Incorporating Uncertainty Into Medical Decision Making: An Approach to Unexpected Test Results

Matt T. Bianchi, MD, PhD, Brian M. Alexander, MD, MPH, Sydney S. Cash, MD, PhD

The utility of diagnostic tests derives from the ability to translate the population concepts of sensitivity and specificity into information that will be useful for the individual patient: the predictive value of the result. As the array of available diagnostic testing broadens, there is a temptation to de-emphasize history and physical findings and defer to the objective rigor of technology. However, diagnostic test interpretation is not always straightforward. One significant barrier to routine use of probability-based test interpretation is the uncertainty inherent in pretest probability estimation, the critical first step of Bayesian reasoning. The context in which this uncertainty presents the greatest challenge is when test results oppose clinical judgment. It is this situation when decision support would be most helpful. The authors propose a simple graphical approach that incorporates uncertainty in pretest probability and has specific application to the interpretation of unexpected results. This method quantitatively demonstrates how uncertainty in disease probability may be amplified when test results are unexpected (opposing clinical judgment), even for tests with high sensitivity and specificity. The authors provide a simple nomogram for determining whether an unexpected test result suggests that one should “switch diagnostic sides.” This graphical framework overcomes the limitation of pretest probability uncertainty in Bayesian analysis and guides decision making when it is most challenging: interpretation of unexpected test results.

Key words: pretest probability; uncertainty; Bayes; unexpected; decision theory.

E ach clinical scenario presents unique challenges related to risks and benefits of testing and therapy, and delivery of appropriate care depends heavily on accurate diagnosis of disease. The increasingly evidence-based context of medicine provides clinicians with important tools to facilitate the diagnostic process, most commonly in the form of the sensitivity and specificity of various tests. Evidence-based diagnosis involves consideration of many factors, including whether a particular patient is sufficiently similar to those reported in studies. Some have advocated incorporating threshold strategies (for deciding whether to test or to treat) into the decision-making process. The challenge of test interpretation is to translate the population statistics of sensitivity and specificity into information relevant to the individual patient. Although formal probability theory is infrequently used in routine clinical practice, the likelihood ratio (LR) method of Bayesian reasoning is a relatively straightforward approach to patient-specific diagnostic interpretations.

One important obstacle to implementing Bayesian reasoning to guide test interpretation is that there is considerable uncertainty in the first and fundamental step—estimation of the pretest probability (pre-TP) of disease. Routine Bayesian strategies (e.g., using the LR-nomogram shown in Figure 1) require designation of a single pre-TP value, even if the clinical impression is qualitative only, or if the epidemiological data are expressed as ranges or intervals. This limitation is crucial because pre-TP estimations are often more appropriately (and intuitively) represented as ranges or intervals of uncertainty. Pre-TP can be estimated from clinical experience, and translation of this
subjective strategy into probability is better described by a range (e.g., high, intermediate, or low probability) than an exact value. Even the rigorous determination of population pre-TP provided by epidemiological studies is commonly presented as a confidence interval, reflecting the uncertainty inherent even in these more formal probability estimations. Finally, aspects of the history and physical findings shaping the clinical assessment may themselves be uncertain or consistent with multiple possible diagnoses. Considering ranges or intervals as a representation of uncertainty is therefore a more intuitive approach that captures the potential complexity of clinical encounters. This framework is compatible with threshold strategies, and it is our specific goal here to demonstrate that it is also compatible with formal diagnostic logic, particularly a Bayesian approach.

Despite the absence of formal probability theory in routine test ordering and interpretation, diagnostic evaluations seem to proceed in the correct direction most of the time. Test interpretation is relatively straightforward when the results agree with the clinical assessment based on history and physical information (i.e., the pre-TP of disease). Although in such circumstances, there is little practical need for detailed statistical decision theory, the interpretation of test results performed in the setting of pre-TP uncertainty may not be so intuitive, particularly those that oppose clinical suspicion. In fact, there are substantial data indicating that variability in physician-generated pre-TP estimations (one type of uncertainty) is not uncommon among physicians. These studies also demonstrate the consequences of test result misinterpretation, particularly when physicians are presented with unexpected test results. The primary source of misunderstanding was either poor estimation of pre-TP or failure to incorporate this information into test interpretation.

We attempt to address two fundamental aspects of diagnostic reasoning that complicate diagnostic test interpretation: uncertainty in the pre-TP and unexpected test results. Decision making under difficult but routinely encountered diagnostic circumstances would benefit from a mechanism for dealing with uncertainty and a guide for weighing clinical suspicion against a test result when the two disagree.

**SENSITIVITY AND SPECIFICITY: FROM POPULATION TO INDIVIDUAL**

The familiar 2 × 2 illustration of disease status versus test result provides the foundation for understanding the clinical utility of diagnostic tests (Figure 1A).
It assumes that disease status can be dichotomized as present or absent and that test results can be either positive or negative. Disease prevalence, or pre-TP, is reflected by the proportion of patients in the “disease present” column compared with the whole population. Whereas sensitivity and specificity address the epidemiologically relevant question of how likely is a positive or negative test result given known disease status, predictive value addresses the following clinically pertinent question: how likely is a given test result true, given unknown disease status? Thinking “horizontally” leads to easy calculation of both positive predictive value (PPV) and negative predictive value (NPV) (Figure 1A).

Unfortunately, the extent to which predictive value depends on disease prevalence in a population (or, for the individual patient, the pre-TP) is not easily apparent using a 2 × 2 box. Indeed, predictive value depends so strongly on pre-TP that the results of a relatively powerful test (e.g., with sensitivity and specificity both > 90%) can be rendered meaningless, and the results of even a random coin toss can seem to have good predictive power (e.g., > 90%) when paired with the appropriate pre-TP. This distinction is also emphasized by the pitfalls of using the common mnemonic “SPin/SNout” as a diagnostic aid (Specific test + Positive result = rule in; Sensitive test + Negative result = rule out). Thus, pre-TP is a critical component of Bayesian reasoning that is required for transitioning from the population concepts of sensitivity and specificity to the predictive value of test results for individual patients.

An alternative to performing predictive value calculations using the 2 × 2 box is to use the Bayes nomogram (Figure 1B). This approach combines test power (sensitivity and specificity) in the form of a single metric, called the likelihood ratio, with the pre-TP, also known as the a priori disease probability. In other words, test results are combined with the best estimation of disease probability (before the test was performed), which is usually based on clinical assessment such as history and physical exam findings, combined with population data from epidemiological studies. We attempt to merge this Bayesian approach with graphical representations of two important clinical challenges to test interpretation: uncertainty in the pre-TP and test results that oppose clinical judgment.

THE LIKELIHOOD RATIO AND BAYES THEOREM

Bayesian analysis uses the LR, which contains information about both sensitivity and specificity of a test, to adjust the probability of disease in an individual patient based on test results. By definition, every dichotomous test has two LRs: one used to adjust disease probability upward (if the result is positive) and one to adjust disease probability downward (if the result is negative). The so-called positive LR, or LR(+), is a value greater than 1 defined by the ratio [sensitivity/(100 − specificity)], whereas the negative LR, or LR(−), is a fraction less than 1 defined by the ratio [(100 − sensitivity)/specificity]. The LR can be considered an adjustment factor applied to the pre-TP to generate a “revised” probability, known as the posttest probability (post-TP). A simple nomogram (Figure 1B) facilitates the use of test results to transform pre-TP into post-TP (circumventing a cumbersome conversion between probability and odds).

From the initial assessment of pre-TP (left column), a straight line through the appropriate LR value (corresponding to the test result) leads to the post-TP of disease. Note that although probabilities are often expressed as fractions between 0 and 1, the nomogram and subsequent discussion will use the corresponding percentage values (0–100). We will consider an example to illustrate the critical importance of pre-TP in Bayesian reasoning.

You are performing a stress echocardiogram on a 42-year-old woman presenting to the emergency department with chest pain, and the result is positive. She asks whether this means that she has acute heart disease, and you correctly defer the answer to this question to the evaluating team for “clinical correlation.” Without more clinical information to establish a pre-TP, the test result cannot be interpreted. If she had no cardiac risk factors and the pain was sharp, brief, and began at rest, her pre-TP might be ∼ 3%. However, if she smokes, has diabetes and hypertension, and the pain was prolonged, exertional, and resolved with rest, her pre-TP could be quite high (we will assume 70% here). The nomogram in Figure 1B shows a graphical representation of the Bayesian approach to this case and the statistical meaning of clinical correlation. The LR(+) of stress echo is estimated at ∼ 6, which is then applied in the nomogram to the low and high pre-TP values that define a potential range of uncertainty about the patient’s baseline risk of coronary disease. This range of pre-TP projects through the LR(+) of the echocardiogram result to define the post-TP window: 15% to 93%. This interval is clearly too large to interpret (and it spans the 50% mark, a potential point of “indifference”) and confirms your decision to defer to clinical correlation. Even if the test had much better discriminative power, with LR(+) = 20 (assuming...
95% sensitive and 95% specific), a similar initial window of pre-TP uncertainty would yield a post-TP window of 33% to 98%—still too wide (uncertain) for meaningful conclusions (Figure 1C). Clearly, a single test result in a given patient (e.g., positive stress test) can yield vastly different certainty of disease depending on the pre-TP. Additional clinical information, combined with relevant epidemiological data, is critical to focus the pre-TP estimation, reduce the uncertainty of this starting point, and thus facilitate accurate interpretation of test results.

UNCERTAINTY IN PRETEST PROBABILITY ESTIMATES

Unfortunately, epidemiological data are not always available or applicable to particular patient presentations, and sometimes the clinical history is poor or complex. Senior clinicians face this sort of uncertainty, where data are lacking, by using intuition based on experience. However, this uncertainty in pre-TP need not represent an obstacle to probabilistic reasoning. Considering a confidence interval or “window” of pre-TP provides an intuitive mechanism for dealing with uncertainty while still implementing Bayesian reasoning. In particular, we emphasize the uncertainty expressed in the pre-TP range of 5% to 25% can increase when the result of a more powerful test, defined here as having an LR(+) of only 0.5, agrees with the clinical suspicion (in this case, a negative result), the uncertainty range is reduced from a baseline pre-TP range of 5% to 25% to a post-TP range of 3% to 15% (Figure 1A). This new range was obtained by using the edges of the pre-TP interval, reading from these 2 points upward to the LR(+) curve for this hypothetical test result, and then reading leftward to specify the post-TP range.

A more accurate (or narrow) pre-TP estimate would have helped to identify the “true” disease probability. A more powerful test also would have helped if it agreed with our clinical suspicion: a negative result from a more powerful test, defined here as having an LR(−) of only 0.1, would collapse the 5% to 25% pre-TP uncertainty window to a small range of post-TP (Figure 1B). This simple graphical method allows one to easily incorporate uncertainty into pre-TP by means of a window or confidence interval. Any test result that agrees with the clinical suspicion will cause that interval to shrink when the post-TP is obtained, and more powerful tests cause more marked collapse of uncertainty ranges (as expected).

Next we consider the impact of an unexpected result using this “window” approach to uncertainty in the pre-TP of disease. The uncertainty expressed in the pre-TP range of 5% to 25% can increase when an unexpected (positive) result is obtained: a poor test, defined here as having an LR(+) = 2, which disagreed with the clinical suspicion, caused the uncertainty window to grow to 8% to 40% (Figure 2B). One might conclude from this hypothetical case that simply using a more powerful test might overcome the issues of uncertainty caused by an unexpected test result (i.e., one could place more confidence in the result of a more powerful test, even if it disagreed with the clinical suspicion). However, when a more powerful test, defined here as having an LR(+) = 10, opposes the clinical expectation, we see that the range of uncertainty is expanded even more than for the less powerful test, to a post-TP range of 18% to...
Note also that this expanded range of uncertainty spanned the 50% mark, which could be viewed as the point of maximal uncertainty.

The reason for the seemingly counterintuitive increase in uncertainty (from the unexpected result of a more powerful test) is easily visualized by considering the shape of the LR curves in the graph (Figure 2).

Unexpected results involve the steep portions of the graph: positive test results opposing low clinical suspicion involve the steep portion of the curves near the top left corner, whereas negative test results opposing high clinical suspicion involve the steep portions of curves near the bottom right corner. As the LR value improves with better test power, the slopes become

Figure 2  Uncertainty in pretest probability estimation. Each panel illustrates the conversion of pretest probability (x-axis) to posttest probability (y-axis) using various likelihood ratio (LR) values representing different test results (curved lines). A “reference” diagonal line is LR = 1 (no change in probability). Better tests for ruling in disease (higher LR+) have curves above this line, whereas better tests for ruling out disease (smaller LR−) have curves below this line. To illustrate a hypothetical patient for whom a disease is thought unlikely, we indicate a range of pretest probability (pre-TP) by a shaded area spanning 5% to 25%. The impact of test results from “poor” (A and B) and “good” tests (C and D) is shown under conditions of agreement with (A and C) or opposition to (B and D) this pre-TP assessment. (A) When a poor test (defined here as having an LR− = 0.5) agrees (is negative) with the expectation (that the patient is unlikely to have disease), the pre-TP uncertainty is narrowed somewhat in the resulting post-TP range. (B) When a poor test (here defined as a LR− = 2) opposes the expectation, the initial range of uncertainty is expanded in the resulting post-TP range. (C) A good test (defined here as having an LR− = 0.1) that agrees with the expectation significantly shrinks the uncertainty window. (D) When the result of a good test (here defined as an LR+ = 10) opposes the expectation, the uncertainty range is expanded substantially (cf. C).
DO ALL UNEXPECTED RESULTS INCREASE UNCERTAINTY?

The phenomenon of amplified uncertainty in disease probability when unexpected test results are encountered depends on 3 factors: 1) the steepness of the LR curve (how powerful the test is), 2) the x-axis position of the “shoulder” of the curve (where the slope of the curve = 1), and 3) the x-axis position of the pre-TP uncertainty window relative to the shoulder (Figure 2). Better tests have LR curves that approach “right angles” at the upper left and lower right corners of this type of graph, and the shoulder moves progressively closer to the extreme probabilities of 0% or 100%. For example, the LR curve of a “superior” test with LR(+) = 500 would approach a right angle in the left upper corner (see Figure 2).

This suggests that in certain circumstances, the pre-TP uncertainty window might not be close enough to the shoulder of the LR curve to produce the paradoxical amplification of uncertainty seen in Figure 2D. Therefore, not all unexpected test results necessarily increase the window of uncertainty. To address this issue, Figure 3 illustrates conditions (combinations of pre-TP estimates and test power) in which an unexpected test result would decrease rather than increase uncertainty.
increase uncertainty in disease probability. We consider 3 window sizes of pre-TP uncertainty—4%, 8%, and 20%—-to reflect potential levels of uncertainty that might be reasonably encountered in practice. We did not consider smaller windows (which approach exact values used in the LR nomogram) or larger windows (which render any meaningful interpretation difficult; see Figure 1B).

To generate the plot of Figure 3, we compared the pre-TP and post-TP uncertainty windows in multiple hypothetical conditions. Each of the pre-TP windows was moved in 1% increments along the x-axis, and the corresponding post-TP window range was obtained for each of 4 hypothetical LR values: LR(+) = 5 (Figure 3A), LR(+) = 20 (Figure 3B), LR(−) = 0.2 (Figure 3C), and LR(−) = 0.05 (Figure 3D). The curves represent the calculated differences between the width of the pre-TP and post-TP windows. Positive values indicate increased uncertainty caused by the test result, whereas negative values represent less uncertainty. In other words, each panel represents a single hypothetical test result, and the curves illustrate how that result affects disease uncertainty depending on the pre-TP estimate (defined by its location on the x-axis and its window size).

The y-intercept of these curves indicates the point at which the test result did not change the range of uncertainty (identical window size before and after the test result). This transition point moves toward the probability extremes (0% or 100%) for higher performance tests, essentially tracking the movement of the shoulder toward the extreme values for more powerful tests, as discussed earlier. The size of the uncertainty window (between 4% and 20%) did not substantially alter the crossover point. As expected, higher LR(+) values left-shift the crossover point from ~30% to ~25% to ~20% for LR(+) values of 5, 10, and 20, respectively (Figure 3A, B).

These graphs illustrate that when the window of uncertainty flanks the shoulder of the LR curves (see Figure 2), an unexpected test result will increase disease uncertainty. More powerful tests may not necessarily circumvent the potential for increased uncertainty from unexpected results, depending on where the pre-TP window of uncertainty lies.

**WHEN SHOULD AN UNEXPECTED RESULT COMPEL YOU TO SWITCH DIAGNOSTIC SIDES?**

We have seen that unexpected test results can increase uncertainty about the presence or absence of disease. From the example in Figure 2D, an unexpected result adjusted disease probability from a range of 5% to 25% to a range of 18% to 78%. Not only did the range increase with the unexpected result, but the post-TP window range also spanned the 50% mark, straddling the diagnostic fence of whether disease is absent or present. To know whether a test result is expected, we used the 50% mark here as a reference point: for pre-TP values above 50%, we would consider a negative result unexpected and vice versa for values below 50%. What if the unexpected result shifted the uncertainty window entirely across this 50% mark or any other defined threshold? When we encounter an unexpected result, we must compare our confidence in the test result with our confidence in the clinical suspicion of disease probability.

This sort of decision making occurs in a qualitative and subjective manner routinely in clinical practice. However, a quantitative approach can be helpful when faced with the challenge of weighing an unexpected result against one’s clinical suspicion. As mentioned earlier, there is abundant evidence from clinical studies that this scenario is precisely where physicians are most prone toward errors of interpretation. One way of phrasing this question is as follows: how good does a test have to be for an unexpected result to move the entire uncertainty window across the 50% mark? This threshold is chosen for illustration here, but any value could be used, tailored to the clinical scenario. In such cases, even if the uncertainty window actually increased with the unexpected test result, the result would nevertheless be helpful in that it would compel the diagnostic suspicion to “switch sides.” In other words, such a situation might represent a criterion for an unexpected test result to outweigh one’s clinical suspicion.

Therefore, we sought a simple graphical method to determine, for any uncertain pre-TP window, how powerful a test would have to be for an unexpected result to force the entire resulting post-TP uncertainty window to the opposite side of the 50% probability threshold. Figure 4 provides a graphical nomogram solution to this question (see figure legend for derivation). One uses the graph in 3 simple steps: 1) define the pre-TP window of uncertainty, 2) draw a vertical line from the more extreme boundary of that window (the end closer to 0% for low pre-TP cases or the end closer to 100% for high pre-TP cases) to the curve, and 3) draw a horizontal line leftward to obtain the minimum power of a hypothetical test that would statistically compel a switch of diagnostic sides.

The graph illustrates the concepts that the closer the pre-TP range is to 0%, the more powerful a test
must be (greater LR(+) for an unexpected test result to move the uncertainty window entirely over the 50% mark (i.e., to switch diagnostic sides). Similarly, the higher the pre-TP is, the smaller the LR(−) must be to compel switching sides. The dotted line indicates the 50% threshold used in this explanation, but any threshold can be used in principle. The gray-shaded region to the left of this threshold indicates LR(+) values that will satisfy this side-switching criterion when the pre-TP is low, whereas the shaded region to the right indicates LR(−) values when the pre-TP is high. In summary, the nomogram provides a simple tool to compare one’s clinical suspicion (in the form of a pre-TP range of uncertainty) with the power of an unexpected test result, to aid in determining which has more weight when there is disagreement.

**DISCUSSION**

In a time when medical practice is increasingly limited by scarce resources, careful consideration must be given to how diagnostic test results will guide management. Understanding Bayes theorem is critical to recognize potential pitfalls of diagnostic interpretations, particularly when unexpected test results are encountered. This important perspective requires that decision making extend beyond the population concepts of sensitivity and specificity, to focus on the predictive value of a given test result for the individual patient. The LR method of Bayesian reasoning incorporates sensitivity and specificity into a strategy that focuses on using tests to adjust disease probability. Given the potential impact of test result misinterpretation on health care delivery and cost, diagnostic strategies should be equipped to deal with the uncertain and the unexpected. Because accurate pre-TP assessment (including clinical features and epidemiological data) remains the cornerstone of diagnosis, progress in epidemiology and improved clinical skills represent important pathways to improve diagnostic acumen at the point of care by decreasing uncertainty in pre-TP estimation. A combination of a threshold framework for testing (or treating) and an appreciation for how tests of various power (sensitivity and specificity) can alter disease probability may help clinicians decide whether to order a diagnostic test, as well as interpret test results in a manner that considers the inherent uncertainty in diagnostic evaluations.

Despite the critical importance of pre-TP on test result interpretation, using this information clinically remains a challenge across multiple levels of medical training. Adding to the complexity is the challenge of accurate pre-TP estimation, which may be better reflected as an interval than a single number, particularly when there are few epidemiology data or when a given patient is not similar to published data. We provide a simple graphical method to incorporate uncertainty into pre-TP estimation that is compatible with a Bayesian approach to diagnostic reasoning. This method is applied to situations in which decision support would be most helpful: unexpected test results. The nomogram provides a visual tool for weighing clinical suspicion against diagnostic test power when there is disagreement. Like the standard Bayes nomogram, it can be used as a “pocket guide” for interpretation of any unexpected test result, provided there is some knowledge of the test’s sensitivity and specificity. Simple calculations allow adaptation of our nomogram to accommodate any threshold value. Although it cannot substitute for the intricacies of clinical reasoning, it provides a semi-quantitative framework to encourage discussion and take advantage of probability theory in situations when decision
support might prove most helpful. We can think of diagnostic tests as serving several statistical purposes: adjustment of disease probability, narrowing windows of uncertainty, and potentially compelling us to change our minds. A graphical aid to this approach complements the wisdom of experience in facing the challenges of diagnostic uncertainty and unexpected results.

REFERENCES