

On Uncertainty in Medical Testing

Robert L. Winkler, PhD, James E. Smith, PhD

There is confusion in the medical decision-making literature about how to handle uncertainty in medical tests. In this article, the authors consider the situation in which there is uncertainty about the pretest probability of a disease in a patient as well as uncertainty about the sensitivity and specificity of a diagnostic test for that disease. They discuss how to calculate posttest probabilities of a disease under such uncertainty and how to calculate a distribution for a posttest probability. They show that given certain independence assumptions, uncer-

*tainty about these parameters need not complicate the calculation of patient positive predictive values: One can simply use the expected values of the parameters in the standard Bayesian formula for posttest probabilities. The discussion on how to calculate distributions for positive predictive values corrects a common and potentially important error. **Key words:** predictive value of tests; sensitivity and specificity; Bayesian analysis; Bayes' theorem; uncertainty. (*Med Decis Making* 2004;24:654-658)*

Bayes' rule is widely recognized as a useful tool for interpreting clinical test results.¹⁻³ In most applications, the sensitivity and specificity of the test and prior (pretest) probability are taken to be fixed parameters. Clinicians and researchers have long worried about the impact of uncertainty in these parameters and the effects of this uncertainty on the calculated posterior probabilities and clinical decision making. For example, Baron^{4(p49)} argued that "it is unlikely that great precision in operating characteristics [of a test] can be achieved" and concluded that because of this uncertainty in operating characteristics, positive predictive values (PPVs) may not be useful for individual patients. Mossman and Berger⁵ and Zou⁶ described methods for calculating confidence intervals for posttest probabilities, quantifying the uncertainty in PPVs due to uncertainty about the parameters.

Mossman and Berger motivated their analysis with an example that considers a hypothetical case of a statistically sophisticated patient, Mr Smith, asking his doctor, Dr Jones, to interpret a positive test result for a condition *D*:

The published estimates of prevalence, sensitivity and specificity are subject to random sampling error, so what I want to know is this: What is the 95% confidence interval for my probability of having *D* given my positive test result and the imprecision in the estimates? Knowing whether the interval is narrow or broad might affect my decisions about getting other tests or choosing treatment.^{5(p498-9)}

Mossman and Berger noted that these intervals may be quite wide: In the Smith-Jones example, a 95% confidence interval for the PPV is (0.431, 0.887).

In this article, we consider the calculation of PPVs and distributions for PPVs when there is uncertainty about a test's characteristics or the prevalence of a disease. First, we show that, given certain independence assumptions, uncertainty about these parameters does not really change the way to calculate a PPV for a particular patient: We can use the standard Bayesian formula for posttest probabilities, replacing the parameters with their expected values. Next, we describe how to calculate distributions for PPVs that reflect the uncertainty in parameters, clarifying some confusion about this in the literature. Our focus is on PPVs, although the same issues and techniques apply with negative predictive values.

Our interest in this problem was motivated by a particularly extreme case of uncertainty about a test that involved Casey Smith, the newborn daughter of one of the authors. Casey was the first ever to test positive for a rare enzyme deficiency in an experimental screening program that had tested some 13,000 newborns; the result turned out to be a false positive. Our analysis of that problem focused on uncertainty in the false-

Received 22 October 2003 from the Fuqua School of Business, Duke University, Durham, North Carolina. The authors are grateful to an anonymous reviewer for helpful comments. Revision accepted for publication 9 August 2004.

Address correspondence and reprint requests to Robert L. Winkler, Fuqua School of Business, Duke University, Durham, NC 27708-0120; e-mail: rwinkler@duke.edu.

DOI: 10.1177/0272989X04271045

positive rate and is discussed elsewhere.^{7,8} Here we consider the problem with more generality with the goal of clarifying how one should proceed when there is uncertainty about a test's characteristics or the prevalence of a disease.

CALCULATING PPVs

Let p represent the prevalence of a disease D in the relevant population or the prior (pretest) probability that the patient has D , and suppose that a diagnostic test yields a positive or negative result about D . Let s represent the sensitivity or true-positive rate, the probability of a positive result for a member of the population who has D . Let t represent the specificity or true-negative rate, the probability of a negative result for a member of the population who does not have D (i.e., who has \bar{D}). Given p , s , and t , the PPV, the posttest probability that the patient has D if the test result is positive, is, from Bayes' rule,

$$P(D|+, p, s, t) = \frac{ps}{ps + (1-p)(1-t)} \tag{1}$$

As Baron⁴ demonstrated, this posterior probability is a highly nonlinear function of the parameters, and small changes in p , s , and t may lead to large changes in PPVs. For example, Figure 1, adapted from Baron, shows PPVs as a function of the prior probability p for a variety of different specificities t , holding the sensitivity s fixed at 0.95. Here we see that with a small p , small changes in p or t can lead to large changes in the posterior probability.

What if we are uncertain about p , s , and t ? In a Bayesian framework, we can describe this uncertainty by assigning a probability distribution over these parameters. Sometimes we have data about p , s , and t in the form of samples from the relevant population that can be used to determine distributions for p , s , and t . Information about p , s , and t may also come in other forms. For example, if a patient is suspected of having coronary heart disease, we might use age and risk factors together with a logistic regression model (perhaps one based on the Framingham heart study⁹) to estimate the probability p that the patient has the condition. Uncertainty in p may include that due to residual error in the logistic regression model and/or uncertainty about the presence of particular risk factors in the patient. Information about p , s , and t may also come in more subjective forms: Does the patient have a condition that might predispose him or her toward a positive reading on the test even if he or she does not have D ? Are there

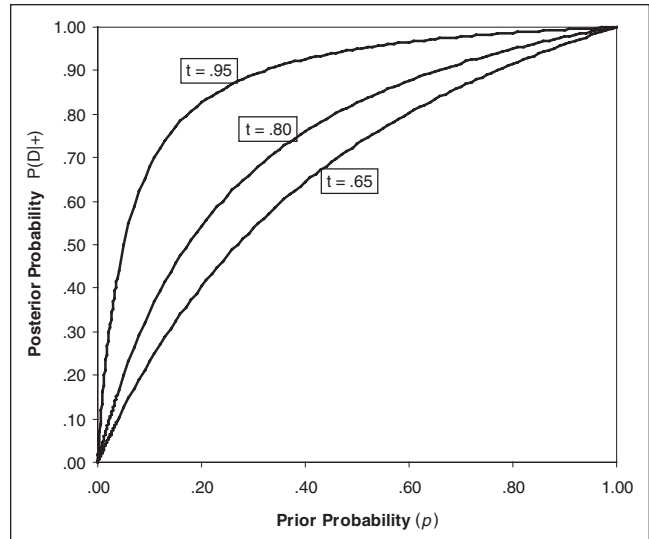


Figure 1 Variation in positive predictive values with variation in prior probability (p) and specificity (t) when sensitivity (s) is 0.95.

biases in the estimates of the test characteristics because of selection biases in clinical evaluations of the test?^{1,10} Such information can have a significant effect on the PPV and should be reflected in the prior distributions for p , s , and t , even if it requires subjective judgment.

If $f(p, s, t)$ denotes the joint density for the parameters, we can find the PPV for a particular patient in several different ways. First, using the definition of conditional probability,

$$P(D|+) = \frac{P(D, +)}{P(+)} = \frac{P(D, +)}{P(D, +) + P(\bar{D}, +)} \tag{2}$$

where $P(D, +)$ and $P(\bar{D}, +)$ may be found by taking the expected value of ps and $(1-p)(1-t)$, integrating out the uncertainty in p , s , and t :

$$P(D, +) = \iiint P(D, + | p, s, t) f(p, s, t) dp ds dt = \iiint psf(p, s, t) dp ds dt \tag{3}$$

and

$$P(\bar{D}, +) = \iiint P(\bar{D}, + | p, s, t) f(p, s, t) dp ds dt = \iiint (1-p)(1-t) f(p, s, t) dp ds dt \tag{4}$$

If p is independent of s and t so that the joint distribution $f(p, s, t)$ may be factored as $f(p)f(s, t)$, we can further simplify the calculation of the PPV by writing

Equations 3 and 4 as $P(D, +) = E(s)E(p)$ and $P(\bar{D}, +) = [1 - E(p)][1 - E(t)]$, where $E(p)$, $E(s)$, and $E(t)$ denote the expected values of p , s , and t . In this case, Equation 2 becomes

$$P(D|+) = \frac{E(p)E(s)}{E(p)E(s) + [1 - E(p)][1 - E(t)]} \tag{5}$$

Comparing Equations 5 and 1, we see that we have simply replaced the uncertain parameters (p , s , and t) appearing in Equation 1 with their expected values. Thus, the presence of uncertainty in the prevalence or test parameters does not really complicate the calculation of patient PPVs.

CALCULATING DISTRIBUTIONS FOR PPVs

Although the PPV for a particular patient is a single number that can be calculated using Equation 5, we can also think of the PPV as an uncertain quantity that reflects the uncertainty in p , s , and t . We can construct the probability distribution for PPVs by generating random samples of (p, s, t) and calculating $q \equiv P(D|+, p, s, t)$ from Equation 1 for each sample. We can then average these sample qs to find the expected q , which would be the PPV for a particular patient. The full distribution for PPVs may also be useful for other purposes, such as constructing confidence intervals or generating ranges to be used in sensitivity analyses in decision-making studies.

This simulation procedure has been implemented in the literature^{3,5} to construct a distribution for PPVs and similar distributions by generating random samples of (p, s, t) from $f(p, s, t)$, effectively using the following formula for the patient PPV:

$$P(D|+) = \iiint P(D|+, p, s, t)f(p, s, t)dp ds dt \tag{6}$$

Several readers of our papers about Casey's problem^{7,8} suggested that we should have calculated the expected PPV this way rather than relying on Equation 5, which we used without discussion or derivation. However, this approach for calculating distributions for PPVs is incorrect because it fails to recognize that the test result itself provides information about the parameters p , s , and t . Intuitively, the distributions $f(p, s, t|+)$ and $f(p, s, t)$ differ because the occurrence of the positive test result not only tells us about the probability that the patient has the disease but could also indicate a higher

prevalence and a higher sensitivity (if it is a true positive) or a lower prevalence and a lower specificity (if it is a false positive).

A correct way to calculate $P(D|+)$ by integrating out the uncertainty in p , s , and t in Equation 1 is to determine

$$P(D|+) = \iiint P(D|+, p, s, t)f(p, s, t|+)dp ds dt, \tag{7}$$

with $P(D|+, p, s, t)$ given by Equation 1 and $f(p, s, t|+)$ given by Bayes' rule as

$$f(p, s, t|+) = \frac{P(+|p, s, t)f(p, s, t)}{\iiint P(+|p, s, t)f(p, s, t)dp ds dt}, \tag{8}$$

where

$$P(+|p, s, t) = ps + (1 - p)(1 - t)$$

is the denominator of Equation 1. These formulas follow directly from the definition of a conditional probability and do not rely on any independence assumptions. To see that Equations 2 and 7 are equivalent, note that substituting Equation 8 into 7, we have

$$P(D|+) = \frac{\iiint P(D|+, p, s, t)P(+|p, s, t)f(p, s, t)dp ds dt}{\iiint P(+|p, s, t)f(p, s, t)dp ds dt} \tag{9}$$

Since $P(D|+, p, s, t)P(+|p, s, t) = P(D, +|p, s, t)$, the numerator of Equation 9 corresponds to $P(D, +)$ as specified by Equation 3. The denominator similarly corresponds to $P(+)$.

We can calculate the distribution for PPVs and evaluate Equation 7 using simulation. With many of the distributions encountered in practice, the simulations are easier if we represent $f(p, s, t|+)$ as a mixture of the distributions, $f(p, s, t|D, +)$ and $f(p, s, t|\bar{D}, +)$, that we would have if we knew the patient had D or \bar{D} , respectively. The mixing probabilities are $P(D|+)$ and $P(\bar{D}|+)$:

$$f(p, s, t|+) = f(p, s, t|D, +)P(D|+) + f(p, s, t|\bar{D}, +)P(\bar{D}|+) \tag{10}$$

To sample from $f(p, s, t|+)$, we first find $P(D|+)$ from Equation 2 or 5 and then draw a uniform random number from $[0,1]$. If the random number is less than or equal to $P(D|+)$, we take the positive result to be a true positive and draw (p, s, t) from $f(p, s, t|D, +)$; otherwise, we take the positive to be a false positive and draw $(p, s,$

t) from $f(p, s, t | \bar{D}, +)$. Then we find $q = P(D | +, p, s, t)$ from (p, s, t) via Equation 1 and repeat this process a large number of times to construct the distribution of q . Averaging these qs gives $P(D | +)$ via Equation 7.

To illustrate, let us consider the Mr Smith/Dr Jones example from Mossman and Berger. In their “objective Bayesian method,” $f(p, s, t)$ is a product of 3 (independent) beta densities:

$$f(p, s, t) = f_{\beta}(p | 10.5, 40.5)f_{\beta}(s | 36.5, 4.5)f_{\beta}(t | 36.5, 4.5),$$

where

$$f_{\beta}(x | a, b) \propto x^{a-1}(1-x)^{b-1}.$$

Information about the parameters is assumed to come from 3 separate samples. The distribution for prevalence, $f_{\beta}(p | 10.5, 40.5)$, is based on an initial noninformative prior followed by a sample of 50 patients, 10 of whom are found to have D and 40 who have \bar{D} . The prior is proportional to $[p(1-p)]^{-1/2}$, which is known as a Jeffreys’ prior or reference prior.^{11,12} The evidence about p can be interpreted as equivalent to having seen 10.5 cases with the disease (D) and 40.5 cases without the disease (\bar{D}) in a sample of 51 patients; the 10 and 40 come from the sample, and the additional 0.5 and 0.5 come from the Jeffreys’ prior. Similarly, the distributions for s and t can be interpreted as the test having correctly categorized 36.5 of 41 patients having the disease and 36.5 of 41 patients not having the disease. With these assumptions, $E(p) = 10.5/51$, $E(s) = 36.5/41$, and $E(t) = 36.5/41$.

We can then calculate Mr Smith’s probability using Equation 5:

$$P(D | +) = \frac{(10.5/51)(36.5/41)}{(10.5/51)(36.5/41) + (40.5/51)(4.5/41)} = 0.678.$$

Having seen a positive reading but not knowing if Mr Smith has D , our updated joint probability distribution for prevalence and test statistics, $f(p, s, t | +)$, is a mixture of 2 densities, as given by Equation 10:

$$f(p, s, t | +) = 0.678f_{\beta}(p | 11.5, 40.5)f_{\beta}(s | 37.5, 4.5)f_{\beta}(t | 36.5, 4.5) + 0.322f_{\beta}(p | 10.5, 41.5)f_{\beta}(s | 36.5, 4.5)f_{\beta}(t | 36.5, 5.5). \quad (11)$$

The first term on the right-hand side of Equation 11 considers the possibility that Mr Smith has D and therefore increases the count for the prevalence (10.5 cases of D becomes 11.5) and the sensitivity (36.5 D cases correctly identified becomes 37.5). The 2nd term

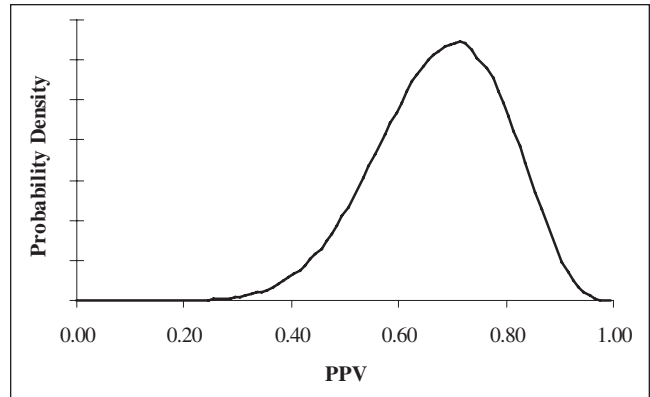


Figure 2 Distribution of positive predictive values (PPVs) for the Smith-Jones example.

corresponds to Mr Smith’s not having D , in which case we increase the counts of patients not having the disease (from 40.5 to 41.5) and the number of patients without the disease who have been categorized incorrectly (from 4.5 to 5.5). Simulating 250,000 values from $f(p, s, t | +)$ and finding q for each value from Equation 1 results in the density for q given in Figure 2.

Although the errors due to using $f(p, s, t)$ and the erroneous Equation 6 instead of $f(p, s, t | +)$ and Equation 7 to determine the distribution for q and its mean, $P(D | +)$, are small in this example, the errors can be much larger in other cases. For instance, Figure 3 shows $f(q)$ based on $f(p, s, t)$ and $f(p, s, t | +)$ when

$$f(p, s, t) = f_{\beta}(p | 10.5, 990.5)f_{\beta}(s | 20.5, 0.5)f_{\beta}(t | 19.5, 1.5).$$

The mean for q is 0.13 from $f(p, s, t | +)$ and 0.20 from $f(p, s, t)$, and the corresponding 95th percentiles are 0.33 and 0.54. In Casey’s problem,^{7,8} we took p and s to be certain with $p = 1/250,000$ and $s = 999/1000$ and assumed t had a beta distribution $f_{\beta}(t | 13999, 1)$. In this case, the correct mean for q from $f(p, s, t | +)$ is 0.05, whereas the incorrect mean for q from $f(p, s, t)$ is 0.14. Such differences in PPVs could lead to different treatments, and the differences in 95th percentiles should be very disturbing for those who worry about the uncertainty in PPVs.

DISCUSSION

Uncertainty in test characteristics or prevalence may make it harder to determine posttest probabilities of a disease, but in our view, it does not suggest abandoning the Bayesian methodology. Indeed, such uncertainty suggests redoubling our attention to Bayesian principles, thinking carefully about what we know

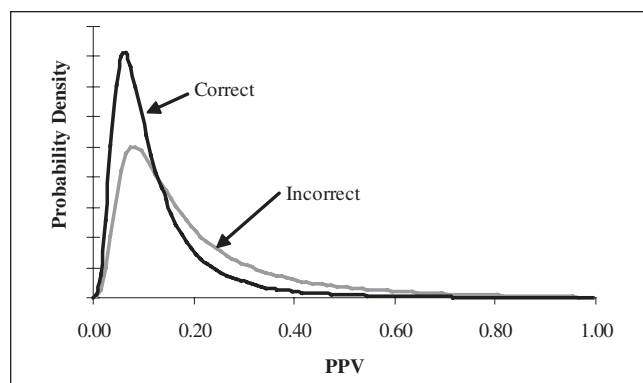


Figure 3 Correct and incorrect distributions of positive predictive values (PPVs).

about the disease and the test, and incorporating this information into our inferences and predictions. Our analysis of the impact of uncertainty about test characteristics or prevalence is fully Bayesian and demonstrates that we can generally calculate the PPV for a particular patient using the standard Bayesian formula for posttest probabilities, simply replacing the parameters with their expected values. This implies that uncertainty about the parameters does not really change the calculation of PPVs. For those interested in how parameter uncertainty translates into uncertainty about PPVs, we show how to calculate a distribution for a posttest probability correctly, thereby clarifying some confusion about this in the literature. The key insight is that the newly observed test result itself provides information about the parameters that must be incorporated in the distribution of PPVs.

Having been in the shoes of Mossman and Berger's "Mr Smith" when Casey was the first to test positive in an experimental screening program, we confess to having worried a great deal about uncertainty about the

test characteristics. In our analysis,^{6,7} this "worry" took the form of performing sensitivity analysis to understand how changes in the assumptions about the test affected the PPV and ultimately led us to develop a Bayesian model of specificity to understand the implications of 13,000 negative results. Although the only output of this model used in the PPV calculation was the expected specificity, considering this uncertainty and constructing the model gave us more confidence in our assessment of this key probability and hence in the calculated result.

REFERENCES

1. Sox HC, Blatt MA, Higgins MC, Marton KI. *Medical Decision Making*. Newton, MA: Butterworth-Heinemann; 1988.
2. Weinstein MC, Fineberg HV, Elstein AS, et al. *Clinical Decision Analysis*. Philadelphia: W. B. Saunders; 1980.
3. Parmigiani G. *Modeling in Medical Decision Making: A Bayesian Approach*. Chichester (UK): Wiley; 2002.
4. Baron JA. Uncertainty in Bayes. *Med Decis Making*. 1994;14:46–51.
5. Mossman D, Berger JO. Intervals for posttest probabilities: a comparison of 5 methods. *Med Decis Making*. 2001;21:498–507.
6. Zou G. From diagnostic accuracy to accurate diagnosis: interpreting a test result with confidence. *Med Decis Making*. 2004;24:313–8.
7. Smith JE, Winkler RL. Casey's problem: interpreting and evaluating a new test. *Interfaces*. 1999;29:63–76.
8. Smith JE, Winkler RL, Fryback DG. The first positive: positive predictive values in the extreme. *Ann Intern Med*. 2000;132:804–9.
9. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837–47.
10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926–30.
11. Jeffreys H. *Theory of Probability*. 3rd ed. Oxford (UK): Clarendon; 1961.
12. Bernardo JM. Reference posterior distributions for Bayesian inference. *J Royal Stat Soc B*. 1979;41:113–47.