intervals that respect the binomial parameter space, such as the Wilson score or the Clopper-Pearson method.

The rule of three [6] provides a simple corrective to the misinterpretation of zero-event samples. That a manufacturer testing 116 negative samples can claim "100% specificity" without acknowledging that the data are equally compatible with a true specificity of 97.4% represents a failure of scientific transparency. While academic reporting guidelines such as STARD [1] mandate disclosure of confidence intervals and sample sizes to prevent over-interpretation in research publications, commercial labeling regulations do not systematically enforce these editorial principles.

The Mayo-Spanos severity framework [7–9] offers a principled approach to quantifying what finite data can and cannot support. Our analysis of the Green Spring test (Case 3), the most transparent example in our sample, revealed that while the data provide strong support for specificity exceeding 97%, their support is weak for claims above 99.5%. The claim of exactly 100% specificity receives zero severity, as it

should: no finite sample can ever provide evidence for a point hypothesis at the boundary of the parameter space. This is not a technical subtlety but a fundamental feature of statistical inference that should be reflected in diagnostic test reporting.

A final pathway by which the 100%-specificity label can enter product communications deserves brief attention, though it requires no additional empirical evidence: the application of standard rounding conventions to point estimates that are close to, but strictly below, unity. A test that yields 2 false positives among 5,000 negative validation samples has a point specificity of 99.96%. To 1 decimal place, this rounds to 100.0%; to an integer, it rounds to 100%. The mathematical consequence is identical to the case of observed perfect performance: the term $(1 - Sp)$ becomes zero, and the PPV formula collapses to certainty.

The distinction between $0.9996 \neq 1.0000$ is not merely semantic. A test with $Sp = 100\%$ is deterministically incapable of producing a false positive; a test with Sp =99.96% will produce, in expectation, approximately one false positive per 2,500 healthy individuals tested. In a national screening program of ten million people, this difference corresponds to approximately 4,000 false positives that would not occur if specificity were truly 100%. When the distinction is erased by rounding in a product label, the consumer cannot perform this calculation. Regulatory standards should therefore specify that any specificity reported as a rounded figure of 100% must be accompanied by the unrounded point estimate and its confidence interval, so that the approximation is transparent.

The Bayesian collapse of PPV under the assumption of perfect specificity is not merely a theoretical curiosity. When consumers, clinicians, or public health officials operate under the belief that a positive test result is definitive, they may forgo confirmatory testing, initiate unnecessary treatments or isolation measures, and misinterpret the epidemiological significance of test results. At a population level, widespread reliance on tests with unqualified specificity claims can distort surveillance data and erode public trust when inevitable false positives occur.

The ECDC has explicitly noted that predictive values depend on prevalence and test performance [2], and Dinnes et al. have demonstrated that at low prevalence, a substantial proportion of positive results may be false [10]. These findings are directly undermined when specificity is communicated as 100% without appropriate qualification.

Limitations

This analysis has several limitations. The case illustrations were selected through a non-systematic search and are not intended to be representative of the entire market. The severity calculations assume perfect study conduct and no structural sources of error beyond sampling variability; in practice, manufacturing variability, operator factors, and antigenic cross-reactivity further erode the justification for 100% claims [11]. Additionally, our analysis focused on specificity, but analogous concerns apply to sensitivity claims of 100% when no false negatives are observed.

Implications for policy and practice

Addressing the problem of 100% specificity claims requires coordinated action across multiple stakeholders. Regulatory agencies should revise their guidance to require that all diagnostic test communications include confidence intervals computed by appropriate methods and prohibit unqualified point estimates of 100%. The European Union's In Vitro Diagnostic Regulation [12] provides a framework for such requirements. Manufacturers should design validation studies with sufficient sample sizes to generate informative confidence intervals even at extreme proportions. Based on severity calculations, achieving 95% severity for the claim that specificity exceeds 99% would require approximately 300 negative samples, a feasible

target that exceeds most current validation studies.

Finally, public health communicators and pharmacists should convey the probabilistic nature of diagnostic test results, particularly in low-prevalence contexts. A positive result is not a diagnosis but an update on the prior probability of infection; this message is systematically undermined by the language of "100% specificity."

CONCLUSION

The repeated promise of "100% specificity" in rapid diagnostic tests does not sit comfortably with what statistical reasoning can genuinely support. When a validation study has already observed at least one false positive, the problem is immediate: the statement of perfection has been openly contradicted by the data. Even when no false positives are seen, the situation is only superficially better. Finite samples, manufacturing variability, and real-world use all leave room for low but meaningful false-positive rates that cannot be ruled out in practice.

The three commercially available tests examined here show that this is not a hypothetical concern. Even in the most transparent example, where confidence intervals are provided and sample sizes are explicit, the headline claim of 100% specificity remains, and the severity calculations suggest only modest support for values above 99.5%. In low-prevalence settings, treating that headline as literally true can substantially inflate positive predictive value and, in turn, influence how clinicians, patients, and public health agencies interpret positive results.

Moving away from this language of perfection does not require radical new methods; rather, it requires more honest use of familiar tools. Confidence intervals based on Wilson or Clopper-Pearson approaches are well described in the literature, and nothing prevents manufacturers from reporting unrounded estimates alongside them. Likewise, validation studies can be designed with sample sizes large enough to probe high-specificity claims with adequate severity, even if they can never prove a point value of 1.00.

Ultimately, a diagnostic test is a probabilistic instrument, not an oracle. When its performance is described as if it ensured certainty, the basic logic of evidence-based medicine is weakened, and expectations are set in a way that real data cannot satisfy. Being explicit about uncertainty is not a matter of statistical pedantry; it is a practical requirement if we want patients and professionals to trust that the numbers on a test box reflect what the test can do, rather than what the label would like them to believe.

AI DISCLOSURE

Generative AI tools were employed exclusively to assist with the editing of language, organisation, and formatting of the manuscript. The author has thoroughly reviewed, verified, and edited all content and assumes full responsibility for the accuracy and integrity of the work. No confidential or identifiable information was input into the AI tools.

REFERENCES

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF for the STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. doi: 10.1136/bmj.h5527.
2. European Centre for Disease Prevention and Control. Options for the use of rapid antigen tests for COVID-19 in the EU/EEA and the UK. Stockholm: ECDC; 2020 Nov 19. 33 p.
3. Xie JW, He Y, Zheng YW, Wang M, Lin Y, Lin LR. Diagnostic accuracy of rapid antigen test for SARS-CoV-2: a systematic review and meta-analysis of 166,943 suspected COVID-19 patients. Microbiol Res. 2022;

265:127185. doi: 10.1016/j.micres.2022.127185.
4. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat. 1998;52(2):119–26. doi: 10.1080/00031305.1998.10480550.
5. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Stat Sci. 2001;16(2):101–33. doi: 10.1214/ss/1009213286.
6. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983;249(13):1743–5. doi: 10.1001/jama.1983.03330370053031
7. Mayo DG. Statistical inference as severe testing: how to get beyond the statistics wars. Cambridge: Cambridge University Press; 2018. 486 p.
8. Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. Br J Philos Sci. 2006;57(2):323–57. doi: 10.1093/bjps/axl003.
9. Mayo DG, Spanos A. Error statistics. In: Bandyopadhyay PS, Forster MR, editors. Philosophy of statistics. Amsterdam: Elsevier; 2011. p. 153–98. doi: 10.1016/B978-0-444-51862-0.50005-8.
10. Dinnes J, Deeks JJ, Adriano A, Berhane S, Davenport C, Dittrich S, Taylor M, Emperador D, Takwoingi Y, Cunningham J, Beese S, Domen J, Dretzke J, Ferrante di Ruffano L, Harris IM, Price MJ, Taylor-Phillips S, Hooft L, Leeflang MMG, McInnes MDF, Spijker R, Van den Bruel A, Arevalo-Rodriguez I, Buitrago DC, Ciapponi A, Mateos M, Stuyf T, Horn S, Salameh JP, McGrath TA, van der Pol CB, Frank RA, Prager R, Hare SS, Dennie C, Jenniskens K, Korevaar DA, Cohen JF, van de Wijgert J, Damen JAAG, Wang J, Agarwal R, Baldwin S, Herd C, Kristunas C, Quinn L, Scholefield B. Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. Cochrane Database Syst Rev. 2021;3(3):CD013705. doi: 10.1002/14651858.CD013705.pub3.
11. World Health Organization. Antigen-detection in the diagnosis of SARS-CoV-2 infection: interim guidance. Geneva: WHO; 2021 Oct 6. 13 p.
12. European Parliament and Council of the European Union. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices. Official Journal of the European Union. 2017 May 5; L117:176–332.