July 18, 2025

Manuscript: *The structure/logic of the Likelihood-Ratio-based formulae used to teach post-test probabilities: derived/explained in words and pictures, rather than by algebra*

Nomogram: *A Shiny modern-age version of Fagan's 1975 classic*

For many years, I was a tutor in the six or so small-group (12-15 students) sessions used in the teaching of epidemiology and statistics to the McGill medical students.[1]. When I started participating, they took the course in the Fall term of their second year (the same term when they get their 'white coats' and had their first encounters with patients on the wards); at some point, the course was moved into the Fall term of their first year, making it even more challenging to teach.[2]

Unlike our McMaster competitor, which stressed 'Clinical' Epidemiology (and wrote the textbook on the topic), our curriculum has focused more on 'Classical' Epidemiology, and starts off with topics such as 'case control' and 'cohort' studies, and confounding. It seemed to be trying to turn the students into 'mini-epidemiologists'. I continued to advocate for a course that starts with the concepts and skills the students would need when they met their first patient, not their first report of an etiologic study. My son did the McGill MDCM program in 2007-2011, and so I had an inside source on what concepts and skills were perceived as most and least relevant.

I was happy when 'Interpreting diagnostic tests' became the first of the six topics addressed in the course, and I always began the corresponding small-group session with a question posed to me by an eminent British pediatrician I met while on sabbatical in Melbourne: what percent of the diagnoses we make are correct? I was surprised at his answer (less than 50%) but not by his statement that 'good medicine starts with getting the diagnosis right'.

Early on, the use of the long formula[3] to calculate positive and negative predictive values (or as I prefer to call them, post-test probabilities) gave the students (and even some tutors) trouble.[4] And so, it was replaced by the (McMaster-recommended) one where sensitivity and specificity are rolled into one Likelihood Ratio, and the separation of these characteristics of the test from the characteristics of the setting (population) is more evident. Also, one could use the Fagan nomogram to do the computations, and instead focus on 'what changes if... '.

The downside, however, was that students who not satisfied with using a black-box formula/tool might ask for its derivation. Most tutors, uncomfortable with the question, merely answered that 'it is based on Bayes Theorem' but a few asked me to help them with the algebra in case a student wanted to know more details.

[1] I taught the plenary sessions in 2002

[2] I regarded this move as further evidence that, in our medical school at least, these topics are becoming less valued by those who design the curriculum, and that our department is indeed seen as 'fringe medicine'.

[3] ppv = $\frac{prevalence \times sensitivity}{prevalence \times sensitivity \ + \ (1-prevalence) \times (1-specificity)}$

[4] They are not alone. See When Doctors Meet Numbers.

Below is a draft of the explanation I produced, along with some more general comments on terminology, and its history. Unfortunately, I haven't been able to convince any of these tutors to join me in getting this into print. So, if you would like to help me revitalize it (or better still, take the lead in doing so) I would be delighted to hear from you! (The earlier version has the tutorial material that prompted me to try to make it clearer).

Here also is a link to my notes on probability, taken from my course to graduate students in biostatistics, on the interpretation of diagnostic tests. They are preceded by material from Clayton and Hills' textbook. Page 8 of my notes has some material (links) on Diagnostic and Screening Tests and Page 12 reproduces Fagan's NEJM letter that introduced his nomogram.[5] The Page 8 inset begins with a link to my Shiny app that has a 'rotated left 90 degrees' version of Fagan's nomogram. The R code itself is provided here.

Readers might also be interested in other Supplementary Exercises.[6]

Sincerely,

James Hanley

webpage: https://jhanley.biostat.mcgill.ca | email: james.hanley@mcgill.ca

[5] It was interesting to me to see how, despite its 'right-to-left' pre-to-post direction, this nomogram kept being reproduced by others, in same cases with the labels edited and retyped. At one point, I wanted the pre-to-post direction to be depicted as a more Cartesian 'left-to-right' direction, and set about doing both the calculations and graphics in R. That's when I found out how tricky the calculations for the placements of the scales were!

I subsequently set the production, from scratch, of a left-to-right Fagan nomogram, as exercise 23 in these Notes on the binomial distribution. [You may have noticed that my chapter headings tend to avoid the technical terms, and instead spoke of the parameter (here a proportion) being pursued]. The exercise includes my own attempt.

[6] such as 2.15: What's the value of a confirmatory PCR test? or Supplementary Exercise 2.1 ('Efron's twins story') which, I told the students, "can be tackled in many ways. Efron uses the odds scale to go from 'pre-' to 'post'-test odds, and then switches back to the probability scale. We do the same when teaching medical students about diagnostic tests. Fortunately, today, with readily accessible apps, there is less emphasis on the calculation, and more on the probabilities themselves. A few pages further in the notes, you can will see what (paper) 'apps' were like in 1975! Fagan's nomogram is still a clever tool, and JH has used it as a starting point for a shiny app cited on the coloured box on the right hand side of page 8 of his Notes. This box gives you links to the 'terminology' for the errors/performance of medical diagnostic tests (If JH had his way, we would never have invented the terms sensitivity and specificity) and the correspondences with statistical tests."

Partial draft 2015.09.15.. comments welcome

Target:

BMJ (they had the right brain article) or Medical Education or J Clin Epi (Sackett legacy)

The structure/logic of the LR-based formulae used to teach post-test probabilities: derived/explained in words and pictures, rather than by algebra.

Summary

The LikelihoodRatio-based formula that converts a pre-test probability, and the sensitivity and the specificity of a diagnostic test, into a post-test probability is increasingly taught, but its algebraic structure is a bit of a black box, and usually skipped over in teaching. We peer behind this formula and use words and pictures to make the algebra more obvious. We also comment on terminology and suggest …

Introduction

The structure of the classical formula that converts a pre-test probability, and the sensitivity and the specificity of a diagnostic test, into a post-test probability is fairly easy to teach. The arithmetic is somewhat tedious, involving two separate multiplications, one addition and one division. Thus, as Berwick et al. documented several decades ago, the formula is not easily remembered or applied quickly. Probability trees help people understand the steps involved, as do different colored dots in a rectangular grid (BMJ right brain article, The Economist article on why Science got it wrong). Once its structure has been firmly understood, calculations carried out by an app do not seem so 'black box.'

In the ensuing decades, more teachers have favoured the formula based on a likelihood ratio, since it cleanly separates the qualities of the test from the setting in which it is applied. Manually, it involves one subtraction and one division to obtain an odds, another subtraction and a division to get the LR, the multiplication of the odds and the LR, and a further addition and division to turn the resulting post-test odds back to a post-test probability. But more important than the longer arithmetic, it rests on less familiar concepts of an odds and a LR. Thus the nomograms, and now the apps, that are widely used to do the calculations, are much more black-box.

The less-algebraicly-transparent structure of this formula poses a challenge for many medical school teachers. They tend to start by presenting elements of the classical formula, using a probability tree, or a box divided two ways, or directly using a 2x2 table with numbers or symbols. If they do present the LR-based approach, they tend to focus on defining the LR, but often merely as a formula involving sensitivity and specificity, rather than in words. They then explain that it makes sense that a ratio above/below 1 should increase/decrease 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. Questions from curious first year medical students as why the formula has the form it has are usually deflected, or answered by saying it stems from Bayes' Theorem.

The objective of this note is to argue that the formula does not have to be so much of a black box, and that the algebra involved in deriving it can actually be made much more intuitive using a graph. Before doing so, we comment on our experiences and examples in teaching these concepts not just to medical students, but to the 3-lawyer panel of CAS judges in the WADA case.

Preliminary Comments on Terminology

Most teachers begin with sensitivity, but we like to emphasize that in developing a diagnostic test, the first task is to agree on a specificity value by setting a cutoff for 'positivity'. If the measurement behind the test is inherently quantitative (e.g., a concentration), this is done by assembling a reference distribution of values from similar-age-sex persons without the target condition, and setting the threshold for 'normal' so that among such persons, the false positive rate (FPR) would be say 5% or 1% (we will use 5% in our example). Until this is set, or there is agreement about what would constitute an 'abnormal' test result (in an image say), one cannot begin to talk about sensitivity.

Lay people understand the concept of setting a threshold, and can turn up of down the threshold for positivity in the spam detector in one's email software, or the motion sensor in a burglar alarm. They also are quite comfortable with the term 'false alarm' and could easily be taught that a false alarm rate of 5% should be taken to mean that of 100 genuine (innocent) emails, 5 would be flagged and put in the spam folder. Clarity on the denominator for a FP rate (FPR) is critical. In the psychophysical experiments with radar in the 1950s, the abbreviation FPR and the term False Alarm Rate were widely and correctly used. Slightly earlier, the statistical community, with a blandness only matched by their economic cousins, gave it a name "The type I Error Rate". To be fair, there was one good reason why they called this one the type "I" error, and the other one (to come) the type "II" error, rather than the other way round: one has to start with the null, or the innocent, or the healthy, and set the cutoff there, before pursuing the state/condition/disease of concern.

The first person (a PhD statistician in 1947[1]) we know who used the terms sensitivity and specificity defined the latter as "*A measure of specificity* or the probability

[1] From Field Studies Section, Tuberculosis Control Division. This paper and the discussion by Professor Neyman, which follows it, were presented before the Institute of Mathematical Statistics, at the twenty-

of correct diagnosis of 'negative'[2] cases." Unfortunately for medical terminology, he chose to focus on the 95% 'True Negative" rate, rather than the 5% False Positive rate, (McNeil NEJM 1975 says FP 'ratio' ??) and to gave it a new name, one that is not that immediately self-evident. Presumably, he chose the word 'specific' because a diagnostic test that targets TB should not be 'fooled' by background noise, and would find TB and only TB (today that in the pursuit of lung cancer, readers of lung CT images that show shadows and TB scars would not interpret them as lung cancer: the ideal system would identify lung cancer, and only lung cancer.

Below, we will use the FPR term whenever we can, and avoid its complement, specificity. But we have no illusions about going back and inventing or adopting a more meaningful term than specificity. It is too late to undo the damage.

Now that a FP rate has been adopted, we can address how frequently the detector detects the target condition. Here again, we think the lay public could easily understand, if we choose our terms well.[3]. The 'hit' rate (to borrow from the radar research[4]) or the success rate or the detection rate will depend on how hidden and perverse and subtle the targets are, but let us say that at the selected FPR, the spam detector 'catches' or 'traps' or 'identifies' 80% of the malicious emails: the 'detection rate' is 80%.

Again, it is unfortunate that the world adopted the term 'sensitivity' for this 80%, given that it was bound to be misunderstood. Many medical students blame the false alarms on an 'overly-sensitive' system that identifies some non-targets, even as it misses (is insensitive to) some of the targets. Here we could have easily stayed with the

eighth annual meeting of the Pacific Division of the American Association for the Advancement of Science, in San Diego, Calif., June 18, 1947,while the author was visiting professor of biostatistics at the School of Public Health, University of California. Professor Newman, one of those we have to thank for the terms Type I and II error, responded to Yerushalmy's analyses, but used the Greek letter pi

[2] Yerushalmy thinks of instances of the presence and absence of the target condition as 'positive' and 'negative' cases; one needs to carefully distinguish this usage from whether the test is positive or negative.

[3] They also know very well that there is no free lunch: if one wants to detect a greater percentage of the malicious emails, one could turn down the criterion, and vice versa, but let's say we agree that a 5% is a good compromise.

[4] It is interesting how usage has changed, and varies from context. The 'hits' produced by Google searches contain both 'false' and 'true' items. The airforce speaks of attempted air strikes, and hits and misses, and sometimes the misses do damage of their own. The radar people also spoke of a miss rate rather than a hit rate.

percentages of targets hit and missed, and adopted the term hit rate or better still (to avoid confusion with the Google meaning) detection rate. But we went instead with Yerushalmy's "*measure of sensitivity*: the probability of correct diagnosis of 'positive' cases." We find it to be the easier of the two to fit to the situation it is meant to describe, but below will again try to avoid it whenever there is a chance to use a more natural term.

From pre- to post-test probability

We now switch directions and come to the 'reverse probability' that is the focus of our note, and that goes under the term "positive predictive value". In a medical diagnostic setting it could more usefully go under its other name, post-test probability. Again, the concept is easy to explain to lay people: some of the examples we like to use are the probability that an email that has been consigned to the spam folder is in fact spam, that a mammogram-prompted biopsy will come back as breast cancer, or that an alarm the fire/police department responds to does in fact involve a genuine fire/burglary. This last one even suggests its form: it is the number of genuine(g) alarms as a fraction/percentage of the sum of the genuine(g) and the false (f) alarms.

We have had trouble locating the first use of the term "positive predictive value" in the diagnostic probability framework. With the help of Google Ngram, we were able to locate instances of the phrase back as far as the 1920s, but in a looser context where psychologists would interpret a predictor "x" that had a positive correlation coefficient with a measured in the future "y" as "having positive predictive value." We found instances where the $g/(g+f)$ fraction or percentage was used in a medical diagnostic probability context only as far back as the late 1960s. Although some of the 'ppv' entries in the Ngram do not all refer to medical diagnosis, it is still disheartening that this much less descriptive term is far more popular than 'post-test probability'.

Before we move on to Likelihood Ratios (one of the two ways of getting from a pre-test probability to a post-test one) we want to clear up a terminology issue in an otherwise excellent and high profile 1979 publication that showed (using the other method) how to make the pre- post probability jump. Unfortunately, throughout their article, Diamond and Forrester referred to what we would today call pre-test and post-test probabilities as pre-test and post-test likelihoods. They were probabilities, expressed as percentages. Neither likelihoods nor their ratios were used.

The usual way from pre- to post-test probability

The top left panel of the Figure was constructed to represent a 50% pre-test probability of being in the darker (target) area. A test with a 5% False Positive Rate will divide the left (lighter) half, producing the upper rectangle ('false alarm' area) comprising 5% of the 50%, or 2.5% of the whole. If it has an 80% detection rate it will divide the right (darker) half, producing an upper rectangle ('genuine' area) that comprises 80% of the other 50%, or 40% of the whole. So in all, some 42.5% (all those inside the black border) will test positive. Of these 42.5, some 40/42.5 or 16/17 will be genuine, so the post test probability is 94%.

This arithmetic is an example of the general post-test probability or PPV formula

$$\frac{\text{PretestProbability} \times \text{Sensitivity}}{(1-\text{PretestProbability}) \times (1-\text{Specificity}) + \text{PretestProbability} \times \text{Sensitivity}}$$

Diamond and Forrester used this formula, as do most modern teachers today. If they wish to illustrate the effect of a different pre-test probability, the entire sequence of steps must be repeated. With the 25% (rather than 50%) pre-test probability shown in the panel on the right, the calculation comes out to (80% of 25%) / (80% of 25% + 5% of 75%) = 20/23.75 or 84%, a step down (from 94%) that is not easily anticipated.

The LR way from pre- to post-test probability

The Likelihood Ratio associated with a positive test (LR+) is often simply defined as "Sensitivity divided by (1 minus Specificity)" or Sensitivity / (1 - Specificity). Sometimes it is stated in words as 'the probability that a person with the target condition will test positive divided by the probability that a person without the target condition will test positive.' [5] In our example, it is 80/5 or 16. In order to apply it, one must first convert

[5] The term comes from the Likelihood Ratio statistic used by Newman and Pearson to compute how much more probable the data (here the test results) are under the alternative

the pre test probability [50% in the left panel, 25% in the right] to the pre test odds [1 in the left, 1/3 in the right], using the definition odds = probability/(1-probability). The LR is multiplied by this pre test odds to produce the post test odds (16 in the left, 16/3 in the right) . In the last step , by reversing the probability to odds equation, one obtains the post test probability as PostTestOdds / (PostTestOdds +1), or 16/17 = 94% in the left and (16/3)/(16/3+1) = 84 on the right.

The Fagan nomogram takes shortcuts through the arithmetic, but in so doing it also hides the logic. But does this have to be? The first conceptual clarification comes from simply referring to '1 minus the specificity' as the False Positive Rate. In both our examples it is 0.05 or 5%. And if we replace the term Sensitivity by the equivalent True Positive Rate (here 80%), we can think of the LR as '16 times whatever the FPR is.' In the leftmost example, where there are just as many with as without the target condition, the TP:FP ratio is also 16:1 (there are 16 TPs for every 1 FP). In the rightmost example, however, there are 1/3 many with as without the target condition, so the TP:FP ratio is no longer 16:1, but 16/3 TPs for every 1 FP. In each example, just like above, it merely remains to convert the 16:1 odds to 16/17 prob (if 50% pretest probability) or (16/3):1 odds to a (16/19) probability (if 50% pretest probability).

And of course' if test is any good, ratio is >> 1 ' -- usual teacher deflection)

NPV

Same example: it should now be obvious what to do, and to derive the formula from scratch using the diagram ..

Also would like to introduce the NPV in case of haemophilia carriers (women) and using each son they have as a test with perfect specificity but 50% sensitivity. So LR+ is infinite… it's a great 'sorting people' example… and LR- works.. its ½ each time a son is born

Ending Remarks

No problem if use nomogram to do it, but at least now teachers know what the LR x pre-odds is and why the LR comes into it. And don't need to invoke any abstruse Bayes Theorem .. it's all grade 6 math and pictures UNDER CONSTRUCTION

hypothesis (the target condition) than under the null (its complement). In model selection contexts, it is called the Bayes Factor.

More to be filled in and expanded, but that is the gist of it

Jh 2015.09.15

Refs

Problems with scientific research: HOW SCIENCE GOES WRONG Scientific research has changed the world. Now it needs to change itself. The Economist, Oct 19th 2013

Loong T-W. Understanding sensitivity and specificity with the right side of the brain BMJ 2003;327:716–9

Berwick DM. Fineberg HV. Weinstein MC. When doctors meet numbers. American Journal of Medicine. (1981) Vol 91 pages 991-998.

Jacob Yerushalmy. Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques : Public Health Reports Vol. 62, No. 40, Tuberculosis Control Issue No. 20 (Oct. 3, 1947), pp. 1432-1449

McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med. 1975 Jul 31;293(5):211-5.

Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med. 1979 Jun 14;300(24):1350-8.

Tutors' Noted for last week's TB Dx small group

(I must also get Marina Klein's lecture notes on how she dealt with PPV

LR etc

1. To understand how to evaluate new diagnostic test
2. To be able to calculate and interpret the characteristics of diagnostic tests including the sensitivity, specificity, PPV, NPV, likelihood ratios
3. Understand the impact of prevalence on predictive values
4. To apply likelihood ratios in clinical decision making

What condition(s) is the diagnostic test being developed to detect?

What is the gold standard against which the test is being compared?

describe in plain English what the following test parameters mean:

a. Sensitivity
b. Specificity
c. Positive predictive value
d. Negative predictive value
e. Likelihood ratio of a positive test
f. Likelihood ratio of a negative test

Eg. Plain English description of
Sensitivity: *The true positive rate; proportion of patients with disease who test positive;*
Specificity: *The true negative rate; proportion of patients who don't have the disease who test negative*
PPV: *Proportion of patients with positive tests who have disease etc.* **Likelihood ratios:** *Of a positive test. The LR of a positive test tells us how well a positive test result does by comparing its performance when the disease is present compared with when it is absent. Represent the probability of disease after a positive or negative test.*
The best test to use for **ruling in** *a disease is the one with the largest likelihood ratio of a positive test. _LR>10 is most helpful at ruling in a disease*

In this instance, the LR for a positive test result means that when a smear is positive, a positive test result would be 120 times more likely to be seen in someone with, as opposed to someone without, TB.

For the Negative likelihood ratio, it is easier to express this probability as 1/LR-. For example here 1/0.02=50 so could say a Negative test is 50x more likely in someone without than someone with TB.

a. Specificity
b. Positive predictive value
c. Negative predictive value
d. Likelihood ratio of a positive test
e. Likelihood ratio of a negative test

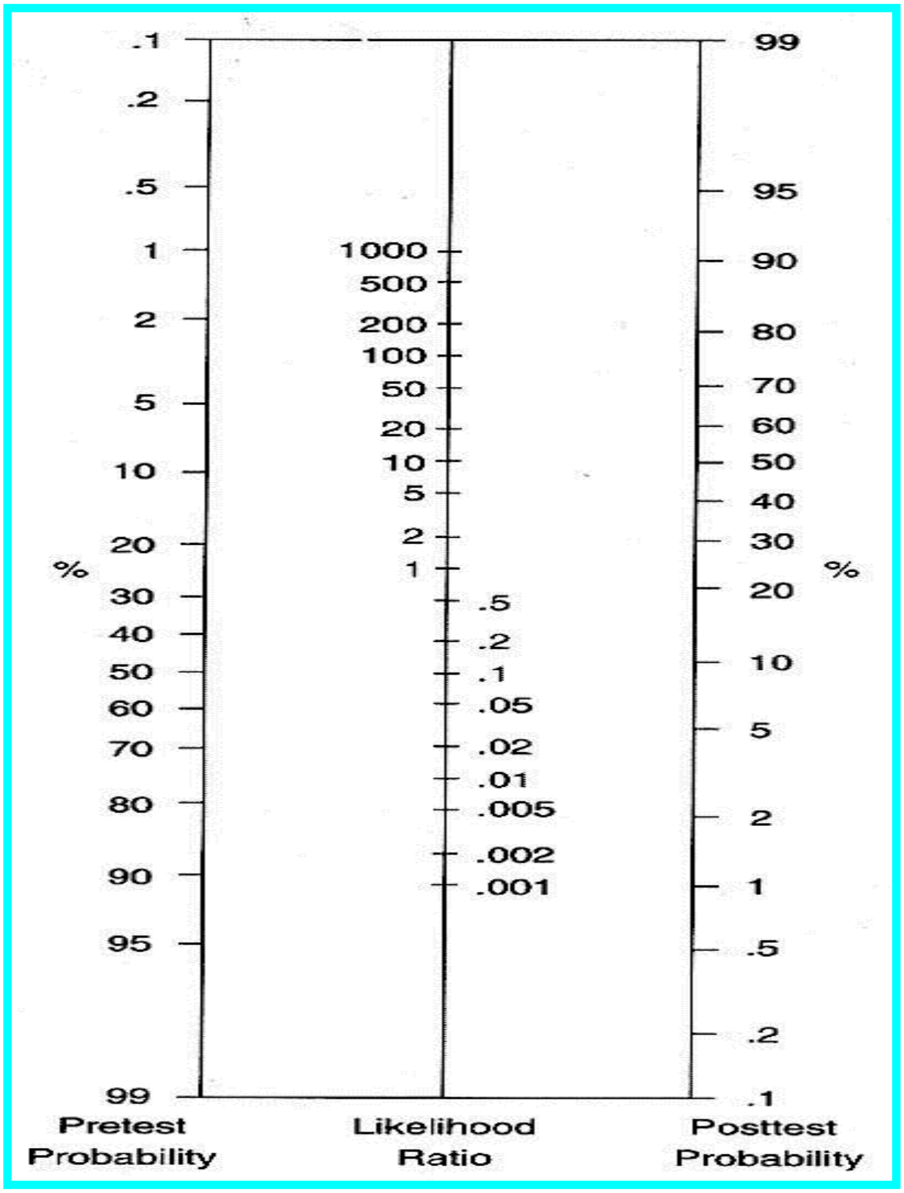
Note the reduction in sensitivity and increased in LR- meaning that we miss more cases and so a negative test is less helpful at **ruling out** disease. In other words they tried harder to find TB and now three times the test is negative making a negative test much more likely rule out the presence of TB.

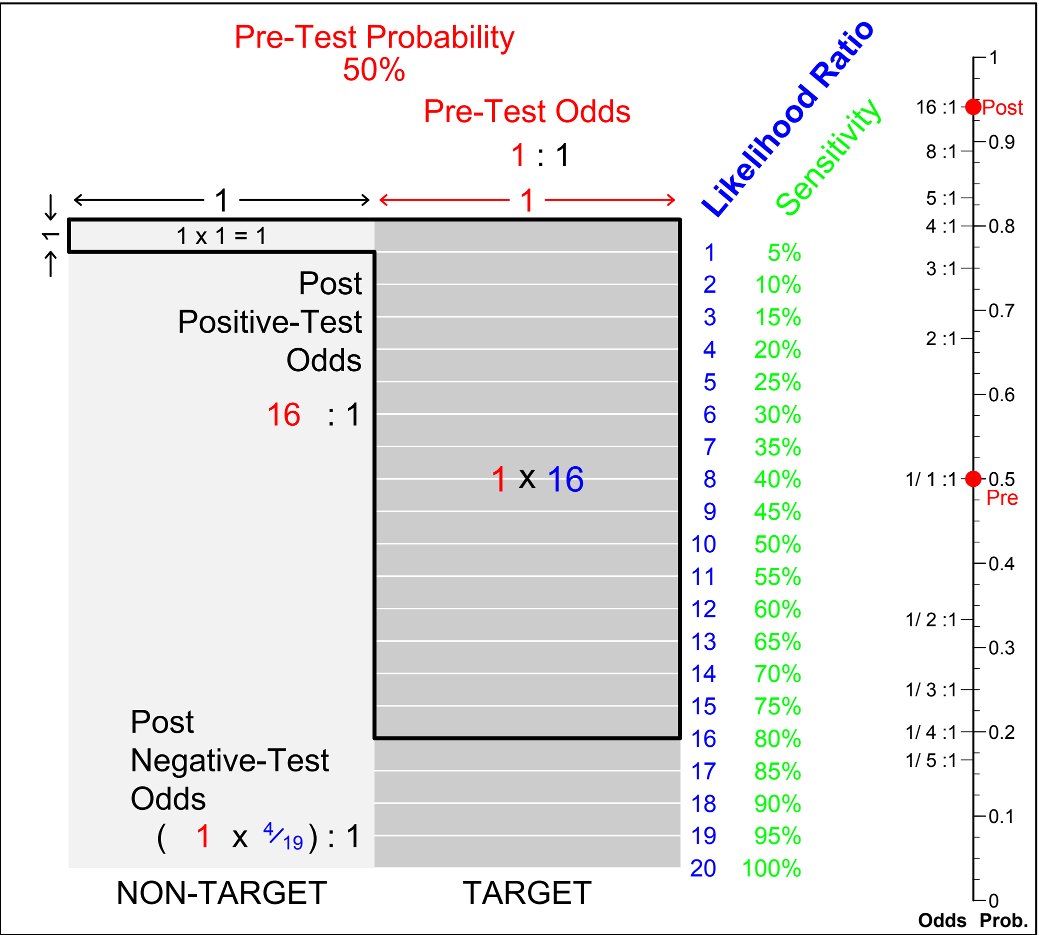
1. Did this affect the sensitivity and specificity of the test at the different sites? Why or why not? **Look at Table 2**

2. What do you think would happen to the positive predictive value if they had evaluated the test in a population where the prevalence of TB is much lower (e.g. in a Canadian hospital where prevalence would be far less than 1%)?
PPV would drop substantially (see table below from lecture).

3. Apply the above results to the following clinical scenarios.

Fagan nomogram from 1975 NEJM letter -- in response to an uglier NEJM one in 1974


.1
.2
.5
1
2
5
10
20
%
30
40
50
60
70
80
90
95
99
Pretest Probability
1000
500
200
100
50
20
10
5
2
1
.5
.2
.1
.05
.02
.01
.005
.002
.001
Likelihood Ratio
99
95
90
80
70
60
50
40
30
%
20
10
5
2
1
.5
.2
.1
Posttest Probability



Pre-Test Probability
50%

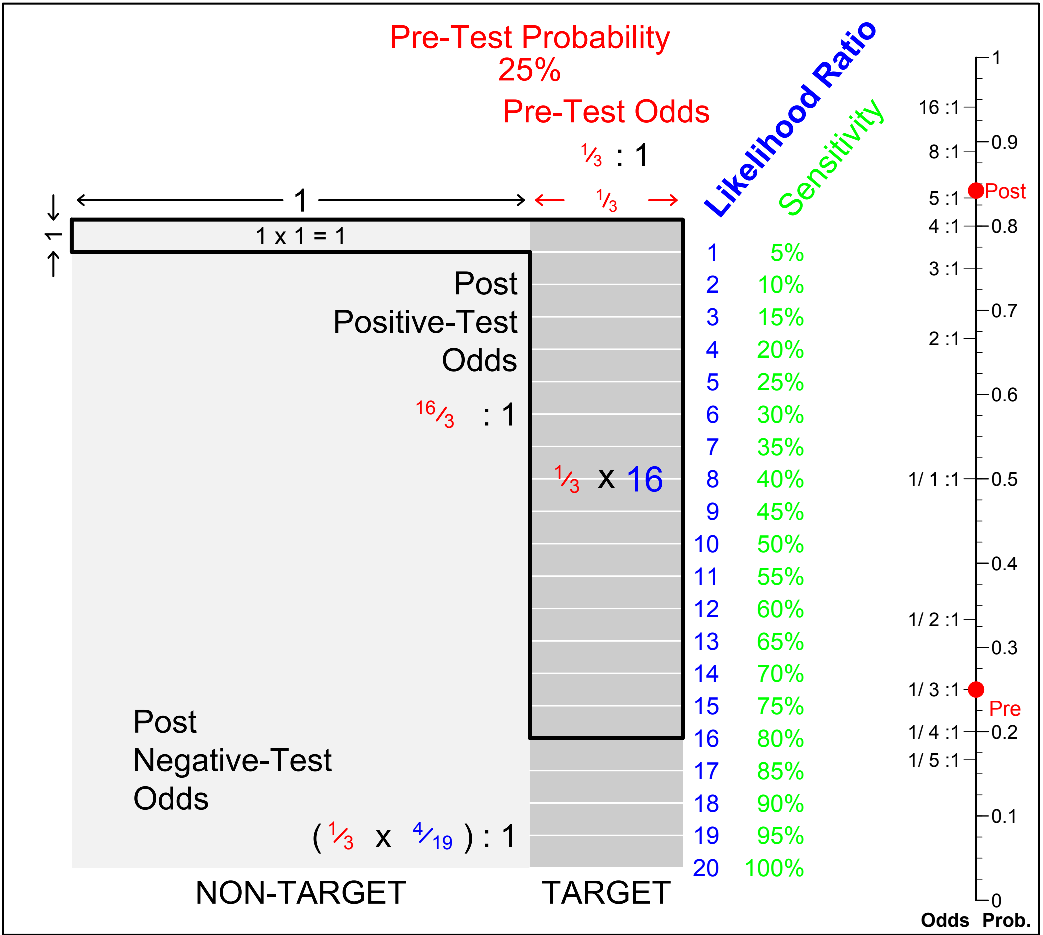
Pre-Test Odds

1 : 1

Likelihood Ratio

Sensitivity

Likelihood Ratio	Sensitivity
1	5%
2	10%
3	15%
4	20%
5	25%
6	30%
7	35%
8	40%
9	45%
10	50%
11	55%
12	60%
13	65%
14	70%
15	75%
16	80%
17	85%
18	90%
19	95%
20	100%



Pre-Test Probability
25%

Pre-Test Odds

$^1/_3$: 1

Likelihood Ratio

Sensitivity

Likelihood Ratio	Sensitivity
1	5%
2	10%
3	15%
4	20%
5	25%
6	30%
7	35%
8	40%
9	45%
10	50%
11	55%
12	60%
13	65%
14	70%
15	75%
16	80%
17	85%
18	90%
19	95%
20	100%

Draft 2019.09.05 Comments welcomed …. james.hanley@mcgill.ca

Target: BMJ (they published the right brain article) or Medical Education or Academic Med or J Clin Epi (Sackett legacy)

The structure/logic of the Likelihood-Ratio-based formulae used to teach post-test probabilities: derived/explained in words and pictures, rather than by algebra.

Summary

The Likelihood-Ratio-based formulae that converts a pre-test probability, and the sensitivity and the specificity of a diagnostic test, into a post-test probability are increasingly taught, but their algebraic structure is somewhat of a black box, and usually skipped over in teaching. We peer behind these formula and use words and pictures to make the algebra more obvious. We also comment on terminology and suggest less ambiguous ones.

Introduction

The structure of the classical formulae that converts a pre-test probability, and the sensitivity and the specificity of a diagnostic test, into post-test probabilities is fairly easy to teach. The arithmetic is somewhat tedious, involving two separate multiplications, one addition and one division. Thus, as Berwick et al. documented several decades ago, the formula is not easily remembered or applied quickly. Probability trees help people understand the steps involved, as do different colored dots in a rectangular grid (BMJ right brain article, The Economist article on why Science got it wrong). Once its structure has been firmly understood, calculations carried out by an app do not seem so 'black box.'

In the last few decades, however, more teachers have favoured the formulae based on Likelihood Ratios (LRs), since they cleanly separate the qualities of the test from the setting in which it is applied. Manually, each formula involves one subtraction and one division to obtain an odds, another subtraction and a division to get the relevant LR (LR+ or LR-), the multiplication of the odds and the relevant LR, and a further addition and division to turn the resulting post-test odds back to a post-test probability. But more important than the longer arithmetic, these newer formulae rest on less familiar concepts: odds and Likelihood Ratios. Thus the nomograms, and now the apps, that are widely used to do the calculations, are much more black-box.

The algebraicly-opaque structure of these two formulae pose a challenge for many medical school teachers. They tend to start by presenting elements of the classical formula, using a probability tree, or a box divided two ways, or directly using a 2x2 table with numbers or symbols. If they do present the LR-based approach, they tend to focus on defining the LR, but often merely as a formula involving sensitivity and specificity, rather than in words. They then explain that it makes sense that a ratio above/below 1 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[ref]. Questions from curious first year medical students as why the formula has the form it has are usually deflected, or answered by saying that it stems from Bayes' Theorem.

The objective of this note is to argue that the formula does not have to be so much of a black box, and that the algebra involved in deriving it can actually be made much more intuitive using a graph. Before doing so, I comment on my experiences and examples in teaching these concepts not just to medical students, but to the 3-lawyer panel of CAS judges in the WADA case[ref].

Preliminary Comments on Terminology

Most teachers begin with sensitivity, but I like to emphasize that in developing a diagnostic test, the first task is to agree on a specificity value by setting a cutoff for 'test positivity'. If the measurement behind the test is inherently quantitative (e.g., a concentration), this is done by first assembling a reference distribution of values from similar-age-sex persons *without the target condition*, and setting the threshold for 'normal' or 'negative' so that among such persons, the *false positive rate* (FPR) would be say 5% or 1% (I will use 5% in the example). Until this is set, or there is agreement about what would constitute an 'abnormal' test result (in an image say), one cannot begin to talk about sensitivity.

Lay people understand the concept of setting a threshold, and can turn up of down the threshold for positivity in the spam detector in one's email software, or the motion sensor in a burglar alarm. They also are quite comfortable with the term 'false alarm' and could easily be taught that a false alarm *rate* of 5% should be taken to mean that of *100 genuine (innocent) emails*, 5 would be flagged and put in the spam folder. Clarity on the *denominator* for a FP rate (FPR) is critical. In the psychophysical experiments with radar in the 1950s, the abbreviation FPR and the term False Alarm Rate were widely and correctly used. Slightly earlier, the statistical community, with a blandness only matched by their economic cousins, gave it a name "The type I Error Rate". To be fair, there was one good reason why they called this one the type "I" error, and the other one (to come) the type "II" error, rather than the other way round: *one has to start with the null*, or the innocent, or the healthy, and set the cutoff there, before pursuing the state / condition / disease of concern.

The first person (a PhD statistician in 1947[1]) I know who used the terms sensitivity and specificity defined the latter as "*A measure of specificity* or the probability of correct diagnosis of 'negative'[2] cases." Unfortunately for medical terminology, he chose to focus on the 95% 'True Negative" rate, rather than the 5% False Positive rate, (McNeil NEJM 1975 says FP 'ratio' ??) and to gave it a new name, one that is not that immediately self-evident. Presumably, he chose the word '*specific'* because a diagnostic test that targets TB should not be 'fooled' by background noise, and would find TB and *only* TB (today, in the pursuit of lung cancer, readers of lung CT images that show shadows and TB scars would not interpret them as lung cancer: the ideal system would identify lung cancer, and *only* lung cancer.)

Below, we will use the FPR term whenever we can, and avoid its complement, specificity. But we have no illusions about going back and inventing or adopting a more meaningful term than specificity. It is too late to undo the damage.

Now that a FPR has been adopted, we can address how frequently the detector detects the target condition. Here again, we think the lay public could easily understand, if we choose our terms well.[3]. The '*hit'* rate (to borrow from the radar research[4]) or the *success* rate or the *detection* rate will depend on how hidden and perverse and subtle the targets are, but let us say that at the selected FPR, the spam detector 'catches' or 'traps' or 'identifies' 80% of the malicious emails: the 'detection rate' is 80%.

[1] From Field Studies Section, Tuberculosis Control Division. This paper and the discussion by Professor Neyman, which follows it, were presented before the Institute of Mathematical Statistics, at the twenty-eighth annual meeting of the Pacific Division of the American Association for the Advancement of Science, in San Diego, Calif., June 18, 1947,while the author was visiting professor of biostatistics at the School of Public Health, University of California. Professor Newman, one of those we have to thank for the terms Type I and II error, responded to Yerushalmy's analyses, but used the Greek letter pi

[2] Yerushalmy thinks of instances of the presence and absence of the target condition as 'positive' and 'negative' cases; one needs to carefully distinguish this usage from whether the test is positive or negative.

[3] They also know very well that there is no free lunch: if one wants to detect a greater percentage of the malicious emails, one could turn down the criterion, and vice versa, but let's say we agree that a 5% is a good compromise.

[4] It is interesting how usage has changed, and varies from context. The 'hits' produced by Google searches contain both 'false' and 'true' items. The airforce speaks of attempted air strikes, and hits and misses, and sometimes the misses do damage of their own. The radar people also spoke of a miss rate rather than a hit rate.

Again, it is unfortunate that the world adopted the term ‘sensitivity’ for this 80%, given that it was bound to be misunderstood. Many medical students blame the false alarms on an ‘overly-sensitive’ system that identifies some non-targets, even as it misses (is insensitive to) some of the targets. Here we could have easily stayed with the percentages of targets hit and missed, and adopted the term *hit rate* or better still (to avoid confusion with the Google meaning) *detection rate*. But we went instead with Yerushalmy’s “*measure of sensitivity*: the probability of correct diagnosis of ‘positive’ cases.” Of the {sensitivity, specificity} pair, we find sensitivity to be better fit to the situation it is meant to describe, but below will again try to avoid it whenever there is a chance to use a more natural term.

From pre- to post-test probability

We now switch directions and come to the ‘reverse probabilities’ that are the focus of our note, and that go under the term “positive predictive value” and “negative predictive value.” In a medical diagnostic setting these could be unified and more usefully go under another name, post-test probability. Again, the concept is easy to explain to lay people: in some of the ‘post-positive-test’ examples we like to use are the probability that an email that has been consigned to the spam folder is in fact spam/malignant, that a mammogram-prompted biopsy will come back as breast cancer, or that an alarm the fire/police department responds to does in fact involve a genuine fire/burglary. This last one even suggests its form: it is the number of genuine(g) alarms as a fraction/percentage of the sum of the genuine(g) and the false (f) alarms. [Conversely, “negative predictive values” are about being reassured that the email is ‘safe’ and that nothing is amiss]

I have had trouble locating the first use of the term “positive predictive value” in the diagnostic probability framework. With the help of Google Ngram, I was able to locate instances of the phrase back as far as the 1920s, but in a looser context where psychologists would interpret a predictor “x” that had a positive correlation coefficient with a measured in the future “y” as “having positive predictive value.” I found instances where the g/(g+f) fraction or percentage was used in a medical diagnostic probability context only as far back as the late 1960s[ref]. Although some of the ‘ppv’ entries in the

Ngram do not all refer to medical diagnosis, it is still disheartening that this much less descriptive term is far more popular than ‘post-test probability’.

Before moving on to Likelihood Ratios (one of the two ways of getting from a pre-test probability to a post-test one) it is important to clear up a terminology issue in an otherwise excellent and high profile 1979 publication that showed (using the other method) how to make the pre- post probability jump. Unfortunately, throughout their article, Diamond and Forrester referred to what we would today call pre-test and post-test probabilities as pre-test and post-test likelihoods. They were probabilities, expressed as percentages. Neither likelihoods nor their ratios were used.

The usual way from pre- to post-test probability

The top left panel of Figure 1 was constructed to represent a 50% pre-test probability of being in the darker (target) area. A test with a 5% False Positive Rate will divide the left (lighter) half, producing the upper rectangle (‘false alarm’ area) comprising 5% of the 50%, or 2.5% of the whole. If the test has an 80% detection rate it will divide the right (darker) half, producing an upper rectangle (‘genuine’ area) that comprises 80% of the other 50%, or 40% of the whole. So in all, some 42.5% (all those inside the black border) will test positive. Of these 42.5, some 40/42.5 or 16/17 will be genuine/correct, so the post test probability is 94%.

This arithmetic is an example of the general post-test probability or PPV formula

$$\frac{\text{PretestProbability} \times \text{Sensitivity}}{(1-\text{PretestProbability}) \times (1 - \text{Specificity}) + \text{PretestProbability} \times \text{Sensitivity}}$$

Diamond and Forrester used this formula, as do most modern teachers today. If teachers wish to illustrate the effect of a different pre-test probability, the entire sequence of steps must be repeated. With the 25% (rather than 50%) pre-test probability shown in the panel on the right, the calculation comes out to (80% of 25%) / (80% of 25% + 5% of 75%) = 20/23.75 or 84%, a step down (from 94%) that is not easily anticipated.

The LR way from pre- to post-test probability

The Likelihood Ratio associated with a positive test (LR+) is often simply defined as "Sensitivity divided by (1 minus Specificity)" or Sensitivity / (1 - Specificity). Sometimes it is stated in words as 'the probability that a person *with* the target condition will test positive divided by the probability that a person *without* the target condition will test positive.' [5] or as the ratio of the probability of a positive test in persons *with* the target condition to the probability of a positive test in persons *without* the target condition. In our example, it is 80/5 or 16. In order to apply it, one must first convert the pre test probability [50% in the left panel, 25% in the right] to the pre test *odds [1 in the left, 1/3 in the right]*, using the definition odds = probability/(1-probability). This pre test *odds* is multiplied by the LR to produce the *post test odds (16 in the left, 16/3 in the right)*. In the last step, by reversing the probability to odds equation, one obtains the post test probability as PostTestOdds / (PostTestOdds +1), or 16/17 = 94% in the left and (16/3)/(16/3+1) = 84% on the right.

The Fagan nomogram takes shortcuts through the arithmetic, but in so doing it also hides the logic. But does this have to be? The first conceptual clarification comes from simply referring to '1 minus the specificity' as the False Positive Rate. In both our examples it is 0.05 or 5%. And if we replace the term Sensitivity by the equivalent True Positive Rate (here 80%), we can think of the LR of 16 as saying that 'the True Positive Rate is *16 times whatever the FPR is*' or – in radar terminology – that 'the hit rate is 16 times the false alarm rate.' In the leftmost example, where there are just as many with as without the target condition, the TP:FP ratio is also 16:1 (there are 16 TPs for every 1 FP). In the rightmost example, however, there are 1/3 many with as without the target condition, so the TP:FP ratio is no longer 16:1, but 16 x (1/3) = 16/3 TPs for every 1 FP. In each example, just like above, it merely remains to convert the 16:1 odds to a

[5] The term comes from the Likelihood Ratio statistic used by Newman and Pearson to compute how much more probable the data (here the test results) are under the alternative hypothesis (the target condition) than under the null (its complement). In model selection contexts, it is called the Bayes Factor.

probability of 16/17 (if 50% pretest probability) or (16/3):1 odds to a probability of 16/19 (if 50% pretest probability).

And of course' if test is any good, ratio is >> 1 ' -- usual teacher deflection)

NPV

Same example: it should now be obvious what to do, and to derive the formula from scratch using the diagram ..

Also would like to introduce the NPV in case of haemophilia carriers (women) and using each son they have as a test with perfect specificity but 50% sensitivity. So LR+ is infinite… it's a great 'sorting people' example… and LR- works.. its ½ each time a son is born

Ending Remarks

No problem if use nomogram to do it, but at least now teachers know what the LR x pre-odds is and why the LR comes into it. And don't need to invoke any abstruse Bayes Theorem .. it's all grade 6 math and pictures UNDER CONSTRUCTION

More to be filled in and expanded, but that is the gist of it

JH 2019.09.05

Refs

The Economist. Problems with scientific research: How Science Goes Wrong. Scientific research has changed the world. Now it needs to change itself., Oct 19th 2013

Loong T-W. Understanding sensitivity and specificity with the right side of the brain BMJ 2003;327:716–9

Berwick DM. Fineberg HV. Weinstein MC. When doctors meet numbers. American Journal of Medicine. (1981) Vol 91 pages 991-998.

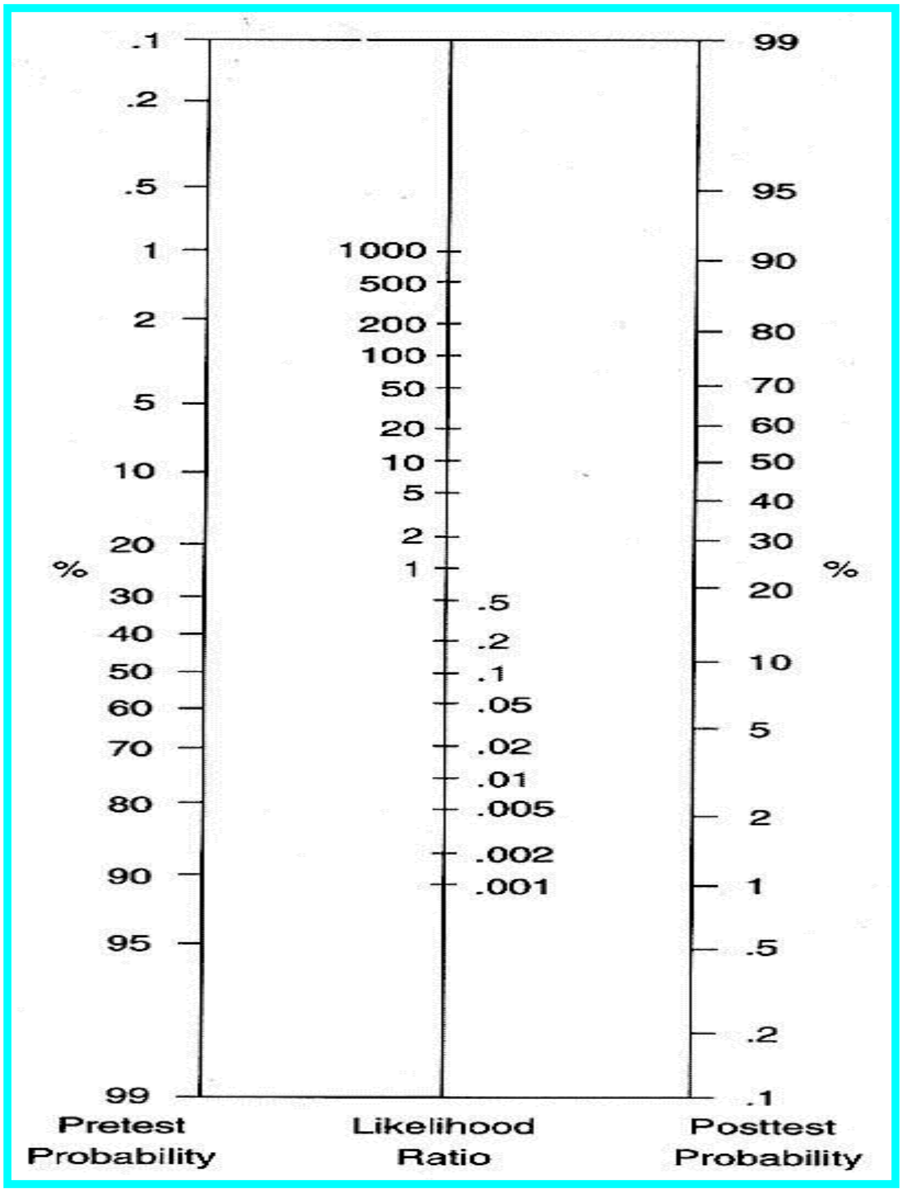
Jacob Yerushalmy. Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques: Public Health Reports Vol. 62, No. 40, Tuberculosis Control Issue No. 20 (Oct. 3, 1947), pp. 1432-1449

McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med. 1975 Jul 31;293(5):211-5.

Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med. 1979 Jun 14;300(24):1350-8.

Fagan nomogram from 1975 NEJM letter -- in response to an uglier NEJM one in 1974

Fagan nomogram

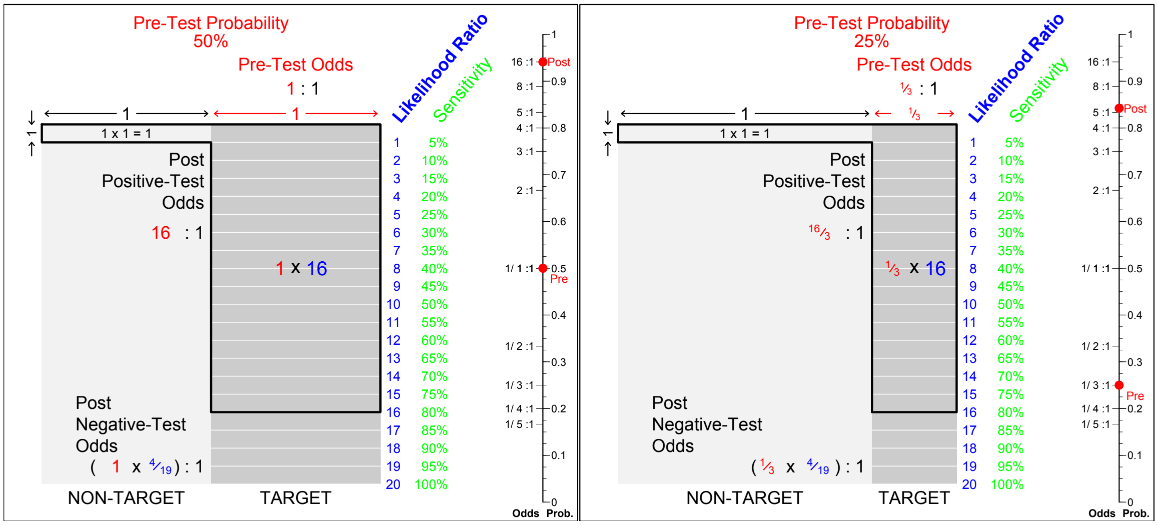


Figure

Mix of *genuinely* positive and *false* positive test results when the target condition is present in ½ or ¼ of those tested.

Likelihood Ratio is the ratio of the number of TRUE: FALSE positive test results, IF the target condition is present in ½ of all instances, i.e., if the pre-test odds were 1:1.

Post test odds must be adjusted according if the prevalence of the target condition is different from this.



From Pre-test to Post-test Probabilities

Test Characteristics

Detection Rate

1 % 80 %

1 11 21 31 41 51 61 71 81 91 99

False Alarm Rate

5 % 99 %

1 11 21 31 41 51 61 71 81 91 99

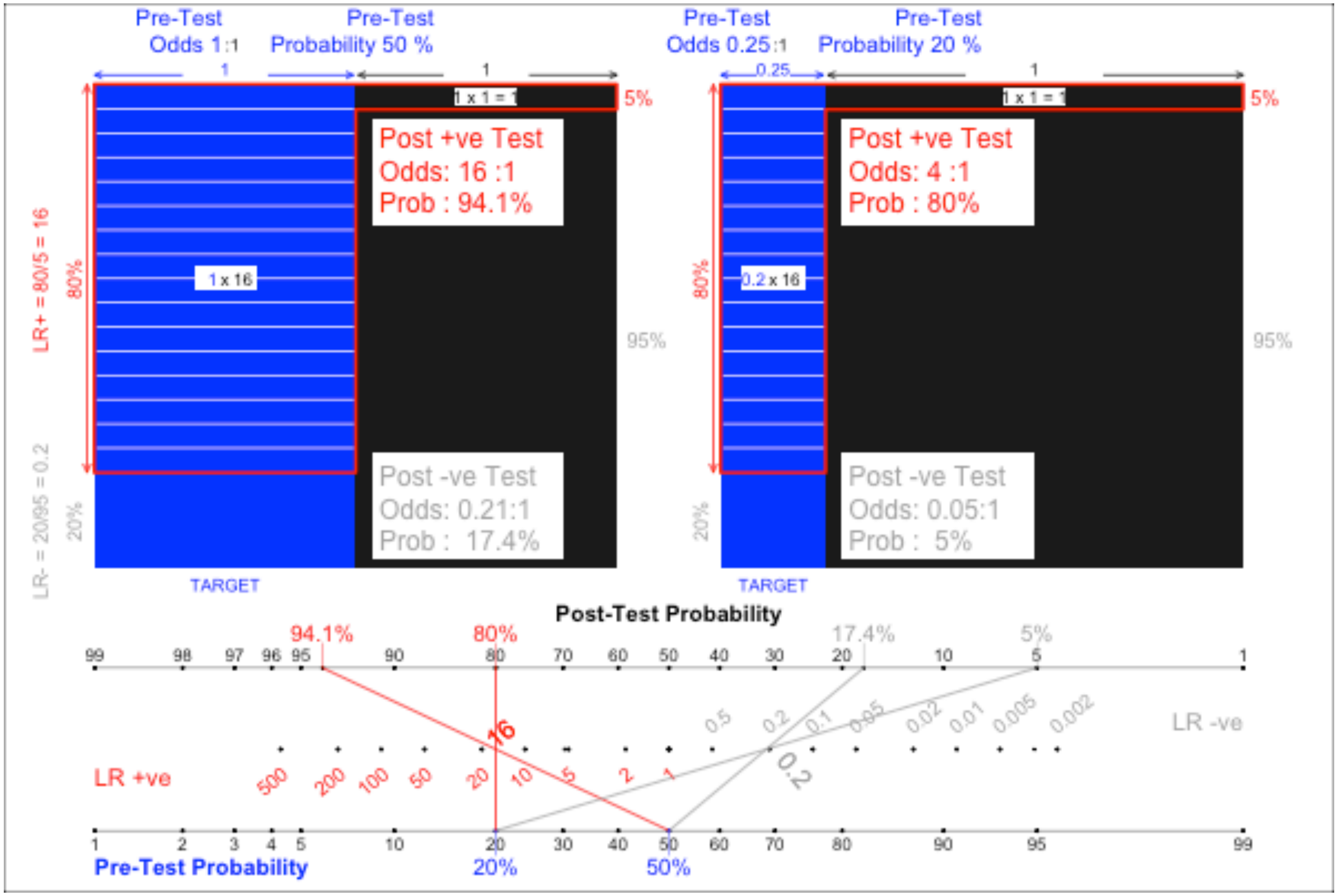
* * * * * * * * * * * * * * * * * *

Setting

Pre-test Probability

20 % 99 %

1 11 21 31 41 51 61 71 81 91 99


Pre-Test
Odds 1:1
Pre-Test
Probability 50 %
1 x 1 = 1
Post +ve Test
Odds: 16 :1
Prob : 94.1%
1 x 16
LR+ = 80/5 = 16
LR- = 20/95 = 0.2
Post -ve Test
Odds: 0.21:1
Prob : 17.4%
TARGET
Pre-Test
Odds 0.25:1
Pre-Test
Probability 20 %
1 x 1 = 1
Post +ve Test
Odds: 4 :1
Prob : 80%
0.2 x 16
Post -ve Test
Odds: 0.05:1
Prob : 5%
TARGET
Post-Test Probability
LR -ve
LR +ve
Pre-Test Probability