2026.09.22 Comments welcomed …. james.hanley@mcgill.ca

TEACHER'S NOTES

FOR

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

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

SUMMARY OF ARTICLE DESCRIBING THE APP

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.

TABLE OF CONTENTS

Teacher's NOTES for PV-LR: an interactive learning app linking the two approaches to updating of pre-test probabilities 1

Summary OF ARTICLE DESCRIBING THE APP 1

NOTES 3

EARLY DETECTION – MASS SCREENING 3

QUANTIFYING THE PERFORMANCE OF A TEST: WHERE TO START 3

GETTING ONE'S HEAD AROUND THE TERM "SPECIFICITY" 4

A MISSED TEACHING MOMENT: AN ALTERNATIVE TO THE TERM "FALSE NEGATIVE" 4

THE ELEPHANT IN THE ROOM: LIKELIHOOD RATIOS 5

SEMANTICS: PRO-GNOSIS = PRE-DICTION? DIA-GNOSIS = POST-DICTION? 6

WHERE DO PRE-TEST PROBABILITIES COME FROM? 6

PREDICTIVE VALUES – IN WORDS AND IN PICTURES 7

A GEOMETRIC (RATHER THAN ALGEBRAIC) BRIDGE BETWEEN THE PV AND LR WAYS 8

AN INTERNET-ERA FAGAN: SAME SCALE, SAME LOGIC, FEWER RULERS 9

Words by which various writers have described a Positive Predictive Value, arranged in no particular order 10

References 11

NOTES

EARLY DETECTION – MASS SCREENING

The examples in the app are mainly from clinical medicine, where patients come one at a time to a physician with an ailment. But, of course, some of the same tests could be used for screening, i.e. for earlier detection and earlier treatment/management of an asymptomatic pre-clinical condition. Examples include pre-natal and perinatal tests, hearing and vision screening in infants and schoolchildren, screening of athletes to identify cardiovascular risks before participation in sports, and of women for neoplasia that can progress to cervical cancer. Others are screening for diabetes, tuberculosis, abdominal aortic aneurysms, and a number of cancers.

While the tests and the terminology may be the same, the prevalence of the target condition is considerably lower than in clinical settings, and raises additional concerns with the burden of following up on the positive test results, many of which will be false alarms. For these uses of the test, the scales in the app could be modified to deal with the higher test performance and the more limited prevalence range.

QUANTIFYING THE PERFORMANCE OF A TEST: WHERE TO START

Understandably, many learners are keen to start with how successful a test is in identifying those who have the target condition. Because of this, or maybe just because the term comes first in the dictionary, they want to know about its sensitivity first. Even though I was trained as a statistician, and was taught that *statistical tests start at the null*, I too was keen to follow the dictionary order when I first began to study the performance of diagnostic tests. Fortunately, this incorrect order was brought home to me several decades ago as I waited at the photocopier machine for the technician to complete his checks on the repairs he had made. When he reached into his bag, I expected he would produce several pages with different test patterns. He did not; instead, he began with a *blank* sheet of paper and checked that the copies it produced were free of any noise or artifacts! Only when satisfied with these did he proceed to the test patterns. This testing sequence is a reminder that even though we (and a best-selling book) speak of the Signal

and the Noise, we need to proceed in the more logical reverse order – just as statisticians do when testing new statistical tests.

GETTING ONE'S HEAD AROUND THE TERM "SPECIFICITY"

Even though the term specificity is the more logical one to start with, I have found it more difficult to explain than the other one of the pair. Thus, when explaining specificity to a naïve or lay audience, I now use a phrase (see panel A) that I have adapted from a commonly used oath. [Just as is the case in perjury cases, the totality requirement (the 'whole' target) is not to be understood literally. Not every omission will be a case of perjury, nor will I insist when discussing sensitivity, that the test detect every instance of the target condition]. Likewise, when discussing specificity, I don't insist that the test never yields a false positive result. Instead, I aim for almost nothing but the target. I have found that this phrase gently eases learners into accepting the word specificity: the test is seldom fooled by irrelevant conditions or by mere noise.

Of the two terms false negative and false positive, I find that learners —and the lay public at large— are much more familiar with the concept of a false positive. In their daily lives, the term false alarm comes up a lot. And so I often skip the term specificity altogether, or explain that a higher specificity just means fewer false positives. This strategy—and the example of 'benign' emails being quarantined by a spam detector—worked well when I had to explain the term specificity to the three-judge tribunal in a doping case heard by the Court of Arbitration for Sport.

A MISSED TEACHING MOMENT: AN ALTERNATIVE TO THE TERM "FALSE NEGATIVE"

Public awareness of the properties/behaviour of diagnostic tests was heightened considerably during the Covid-19 pandemic. And while epidemiologists did not speak much about the (few) false positives, they used the term 'false negatives' a lot. I am not sure how well this term was understood by the public, and it may have been a missed opportunity to stress the consequence of the result, and to use the more comprehensible term "missed cases."

Incidentally. statisticians have given them an even less self-explanatory name: Type II errors !

that, unlike specificity and sensitivity, which have a direct interpretation on their own even if we never used them to update probabilities, the only raison d'être for likelihood ratios is in doing exactly that. They are mathematical re-expressions of specificity and sensitivity designed expressly and solely to make the updating formulae more parsimonious, albeit on a scale that few other than statisticians and bookmakers use.

Perhaps, then, one way to motivate one of the LRs is to use a worked example of one of the two PV formulae, and re-arrange it to arrive at the LR version. So, let's take a scenario where the specificity is 90% (a 10% false positive rate), the sensitivity is 80%, the pre-test probability is 20%, and the test returns a positive result.

The traditional PPV calculation is

Probability[Person who tested positive has the target condition] = $\frac{0.8 \times 0.2}{0.8 \times 0.2 + 0.05 \times 0.8} = \frac{0.16}{0.16+0.04}. = 4/5.$

Now, change to that other (odds) scale, where a pre-test probability of 20% becomes an odds of 1:4 or 1/4,

i.e. instead of a probability of $\frac{\text{FOR}}{\text{FOR+AGAINST}}$, we write it as the odds $\frac{\text{FOR}}{\text{AGAINST}}$.

Thus, in this new scale, a pre-test probability of = $\frac{1}{5} = \frac{1}{1+4}$, becomes the pre-test odds $\frac{1}{4.}$

and a post-test probability of $\frac{0.80 \times 0.20}{0.80 \times 0.20 + 0.05 \times 0.80}$ becomes the post-test odds $\frac{0.80 \times 0.20}{0.05 \times 0.80}$.

Now, split the post-test odds $\frac{0.80 \times 0.20}{0.05 \times 0.80}$ into its left-hand-side and right-hand side parts, ie., $\frac{0.80}{0.05} \times \frac{0.20}{0.80}$.

Then revert to names instead of numbers, to get

post-test ODDS = $\frac{\text{Sens}}{1-\text{Spec}} \times \frac{0.20}{0.80}$ = LR+ × pre-test ODDS.

Thus, operationally, we can think of the LR+ as the function of the two test characteristics that, by a single multiplication, turns the pre-test odds into the post-test odds. Unlike the scale used in the PPV formula, the odds scale enables an updating formula that cleanly separates the test and the patient.

Of course, unless the learner/clinician is comfortable with a post-test odds of $16 \times \frac{1}{4} = 4$, this 4:1 needs to be converted back to a post-test probability of $\frac{4}{4+1} = 80\%$.

SEMANTICS: PRO-GNOSIS = PRE-DICTION? DIA-GNOSIS = POST-DICTION?

In a mini-dictionary of epidemiological concepts and terms, its single author noted that "the [predictive-value] term is seriously malformed. At issue is not prediction (of the future) but knowing about the present." We might argue that he was being overly semantic. Another observation of his was more useful: *pro*-gnostic probabilities rely on determinants of the risk of a *future* event, and so it is logical, when fitting risk *pre*-diction models, to put these X's on the right hand side of a regression equation, and the risk (probability, E[Y|X]) on the left hand side. On the other hand, the indicators used in a diagnostic probability model are *manifestations of*, and were produced by, the present illness (or non-illness). Thus, even if the same-direction X to Y model is used for both, the arrows in the directed acyclic graphs for *pro*-gnostic probabilities should point from X to Y, and from X to Y in the *dia*-gnostic (*post*-dictive) model.

WHERE DO PRE-TEST PROBABILITIES COME FROM?

The earlier applications of the formulae were in mass screening, and so the estimated *prevalence* in the population to be screened was used to calculate the expected prevalences in the two sub-populations produced by the test results. In a clinical context— the primary focus of this note—there will be different 'profiles,' and so the prevalence in the screened population gets to be replaced by a profile-specific pre-test probability (one could think of it as the theoretical prevalence in patients with the same profile as the patient seeking care).

In many medical school curricula, the topic of diagnostic tests and diagnostic probabilities is covered quite early on, before students have encountered patients. At this starting out stage, they are just *given* the three inputs to the updating formulae. Some ask: if sensitivity and specificity come from elsewhere, where do pretest probabilities come from? I like Sackett's answer "Usually they are derived from your own accumulating clinical experience, specific for the setting in which you work and the sorts of patients you see. As a result, pretest probabilities for the same target disorder can vary widely between and within countries and between primary, secondary and tertiary care."

Increasingly, learners and starting out clinicians can learn them from medical calculators that, like the Framingham score for pro-gnostic probabilities or risks, compute scores that are then converted into dia-gnostic probabilities.

PREDICTIVE VALUES – IN WORDS AND IN PICTURES

The terms *positive predictive value* and *negative predictive value*, and the accompanying formulae, were introduced some sixty years ago, notably in a mass-screening context [ref]. In a footnote, the author apologized for being forced to use the term *predictive value*, noting that " 'predictiveness' would be a better term, but is, unfortunately, nor recognized in Webster's Unabridged Dictionary.''

Appendix 2 lists the *phrases* by which various writers since then have described a positive predictive value. They vary considerably in clarity and sharpness, and some even run the risk of being confused with sensitivity. I prefer ones that start by explicitly describing the persons the probability statement applies to (the 'who') and then describes the 'what'. In other words, I start with what the mathematical statisticians put *after* the vertical bar, i.e., after the "|" or the "given that" which tells the reader the context they are 'conditioning on' or restricting themselves to. To reinforce this, I ask learners to imagine a physician starting his/her day by reviewing the results of the tests he had ordered yesterday. For simplicity, I limit the discussion to tests that yield a binary result, so that the reports can be sorted into just two piles, the ones where the test results were positive and negative respectively. The physician deals with the positive results first, and calls each of these patients to discuss the next step. Before talking to the patient, he uses the positive test result to update the pre-test-probability he arrived at yesterday. His next step will of course depend on the context: a 35% post-abnormal-mammogram probability of breast cancer will not have the same 'proceed directly to treatment' as a 35% post-positive-test probability of appendicitis or strep throat or urinary infection or multiple sclerosis might have.

Pictures are a common way to convey the probabilities involved in interpreting the results of diagnostic tests. As computer-generated infographics have become easy to make, simple rectangles or squares divided into four shaded areas have given way to icon arrays (pictographs, pictogram charts, and 'people graphs'). In the app, we have stayed

with shaded rectangles, particularly since the area of each is a product of a quantity to do with the *test* (the height) and am unrelated quantity to do with the *patient* (the width).

The difference between general-population prevalence and a patient's clinical pre-test probability is easy to overlook, catching even seasoned medical editors off guard. A classic case of a sub-editing blunder occurred in a BMJ teaching article on this topic [Loong, 2003] that I would have liked to show to the judges in the doping case. The author's text used several pictures to explain why, among other things, a clinical patient's probability would be much higher than the general population's. Unfortunately, a well-meaning layout editor completely missed the point. In a rush to create a punchy hook, they added an uncredited promotional blurb asking: "Can you explain why a test with 95% sensitivity might mean that only 1% of those who test positive are actually affected? The visual approach in this article should make the reason clearer"— awkwardly inverting the definition of sensitivity and stumbling into the very trap the author was warning against. The BMJ never did admit the editorial-insert-error, but this major gaffe is a reminder of why we explicit visual tools can help us keep our logic straight.

The 2013 article in the Economist, which I have routinely used when teaching this topic, makes very effective use of pictures to highlight the limitations of tests in a low-prevalence context, albeit at a research-study rather than individual-patient, level.

A GEOMETRIC (RATHER THAN ALGEBRAIC) BRIDGE BETWEEN THE PV AND LR WAYS

Some learners prefer geometry over algebra. To appreciate how it can hep here, consider patients whose tests that return a positive result. The multiplicative relation between the pre- and post-test odds can be derived directly by looking at the relative areas of the rectangles representing the numbers of positive tests: *T* and *F* when the testing the test in persons known to have/not have the target condition; *t* and *f* when the target is *latent*, and the test result is the '*known*.'. In the former, for simplicity, the bases of the two enclosing rectangles are in the ratio 50:50, so the *T* area is 80 (height) × 50 (base) and the *F* area is 5 (height) × 50 (base), So, because the two 50s cancel, their ratio is just $\frac{80}{5}$.

But what happens when we are in a real-life 20:80 mix of target present:absent? The heights of the rectangles containing *t* and *f* are the same as they were in the 'testing the

test' situation, but their bases have altered from 50:50 to 20:80. So, to get the *t:f* ratio, we need to multiply the *T:F* ratio by 20 and divide it by 80. This gives a *t:f* ratio of (T:F) × (20:80). Or, writing an A:B ratio as $\frac{A}{B}$, we get

$$\text{Post-positive-test odds} = \frac{t}{f} = \frac{T}{F} \times \frac{20}{80} = LR_+ \times \text{Pre-test odds.}$$

This way of deriving the post-positive test odds also gives a more concrete meaning to the LR_+. It is the post-positive-test odds when the pretest odds are 50:50 !

AN INTERNET-ERA FAGAN: SAME SCALE, SAME LOGIC, FEWER RULERS

The more mathematically observant/curious will have noticed that the spacing of the probability values is the same as in the two probability rulers in Fagan's nomogram. Fagan did not have space to explain how he constructed it, why he used that scale, why he had to use two probability rulers, or why in one the probabilities start at the bottom and in the other at the top. We explain the spacing of the probability scale, and how a 'virtual slide rule' avoids the constraints of Fagan's 'fixed on paper' choice of medium.

The central idea in both the paper and virtual platforms goes back to the law stating that the log of a product is the sum of the individual logs; in our case, the product is that of the pre-est odds and the relevant LR. Most applications of this law to diagnostic quantities use the natural log scale, but, as can be seen in the app, logs to the base 2 are far more friendly. If the odds shown on the ruler are expressed as powers of two, but plotted on a log-to-the-base-2 scale, then odds of ¼, ½, 1, 2, 4, 8, etc. are equally spaced. We can do the same with the LR values, starting at 1, and doubling them at each step (2, 4, 8, 16 etc) or halving them at each step (½, ¼, , …). Then, when (say) the LR+ is 16 = 2^4, the post-test odds is located 4 'notches' to the right of the pre-test odds. And similarly an LR- of say 1/8 = 2^{-3} puts the post-test odds 3 notches to the left of the pre-test-odds. To respect the law, we maintained the same distance (in pixels) between successive notches for the LRs in the left panel and the odds shown in the probability ruler. This consistency allows the '4 to the right and 3 to the left' shifts to apply right across the ruler, no matter what the starting odds is. With this ensured, the only remaining ruler-drawing task was to put in the (unequally spaced) notches for the probabilities.

WORDS BY WHICH VARIOUS WRITERS HAVE DESCRIBED A POSITIVE PREDICTIVE VALUE, ARRANGED IN NO PARTICULAR ORDER

The percentage of times that a positive test will detect a diseased individual

If you are given an abnormal result for a test, what are the chances that the abnormal result indicates disease?

The percentage of positive results that are true positives when the test is applied to a population containing both healthy and diseased subjects.

The probability of disease in a patient with an abnormal test result

The probability that his patient has the disease of interest on the basis a positive test result

Probability of disease when the test is positive

The probability of having the disease if the test result is positive

The percentage probability that a patient who has tested positive has the disease

Concerning a positive result of a diagnostic test (in clinical medicine), the probability that this implies for the presence of the illness at issue

Probability that a patient has a particular disease if the patient has a positive test.

The proportion of positive results that are true positive results

The probability that a person who tests positive for a disease actually has the disease.

The presence of the target disorder among positives.

The probability that a patient has a particular disease if the patient has a positive test.

The probability that a patient has the disease after a positive test

REFERENCES

Berwick DM. Fineberg HV. Weinstein MC. When doctors meet numbers. American Journal of Medicine. 1981 Vol 91 pages 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's Theorem (letter). N Engl J Med 1975;293:257. DOI: 10.1056/NEJM197507312930513

Galen, RS, Gambino. Beyond normality: the predictive value and efficiency of medical diagnoses. 1975. John Wiley & Son.

Loong T-W. Understanding sensitivity and specificity with the right side of the brain BMJ 2003;327:716–9 https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/RightSideLeftSide.pdf

The Economist. Problems with scientific research: how science goes wrong. scientific research has changed the world. Now it needs to change itself. Oct 19, 2013 https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/HowScienceGoesWrong.pdf

Vecchio, TJ. Predictive value of a single diagnostic test in unselected populations. New Engl. J. Med. 274, 1171-1173, 1966.