

PV-LR: AN INTERACTIVE LEARNING APP LINKING THE TWO APPROACHES TO UPDATING OF PRE-TEST PROBABILITIES

SUMMARY

Measures of diagnostic-test performance and the calculations used to update a pre-test probability remain difficult for many learners to connect and apply. The difficulty reflects unfamiliar terminology, the coexistence of predictive-value (PV) and likelihood-ratio (LR) formulations, and arithmetic that can obscure their shared logic. We describe an interactive learning app (PV-LR) designed to explain the concepts and to connect the two approaches. It is intended for guided exploration of diagnostic reasoning, rather than rapid point-of-care calculation. Learners first construct an imagined diagnostic test by choosing an acceptable false-positive rate and a desired detection rate (sensitivity). The app displays the selected specificity and sensitivity, and the resulting positive and negative likelihood ratios, and expected distribution of test results in words, numbers, and pictures. Learners then apply the same test to clinical scenarios with different pre-test probabilities and observe the corresponding post-test probabilities after positive and negative results. By keeping this sequence on one interactive page, the app links test performance, clinical context, and probability updating without treating either formulation as a black box.

Link to app: <https://jhanley.biostat.mcgill.ca/med2/PV-LR.html>

INTRODUCTION

The terms, concepts and calculations involved in using measures of diagnostic test performance to update pre-test probabilities continue to challenge both teachers and students. In part it is because of the terminology itself;¹ in part it is because there are two seemingly-very-different approaches to the calculations; and in part it is because of the mental (or calculator-based) arithmetic involved in these formulae.

More than a hundred years on from when they were coined in the context of testing for syphilis [Binney, 2001], we are unlikely to replace the problematic-for-some terms *sensitivity* and *specificity*. Just like earlier writers such as Galen and Gambino [1975], we may nevertheless help learners by explaining how these terms relate to alternative terminology that has appeared in the teaching literature.²

One way of calculating post-test probabilities uses predictive values; another uses likelihood ratios. These are two seemingly different approaches to using these operating characteristics of the test to update from a pre-test to a post-test probability; and they go back a long way. On the face of it, the calculations via the *predictive value* approach seem simpler, but the formula itself is not easily recalled. Thus, it was no surprise that when their skills were tested, most participating physicians and medical students gave incorrect answers. [Berwick, 1981] The major drawback of the approach is that the two quite separate components (the pre-test probability and the intrinsic properties of the test itself) are handled *simultaneously*.

The mathematically equivalent *likelihood-ratio formulation* does better in this regard, by cleanly separating these two components. However, to use this more parsimonious

¹ “It is unfortunate that we are saddled with words that lack sharpness, but as much as we'd like to dispose of these words they are so established in the special literature of public health that we cannot eliminate them.” [Galen and Gambino 1975]

² “On the other hand, we do suggest some synonyms, and by example we hope to define more clearly what is meant when these words are used and why we must distinguish between them.” [ibid]

approach, one had to be agile enough to carry out the intermediate calculations on the odds, rather than the probability, scale. Thus, despite many further expositions, it was not until Fagan's ingenious "Nomogram for Bayes Theorem" [1975] that the need for explicit intermediate arithmetic on the odds scale was removed. However, the nomogram itself is very much a black box. Fagan did not use the term likelihood ratio; nor did he have the space, in the few hundred words his letter was allotted, to explain why the nomogram worked, and why it was mathematically equivalent to the other approach. Despite the substantial efforts of subsequent writers (e.g., the McMaster Group in their tutorials in the Canadian Medical Association Journal in the early 1990s) to make it less of a black box, the LR approach continues to pose a challenge for many medical-school teachers.

I have noticed that, when defining an LR, some teachers focus on the formula rather than explaining its meaning in words. They do explain that it makes sense that a ratio above/below unity should move up/down the pre-test probability or odds, that the formula is a bit complicated, but that the steps can be avoided by using the Fagan nomogram. The mathematical bridge between the two prevailing approaches tends to be skipped. One reason may be to avoid confusion: no matter the test result, the LR approach stays focused on the probability that the *condition is present*. In the case of a *positive test result*, the resulting post-test probability coincides with the positive predictive value. But in the case of a *negative* result, the resulting post-test probability is *pushed down*, whereas the negative predictive value, which focuses on *the 'correctness' of the result*, is *close to 1*.

In addition, queries from curious first year medical students as to what the Fagan nomogram is based are usually deflected, or answered by just saying that it 'uses Bayes' Theorem'— as if predictive values approach do not!

To address these teaching challenges, we developed an interactive learning app. Using words, diagrams, and numbers, it first invites learners to construct an imagined diagnostic test with a chosen level of performance. The resulting sensitivity, specificity, and likelihood ratios become concrete and connected. Learners then change perspective and apply the same test to clinical scenarios with different pre-test probabilities.

THE APPLICATION: ARCHITECTURE AND USE

The app is built on a single canvas where user inputs, numbers, diagrams, and text dynamically annotate one another. To minimize cognitive load, the interface uses an interactive sequence that guides the learner through three distinct phases (see Figures 1-4).

PHASE 1: ESTABLISHING THE INTRINSIC TEST PERFORMANCE

The app begins by presenting examples of latent conditions, noting that prior to clinical adoption, diagnostic tests are evaluated in two sub-populations: people known not to have the target condition and people known to have it. The ‘unknown’ (the probabilistic aspect) is the test result in each person in each condition.

Setting the false-alarm rate first: The learner selects the specificity first, looking only at how the test performs in people *without* the condition. It explains why this order is not arbitrary, or lexical; it is a fundamental principle. Only then does the user move on to its performance in the other sub-population.

To accommodate different cognitive styles, *the app integrates two viewpoints*: it explicitly frames performance metrics both as error frequencies (the statistician’s focus on false alarms and misses) and success frequencies (the developer’s focus on specificity and sensitivity). Once the user inputs sensitivity and specificity, the app automatically calculates and displays their mathematical equivalents: the positive and negative likelihood ratios (LR+ and LR-). As soon as the user indicates they want to apply the test, the levels of the simulated performance of the test per se become locked in.

PHASE 2: SETTING THE CLINICAL CONTEXT

After establishing these fixed operating characteristics, the user is prompted to tap on a probability scale to set the clinical pre-test probability. In anticipation of some learners wishing to explore the likelihood-ratio approach, the app simultaneously displays this probability value on the odds scale.

PHASE 3: DUAL PROBABILITY UPDATING

Clicking on the pre-test probability opens the predictive value panel and the likelihood ratio panel at the same time, showing that both updating frameworks lead to the exact same result.

For both positive and negative results, the app simultaneously displays in numbers and pictures the updated post-test probabilities through these two side-by-side frameworks:

The Predictive Value Panel shows what percentage of the positive or negative results are true matches (PPV and NPV) and thus come from the ‘condition-present’ and ‘condition-absent’ sub-populations respectively.

The Likelihood Ratio-based Panel illustrates how the pre-test odds (and thus the probability) that the patient has the condition is moved up or down the odds scale (and thus the probability) scale by the test result.

Learners can then tap on other pre-test probabilities to simulate other clinical scenarios, but the underlying test characteristics remain locked. This interactive separation reinforces the fundamental concept that while a test's physical capabilities are considered fixed, its predictive power relies entirely on the clinical population to which it is applied.

The app is designed to be self-contained. However, an online file, found here, https://jhanley.biostat.mcgill.ca/med2/PV-LR_TeachersNotes.pdf contains additional commentary, some of which the teacher may wish to incorporate/adapt as he/she accompanies students through the app.

IN CLOSING

The interactive app is aimed at beginning medical students and their teachers. Its mix of the verbal, the numerical and the visual is meant to help them ‘own’ the concepts. The app addresses both the older PPV approach and the somewhat newer LR approach—but without taking sides. It is not designed to be a slick smartphone app that provides rapid numerical calculations for busy clinicians; rather, it is for learners who wish to take a little time to understand what the unfamiliar terminology and the associated numbers mean.

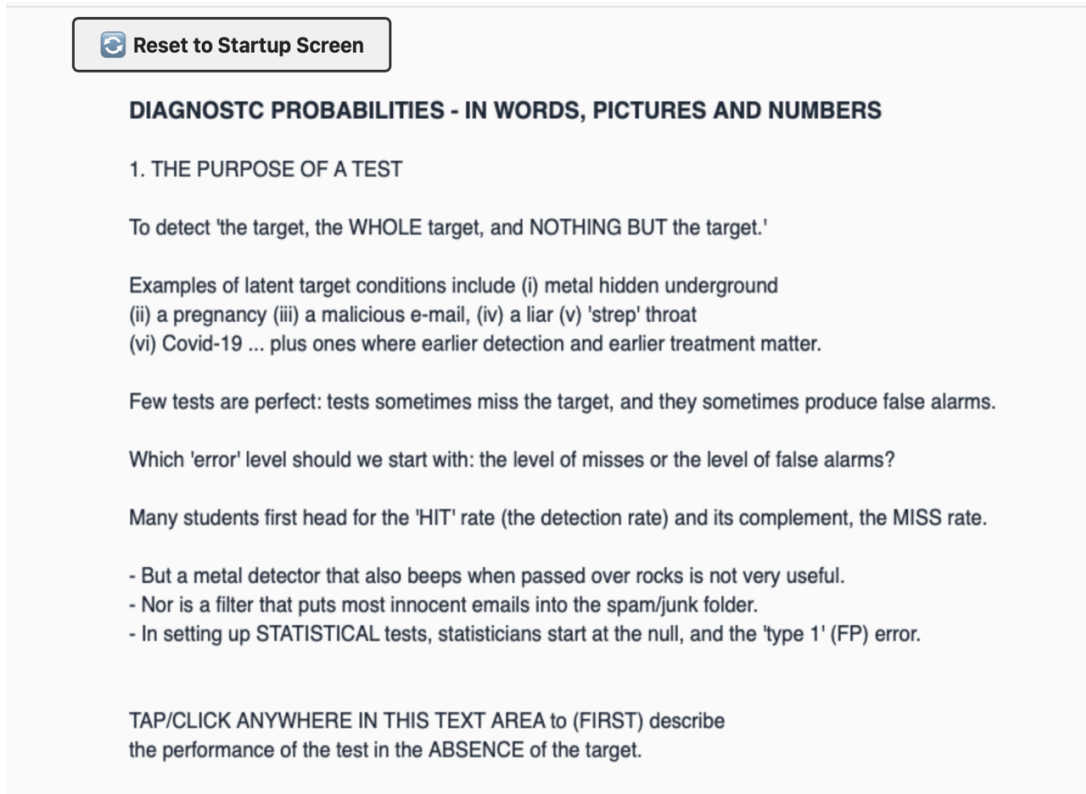


Figure 1: The opening screen of the learning application.



Figure 2: First, the learner sets the false-alarm rate (or its complement, the specificity), i.e., describes how badly/well the test performs in people *without* the condition.

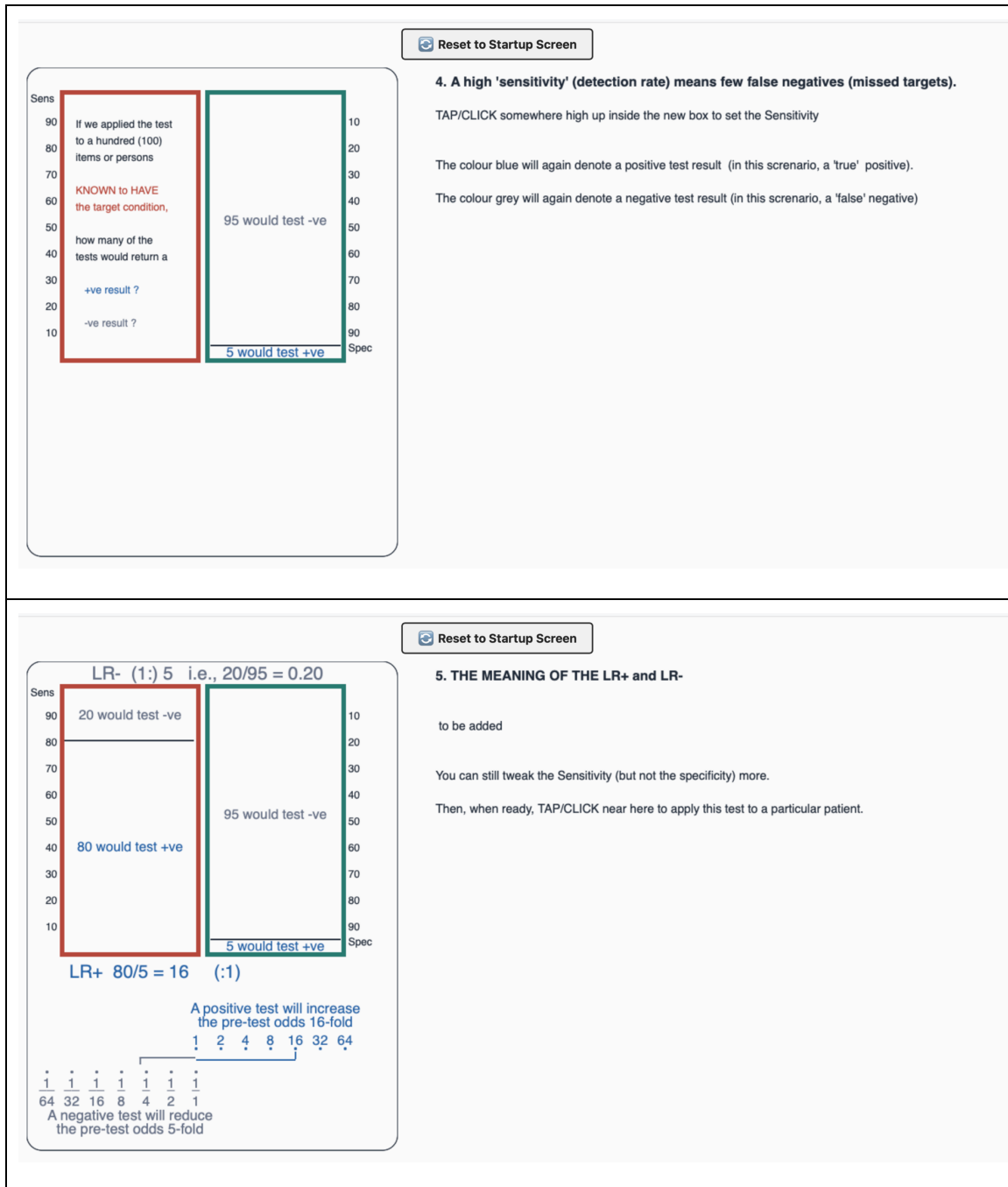


Figure 3: *Second*, the learner sets the sensitivity, i.e., describes how well the test performs in people *with* the condition. Once this 'hit rate' is selected, the LRs are calculated and displayed.

REFERENCES

- Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. *American Journal of Medicine*. 1981;71:991-998.
- Binney N, Hyde C, Bossuyt PM. On the origin of sensitivity and specificity. *Annals of Internal Medicine*. 2021;174:401-407. doi:10.7326/M20-5028.
- Fagan TJ. Nomogram for Bayes theorem. *New England Journal of Medicine*. 1975;293:257. doi:10.1056/NEJM197507312930513
- Galen, RS, Gambino. *Beyond normality: the predictive value and efficiency of medical diagnoses*. 1975. John Wiley & Son.
- Galen RS, Gambino SR. *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*. New York, NY: John Wiley & Sons; 1975.
- Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. *New England Journal of Medicine*. 1966;274:1171-1173.