The Post Hoc Pitfall: Rethinking Sensitivity and Specificity in Clinical Practice

José Nunes de Alencar Neto, MD[1], *and Leopoldo Santos-Neto, MD, PhD*[2]

[1]Instituto Dante Pazzanese de Cardiologia, São Paulo, Brazil; [2]Universidade de Brasília, Brasília, Brazil

J Gen Intern Med
DOI: 10.1007/s11606-024-08692-z

INTRODUCTION

In medical education, sensitivity and specificity are often emphasized as essential criteria for evaluating the efficacy of diagnostic tests.[1] While these measures are pivotal, their application in isolation, as posited by the Spin and Snout mnemonics[2], is not without limitations in the clinical environment. The article unfolds the shortcomings of this reliance, highlighting their post hoc nature and the disconnect this creates in the context of pre hoc, or forward-looking, clinical diagnostics.

Subsequently, we will delve into the subject through hypothetical illustrative scenarios, postulating that likelihood ratios (LRs) present compelling alternatives. We will examine how, cognitively, LRs necessitate probabilistic thinking from clinicians by their very definition—a critical aspect often underappreciated in medical diagnostics.

THE POST HOC NATURE OF SENSITIVITY AND SPECIFICITY

Sensitivity and specificity are metrics calculated from studies where participants' health status is already known. In contrast, clinical practice often requires "pre hoc" or "forward-looking" diagnostic tests to determine an unknown health outcome.[3] This creates a significant disconnect between the retrospective nature of these metrics and the prospective needs of clinical practice.[4]

Sensitivity measures how well a test identifies true positives among those with the disease. Specificity gauges the test's ability to correctly identify true negatives among healthy individuals. Mathematically:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{People with Disease}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{Healthy Individuals}}$$

Received October 6, 2023
Accepted February 20, 2024

Published online: 27 February 2024

As an example, suppose a doctor prescribes a diagnostic test that possesses a sensitivity and specificity of precisely 90% in order to identify a specific disease. It is tempting to assume that the patient has the disease with a 90% probability when the test is positive. This reasoning is fallacious. Sensitivity and specificity are not derived from the uncertain clinical scenarios to which these tests are frequently applied, but rather from populations with known disease status. Indeed, the doctor is employing the exam specifically to ascertain the patient's unidentified health condition.

THE PROPOSED ALTERNATIVE: ADVOCATING FOR A WIDER USE OF LIKELIHOOD RATIOS IN CLINICAL DECISION-MAKING

In contrast to sensitivity and specificity, likelihood ratios are more naturally applied in a "pre hoc" manner, allowing clinicians to update their diagnostic probabilities based on new evidence. Mathematically, the LR + and the LR − are defined as follows:

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

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

It is crucial to acknowledge that likelihood ratios are derived from sensitivity and specificity. However, its application offers distinct advantages. The perspective provided by likelihood ratios is advantageous because it advocates for a different view than usual: given that a test result is positive, by how many times does the chance of the patient having the disease increase? And if it is negative, by how many times will this chance decrease?

Notice that this perspective, in terms of probability, requires the physician to take a step back and think about the chance (or probability) of the patient having the disease in question. This insight is not provided by sensitivity and specificity alone.

To provide an illustration, consider a test characterized by a sensitivity of 20% and a specificity of 90%. A physician might be tempted to conclude, based solely on this information, that a positive test result signifies a 90% chance that the patient has the disease; however, this is not the case. Conversely, by focusing on likelihood ratios, they will ascertain that the LR + is 2.0,

signifying that a positive test result will result in a doubling of the patient's probability of contracting the disease.

$$LR+ = \frac{\text{Sensitivity}}{(1 - \text{Specificity})} = \frac{0.20}{(1 - 0.90)} = \frac{0.20}{0.10} = 2$$

$$LR- = \frac{(1 - \text{Sensitivity})}{\text{Specificity}} = \frac{(1 - 0.20)}{(0.90)} = \frac{0.80}{0.90} = 0.89$$

But double from what to what? If the disease probability is 5%, it will be 10% after the test, not 90% as determined by the specificity. This is where the use of likelihood ratios "forces" probabilistic thinking. They compel physicians to consider the pre-test probability or the baseline rate of a disease in a given population, a step often overlooked when relying solely on sensitivity and specificity.[5,6]

BAYESIAN REASONING IN CLINICAL PRACTICE: A DYNAMIC APPROACH TO DIAGNOSING

In medicine, the probabilistic nature of diagnosis is often overlooked, leading to a cognitive bias known as base-rate neglect.[7] Clinicians may focus too intently on the sensitivity and specificity of a test, neglecting the initial likelihood or actual prevalence of a disease in the population. This oversight can distort the application of Bayesian reasoning,

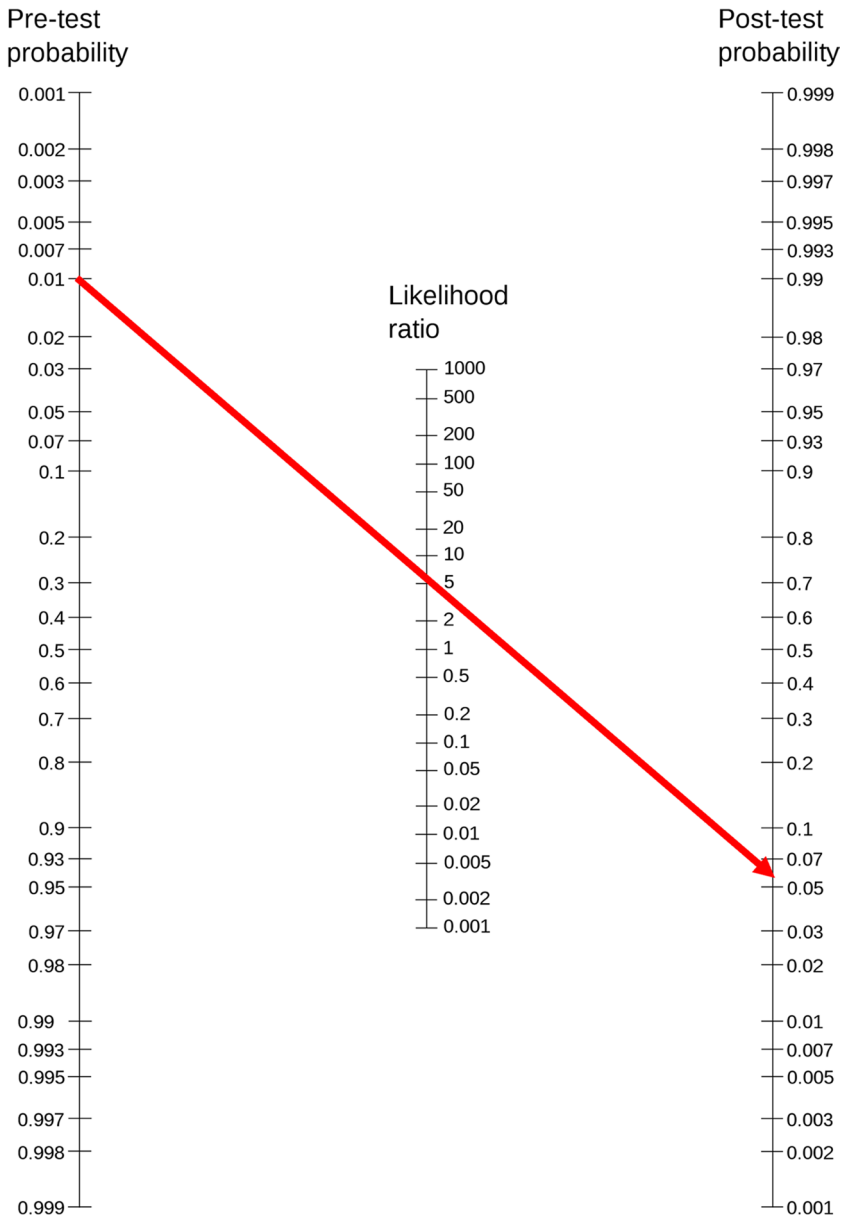

Figure 1 The calculation of the post-test probability for an acute myocardial infarction in an asymptomatic patient exhibiting ST-segment elevation is illustrated using Fagan's nomogram. The derived post-test probability is 6.5%, which is obtained by intersecting a pre-test probability of 1% with an LR + value of 6.83. The visual depiction underscores the importance of integrating likelihood ratios when enhancing diagnostic probabilities.

resulting in flawed clinical decisions. Bayesian reasoning represents a dynamic framework in medical decision-making. This approach integrates prior probabilities and incorporates the diagnostic performance of a test. The strength of Bayesian reasoning lies in its ability to constantly update and adapt to new information, thereby offering a more nuanced and patient-centered diagnostic process.[8]

To illuminate the application of Bayesian reasoning, let us contemplate an alternative scenario. Consider a patient who exhibits ST-segment elevation during an electrocardiogram (ECG) for screening purposes. Nevertheless, the patient exhibits no clinical symptoms or symptoms consistent with acute coronary syndrome (ACS). An accuracy study reported a sensitivity of 41% and a specificity of 94% for ST-segment elevation when diagnosing occlusion myocardial infarction.[9] At first glance, this finding might be interpreted as suggesting a 94% probability of the patient having the disease when ST-segment elevation is present. Using standard formulas, we find:

$$\mathrm{LR}+ = \frac{\text{Sensitivity}}{(1-\text{Specificity})} = \frac{0.41}{(1-0.94)} = \frac{0.41}{0.06} = 6.83$$

$$\mathrm{LR}- = \frac{(1-\text{Sensitivity})}{\text{Specificity}} = \frac{(1-0.41)}{(0.94)} = \frac{0.59}{0.94} = 0.62$$

Given the patient's asymptomatic status, the clinician estimates the pre-test probability of occlusion myocardial infarction to be about 1%. Using the Fagan nomogram[10] and applying the LR + of 6.83 to this pre-test probability, the post-test probability is calculated to be around 6.5%. This means that despite the ST-segment elevation, there is approximately 93.5% chance that the patient is not experiencing an acute coronary occlusion (Fig. 1).

Alternatively, this post-test probability can be calculated through the following steps:

1. Convert pre-test probability to odds:

$$\text{Odds (Pre}-\text{test)} = \frac{\text{Probability}}{1-\text{Probability}}$$

2. Multiply by LR:

$$\text{Odds (Post}-\text{test)} = \text{Odds(Pre}-\text{test)} \times \text{LR}$$

3. Convert post-test odds to probability:

$$\text{Probability (Post}-\text{test)} = \frac{\text{Odds (Post}-\text{test)}}{1+\text{Odds (Post}-\text{test)}}$$

NATURAL FREQUENCIES: ANOTHER WAY TO USE BAYESIAN REASONING

Another pertinent approach within Bayesian reasoning is the use of natural frequencies, a method that involves constructing a decision tree to visually represent how diagnostic tests

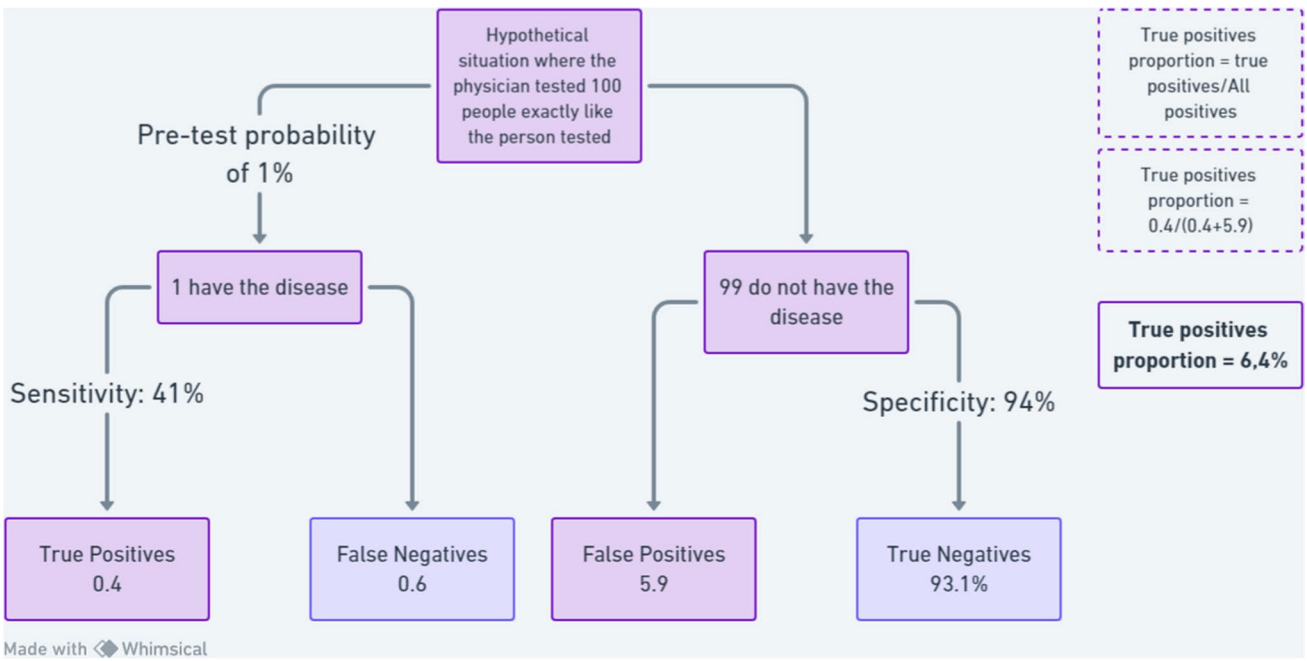

Figure 2 By employing natural frequencies and starting with the base rate, a physician can more accurately determine the post-test probability of a positive test result being true. In this model, we consider a hypothetical cohort of 100 individuals who closely resemble the patient under investigation in terms of age, comorbidities, and symptoms. Based on the physician's estimated pre-test probability of 1%, 1 individual in this cohort is assumed to have the disease. In a population consisting of 1 diseased and 99 healthy individuals, it becomes evident that the proportion of true positives among all positive test results is 6.4%. This value represents the post-test probability and signifies the likelihood that a positive test result is indeed accurate.

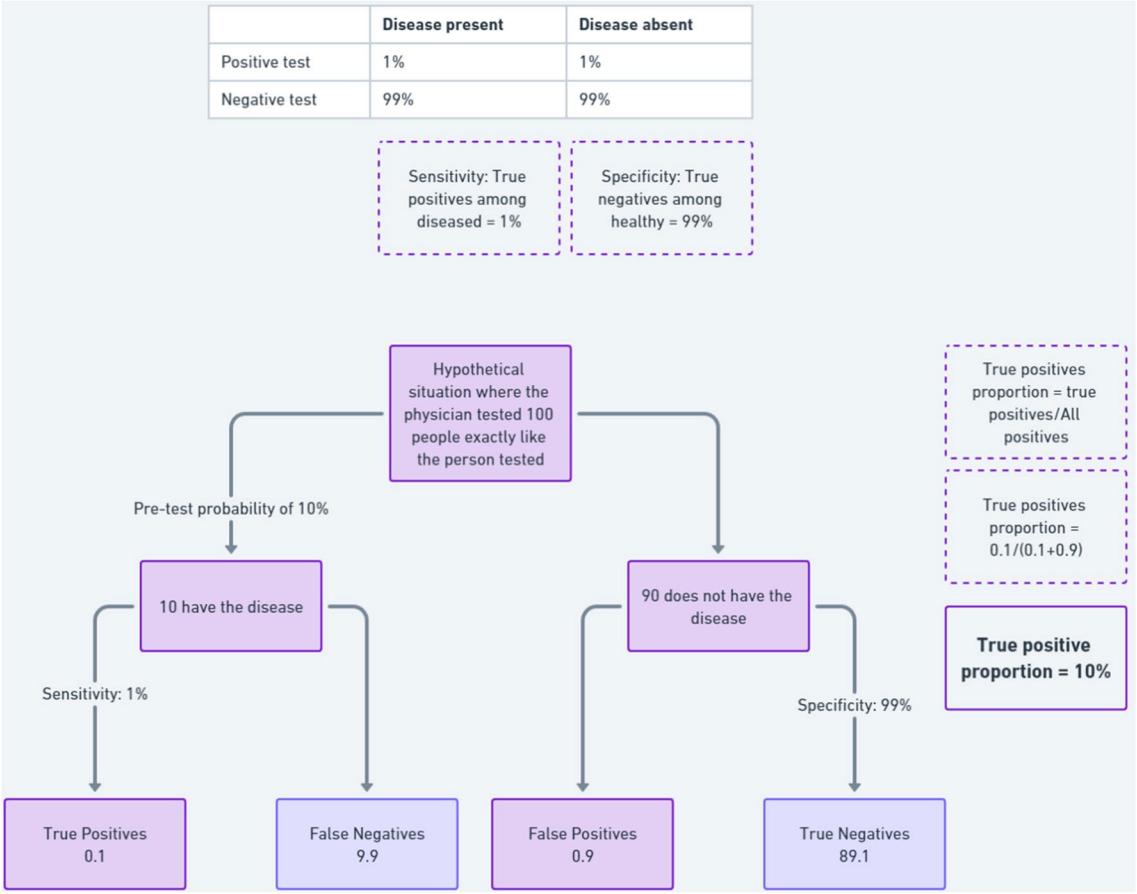

Figure 3 The concept of natural frequencies exemplifies the utilization of a diagnostic test whose prevalence is equivalent between healthy and diseased individuals, thereby making it useless in clinical settings to ascertain the presence or absence of disease. The issue of sensitivity and specificity adding up to 1.0 becomes apparent in such circumstances. Notwithstanding the test's apparent high specificity (99%), its clinical utility is rendered futile on account of its likelihood ratio of 1.0. This means that upon receiving a positive test result, the probability of disease presence is effectively multiplied by 1.0, while the post-test probability remains unchanged from the pre-test probability.

interact with pre-existing probabilities, thereby enhancing our understanding of a patient's health status.[11,12] The approach begins with the pre-test probability, which is derived from epidemiological data or the clinician's evaluation of the likelihood of the disease prior to conducting the test. A decision tree then divides into "Diseased" and "Healthy" branches, which further subdivide True Positives, False Negatives, True Negatives, and False Positives in accordance with test outcomes (Fig. 2).

While natural frequencies can intuitively convey the probabilistic nature of test interpretations, their integration into clinical practice is not straightforward. Often, this methodology does not align with the typical cognitive framework of practitioners, leading to underutilization in actual patient care. Despite their potential to demystify complex statistical concepts, natural frequencies remain an underemployed strategy in the diagnostic process.

The Case of Tests with Sensitivity + Specificity = 1

When the sum of sensitivity and specificity equals 1.0, an intriguing instance of curiosity arises. Consider, for example, a test whose sensitivity is 5% and its specificity is 95%. Upon initial examination, the test's high specificity may indicate its efficacy in definitively diagnosing the disease. But this could not be further from the truth. The mathematical expressions for calculating the LR + and LR − are as follows:

$$\mathrm{LR}+ = \frac{\text{Sensitivity}}{(1 - \text{Specificity})} = \frac{0.05}{(1 - 0.95)} = \frac{0.05}{0.05} = 1$$

$$\mathrm{LR}- = \frac{(1 - \text{Sensitivity})}{\text{Specificity}} = \frac{(1 - 0.05)}{(0.95)} = \frac{0.95}{0.95} = 1$$

A value of 1.0 is produced by both LR + and LR −, signifying that the test has no effect on the pre-test probability of the disease. Alternatively stated, a patient's post-test probability would remain at 10% regardless of the outcome of the test, whether it be positive or negative, if their pre-test probability is 10%. Notwithstanding its notable specificity, the test is fundamentally ineffective in either validating or excluding the disease.

This phenomenon occurs when a test has no actual diagnostic power for the disease in question. Since the disease

has no correlation with the test, both groups—those labeled as diseased and those labeled as healthy—are essentially composed of the same individuals and will test positive or negative merely by chance. Consequently, the prevalence of the disease will naturally be similar in both groups.

Interestingly, the rarer the disease under study (which, again, has no actual correlation with the test), the more inflated the specificity will appear. This is because there will be more true negatives in the sample defined as healthy, artificially boosting the specificity (Fig. 3).

It is important to clarify that while likelihood ratios (LRs) are derived from sensitivity and specificity, they reframe this information in a manner that is more directly applicable to clinical decision-making. Emphasizing this point, it becomes evident that knowing the exact values of sensitivity and specificity is less critical than understanding how LRs should be interpreted. A value of 1.0 for an LR means multiplying the chance by 1, essentially keeping it the same. In contrast, a specificity of 90% might seem appealing based on the "Spin and Snout" mnemonic, but if the sensitivity is only 10%, the test will not be useful. This conclusion is not obvious when analyzing sensitivity and specificity in isolation.

CONCLUSION

In synthesizing our findings, this article reaffirms the value of likelihood ratios (LR+) and (LR−) in clinical practice, not simply as substitutes for traditional sensitivity and specificity, but as cognitively superior tools for diagnostic reasoning within a Bayesian framework. This assertion rests on the premise that while LRs are indeed derived from sensitivity and specificity, their utilization promotes a probabilistic mode of thinking that is not inherently elicited by sensitivity and specificity alone.

The central argument of this paper is that LRs facilitate a cognitive shift towards probabilistic reasoning, thereby enhancing the physician's ability to calibrate diagnostic hypotheses more effectively. This shift is critical, as it moves beyond the raw metrics of test accuracy to encompass the nuances of clinical context and patient-specific probabilities. The definition of LRs themselves—quantifying how much a positive or negative test result shifts the odds of having a disease—inherently guides clinicians to consider the magnitude of change in disease probability, a conceptual leap that is less apparent when considering sensitivity and specificity in isolation.

Furthermore, while the construction of natural frequencies offers an alternative Bayesian approach, it is not as intuitively accessible as the straightforward calculation and interpretation of LRs. Therefore, we argue for the broader adoption of LRs as an essential component of a comprehensive diagnostic strategy, one that better navigates the complexities and uncertainties of medical practice. Through this lens, LRs are not merely mathematical derivatives but pivotal instruments that prompt clinicians to engage more deeply with the probabilistic nature of diagnosis and treatment decisions.

Corresponding Author: *José Nunes de Alencar Neto, MD; , Instituto Dante Pazzanese de Cardiologia, São Paulo, Brazil (e-mail: jose.alencar@dantepazzanese.org.br).*

Declarations

Conflict of interest *The authors have no conflicts of interest to declare.*

REFERENCES

1. **Altman DG, Bland JM.** Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994;308(6943):1552. https://doi.org/10.1136/bmj.308.6943.1552
2. **Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M.** Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ. 2004;329(7459):209-213.
3. **Naeger DM, Kohi MP, Webb EM, Phelps A, Ordovas KG, Newman TB.** Correctly Using Sensitivity, Specificity, and Predictive Values in Clinical Practice: How to Avoid Three Common Pitfalls. Am J Roentgenol. 2013;200(6):W566-W570. https://doi.org/10.2214/AJR.12.9888
4. **Moons KGM, Harrell FE.** Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol. 2003;10(6):670-672. https://doi.org/10.1016/s1076-6332(03)80087-9
5. **Deeks JJ, Altman DG.** Diagnostic tests 4: likelihood ratios. BMJ. 2004;329(7458):168-169. https://doi.org/10.1136/bmj.329.7458.168
6. **Cahan A, Gilon D, Manor O, Paltiel O.** Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? QJM Int J Med. 2003;96(10):763-769. https://doi.org/10.1093/qjmed/hcg122
7. **O'Sullivan ED, Schofield SJ.** Cognitive bias in clinical medicine. J R Coll Physicians Edinb. 2018;48(3):225-232. https://doi.org/10.4997/JRCPE.2018.306
8. **de Alencar Neto JN.** Applying Bayesian reasoning to electrocardiogram interpretation. J Electrocardiol. Published online October 17, 2023. https://doi.org/10.1016/j.jelectrocard.2023.10.006
9. **Meyers HP, Bracey A, Lee D, et al.** Accuracy of OMI ECG findings versus STEMI criteria for diagnosis of acute coronary occlusion myocardial infarction. IJC Heart Vasc. 2021;33:100767. https://doi.org/10.1016/j.ijcha.2021.100767
10. **Fagan TJ.** Letter: Nomogram for Bayes theorem. N Engl J Med. 1975;293(5):257-257. https://doi.org/10.1056/NEJM197507312930513
11. **Binder K, Krauss S, Schmidmaier R, Braun LT.** Natural frequency trees improve diagnostic efficiency in Bayesian reasoning. Adv Health Sci Educ Theory Pract. 2021;26(3):847-863. https://doi.org/10.1007/s10459-020-10025-8
12. **Gigerenzer G, Hoffrage U.** How to improve Bayesian reasoning without instruction: Frequency formats. Psychol Rev. 1995;102(4):684-704. https://doi.org/10.1037/0033-295X.102.4.684