The First Positive: Computing Positive Predictive Value at the Extremes

James E. Smith, PhD; Robert L. Winkler, PhD; and Dennis G. Fryback, PhD

Computing the positive predictive value (PPV) of a well-known test for a relatively common disease is a straightforward exercise. However, in the case of a new test for a rare disorder, the extreme numbers involved—the very low prevalence of the disorder and the lack of previous false-positive results—make it difficult to compute the PPV. As new genetic tests become available in the next decade, more and more clinicians will have to answer questions about PPVs in cases with extreme prevalence, sensitivity, and specificity. This paper presents some tools for thinking about these calculations. First, a standard PPV calculation with rough estimates of the prevalence, sensitivity, and specificity is reviewed. The "zero numerator" problem posed by not having seen any false-positive results is then discussed, and a Bayesian approach to this problem is described. The Bayesian approach requires specification of a prior distribution that describes the initial uncertainty about the false-positive rate. This prior distribution is updated as new evidence is obtained, and the updated expected false-positive rate is used to calculate PPVs. The Bayesian approach provides appropriate and defensible PPVs and can be used to estimate failure rates for other rare events as well.

Ann Intern Med. 2000;132:804-809.

For author affiliations and current addresses, see end of text.

A genetic disorder—serious, perhaps fatal without treatment. Very rare, an incidence of maybe 1 in 250 000. A new test, thought to be highly sensitive and specific. To date, 13 000 newborns have been screened, with no positive results. Until now. "What is the chance Casey has it, doctor?"

In situations like this, the lack of previous false-positive results and the rarity of the suspected disorder make the positive predictive value (PPV) of the test result difficult to estimate. As the father of baby Casey, one of the authors asked the above question and found that the calculated PPVs varied widely—from 1.7% to 100%—as we varied the false-positive rate over reasonable ranges. What, then, is the probability that Casey has the condition? How should we calculate PPVs in such situations? As new genetic tests become available in the next decade, more and more clinicians will have to deal with such questions, which also arise with evaluation of the effectiveness of new tests and choosing which tests to administer. Similar questions arise in other situations in which we must consider probabilities of failures before they have occurred. For example, what is the probability that a new drug has a side effect that was not seen in clinical trials? What is the chance of a particular complication during a new surgical procedure?

In this paper, we describe the approach that we used in Casey's situation and indicate how it could be used in other, analogous situations. We begin by reviewing the calculation of PPVs, highlighting the difficulties encountered with new tests. Next, we consider approaches for dealing with the "zero-numerator problem"—estimating proportions when there are zero observations in the numerator. We consider a standard approach to the "zero-numerator problem" and then describe a simple Bayesian approach. We conclude by considering some of the benefits of the Bayesian approach and barriers to its use. In this paper, we focus on the generic aspects of Casey's situation; for more specific information, see reference 1.

Computing Positive Predictive Value

Three estimates are required to compute the positive predictive value (PPV) of a particular test and disorder (2–4): the *prevalence* of the disorder (the pretest probability that the patient being tested has the disorder), the *sensitivity* of the test (the probability that the test result is positive for someone who has the disorder), and the *specificity* of the test (the probability that the test result is negative for someone who does not have the disorder).

To calculate the PPV, we first need to consider the probability of receiving a positive test result, either true or false. The probability of a true-positive result is given by the probability that the patient has the disorder times the probability that the test result is positive for someone who has the disorder: prevalence × sensitivity. The probability of a false-positive result is given by the probability of not having the disorder times the probability of a positive test result for a patient without the disorder: (1 − prevalence) × (1 − specificity). The PPV is then:

$$\text{PPV} = \frac{\textit{prevalence} \times \textit{sensitivity}}{\textit{prevalence} \times \textit{sensitivity} + (1 - \textit{prevalence}) \times (1 - \textit{specificity})}$$

In other words, the PPV equals the fraction of patients who truly have the disease among all of those who receive positive test results.

In Casey's situation, suppose we assume that the disorder occurs in 1 in 250 000 newborns (that is, the prevalence is 1/250 000) and assume that that the sensitivity of the test is 100%. What about the specificity of the test? Should we assume that it too is 100%? After all, the test has been negative in the past 13 000 births, and we have never seen a false-positive result. If we assume that the sensitivity and specificity are both 100% (that is, the test is perfect), we find a PPV of (1/250 000)/(0 + 1/250 000) = 100%.

But the test may not be perfect, and even a small false-positive rate will change the computed probability dramatically. For example, if the false-positive rate, $f = 1 -$ specificity, is $1/13\,000 = 0.0077\%$ instead of 0/13 000, the PPV decreases dramatically, from 100% to 4.9%. Because the prevalence is so low (1 in 250 000), the PPV is extremely sensitive to small changes in the false-positive rate. The results are less responsive to changes in the sensitivity: If we assume a sensitivity of 90% instead of 100% and maintain the assumption of perfect specificity, the PPV remains 100%. If we assume a sensitivity of 90% instead of 100% and assume a false-positive rate of 1/13 000, the PPV decreases from 4.9% to 4.5%.

The Problem of Zero Numerators

How should we estimate the false-positive rate, *f*, when we have never seen a false-positive result? In many contexts, this rate is estimated by calculating the ratio of the number of times the event has occurred (r) and the number of trials (n): $f = r/n$. In classic statistics, this is called the *maximum likelihood estimate* because the value of f maximizes the probability of observing the given data (r occurrences in n trials). With no false-positive results in 13 000 trials, this leads to an estimated false-positive rate of $f = 0/13\,000$—that is, perfect specificity—which results in a PPV of 100%.

To be more conservative in our estimate of the false-positive rate, we might consider a range of estimates. For example, the "rule of three" states that "if none of n patients shows the event about which we are concerned, we can be 95% confident that the chance of this event is at most three in n (i.e., $3/n$). In other words, the upper 95% confidence limit of a $0/n$ rate is approximately $3/n$" (5). If n is greater than 30, this approximate rule agrees with the exact calculation of the upper confidence limit to the nearest percentage point. If we apply the rule of three to the false-positive rate, the upper confidence limit is 3/13 000 (the exact value is 2.9954/13 000 [6]), which leads to a PPV of 1.7%. The lower 95% confidence limit on f given zero occurrences is 0%, yielding a PPV of 100%.

The 95% confidence limits thus give false-positive rates ranging from 0/13 000 to 3/13 000 with corresponding PPVs ranging from 100% to 1.7%. Although it is possible that the same actions would be recommended for PPVs in this range (perhaps follow-up tests), in many cases the indicated tests or treatments would vary across such a broad range; the precise thresholds would depend on the potential benefits, harms, and costs of the treatment (7). In any event, the range is so broad that it does not convey much information to the concerned patient or family. How can we do better?

A Bayesian Approach

A Bayesian approach to this problem is summarized in **Figure 1**. To describe the uncertainty surrounding the false-positive rate, we must specify two things. First, we must specify the likelihood of observing the actual results as a function of the false-positive rate. In this setting, we assume that each healthy baby has the same probability f of receiving a false-positive result and that the test results for different babies are independent. This allows us to calculate the probability of seeing, for example, 0 false-positive results in 13 000 tests with the binomial probability formula (Appendix). This assumption seems natural and appropriate in this setting, and the maximum likelihood estimates and the rule of three are also based on this assumption.

Second, we must specify a prior probability distribution for the false-positive rate *f* that describes the uncertainty about *f* before seeing the results of tests. Protocols for assessing probability distributions are discussed in detail elsewhere (8, 9); these typically involve asking experts a series of questions, such as "What is the probability that the false-positive rate is less than 1 in 1000?" or "Would you rather bet on heads in a toss of a fair coin or the false-positive rate being less than 1 in 1000?" From these assessments, one can sketch a probability distribution or fit some mathematical form. Ideally, these prior probability distributions should be assessed before the experimental screening program begins by questioning the persons who are most knowledgeable about the test. Their assessments should reflect their understanding of the biochemical or genetic methods used in the new test, their experience in preliminary laboratory uses of this test, and experience with other tests using similar methods. Moreover, the assessments should consider the possibilities for human error; with highly sensitive and specific tests, human error may become a much more important determinant of reliability than the workings of the underlying test itself. If we consult multiple experts, they may disagree about the prior probabilities. One could attempt to combine the different assessments into an aggregate assessment from a panel of experts (10) or, alternatively, report the whole set of prior probabilities and the corresponding posterior probabilities and PPVs. It may well be that the same treatments or follow-up actions are indicated in all cases and the disagreement need not be resolved.

The Bayes theorem is used to update the prior distribution to a revised (posterior) distribution after seeing the results from repeated uses of the test. We can then use this posterior distribution on the false-positive rate to calculate the PPV for a particular patient, such as Casey. To do this, we need to determine the mean of the posterior distribution; this "expected probability" is the probability we would assign to the next patient (Casey) having a false-positive result and can be used in the usual way in the formula for calculating PPVs. Although this whole procedure can be applied with any prior distribution, if the prior distribution is a beta distribution, the procedure is very easy to apply. This form can represent a wide variety of shapes of distributions (some examples are shown in **Figures 2** and **3**) and leads to a posterior distribution that is also a beta distribution with updated parameters. The Bayes theorem and the beta distribution are discussed in more detail in the Appendix.

In our situation, no formally assessed prior distribution was available that could be used to determine Casey's PPV. To get a sense of the range of possible PPVs, we considered a variety of prior distributions on the basis of our understanding of the test and consultations with the researchers who developed the test. Our assumptions and results are summarized in **Figure 2**. To improve the readability of the graphs, the scale of the horizontal axis is given in terms of the logarithm of *f* and the probability density plots are densities for log *f*. The priors are shown in the left column of the figure, with our "base-case" prior in the middle row; this is a beta distribution with the parameters indicated in the figure. These parameters imply that there is a 5% chance that *f* is less than 0.0000513 (1 in 19 477), a 50% chance that *f* is less than 0.000694 (1 in 1442), and a 95% chance that *f* is less than 0.00299 (1 in 334). This prior distribution has a mean of 1 in 1000, meaning that the first patient tested in the experimental screening program has a 1 in 1000 chance of having a false-positive result.

The second column of **Figure 2** shows the probability of observing the results of the experimental screening program—no false-positive results in 13 000 tests—as a function of the false-positive rate. In Bayesian statistics, this is called the *likelihood function*. The function values are calculated by using the binomial probability formula described in the Appendix. Here we see that this experimental result would be very unlikely if the false-positive rate were, say, 1 in 1000 or greater, but not at all surprising if the false-positive rate were, say, 1 in 100 000.

The third column shows the corresponding posterior distributions. After updating the base-case prior distribution with the observed data, we find that there is a 5% chance that *f* is less than 0.00000366 (1 in 272 922), a 50% chance that *f* is less than 0.0000495 (1 in 20 197), and a 95% chance that *f* is less than 0.000214 (1 in 4673); the revised mean is 1 in 14 000. As would be expected, the evidence of no false-positive results in 13 000 trials has shifted

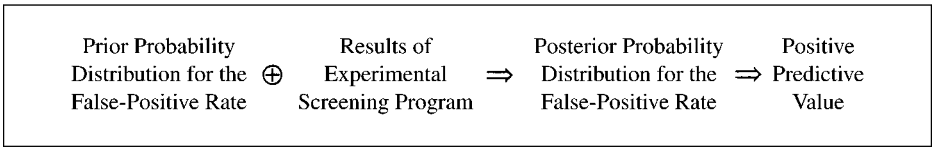
Prior Probability Distribution for the False-Positive Rate
⊕
Results of Experimental Screening Program
⇒
Posterior Probability Distribution for the False-Positive Rate
⇒
Positive Predictive Value

Figure 1. A Bayesian approach.

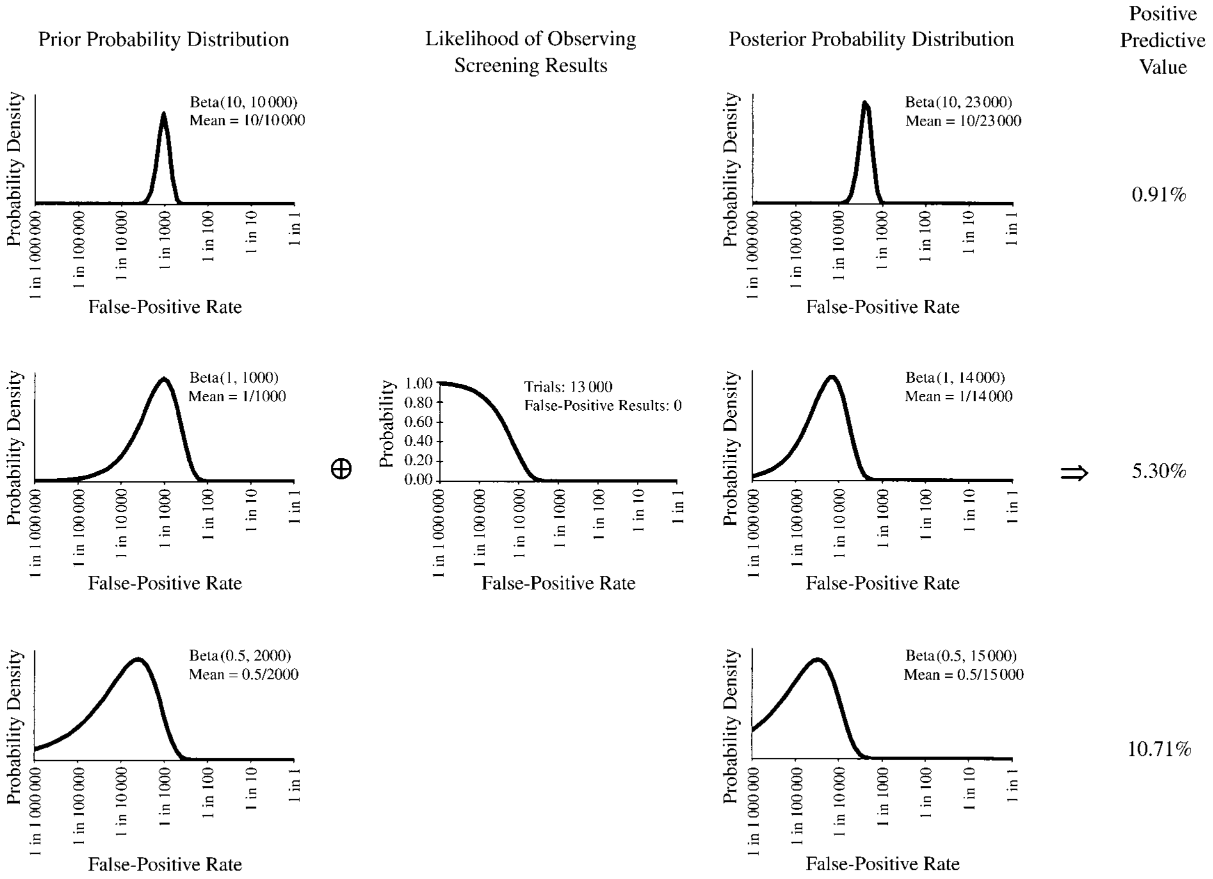

Figure 2. Results of the Bayesian procedure using different prior distributions.

the distribution to the left, making low values of *f* more likely and high values of *f* less likely.

The final column of **Figure 2** shows the resulting PPV. Applying the PPV formula with the mean of this posterior distribution (maintaining our earlier assumption about the prevalence and assuming perfect sensitivity), we find that Casey has a 5.3% probability of having the disorder. This calculated PPV lies within the range of values calculated earlier (1.7% to 100%) but is clearly toward the left end of this range.

The other rows in **Figure 2** correspond to alternative choices of prior distributions. The prior distribution in the top row assumes the same prior estimate of the false-positive rate (1 in 1000) but with greater confidence in this estimate, as shown by the narrower prior distribution. The corresponding posterior distribution is narrower than the base-case distribution, and the mean does not change as much after seeing the results of the screening trial (the mean is 1 in 2300 instead of 1 in 14 000). The higher posterior estimate of the false-positive rate leads to a lower PPV. The bottom row corresponds to a case in which we have a lower initial estimate of the false-positive rate (1 in 4000) and more uncertainty about the estimate. This leads to a posterior distribution that is shifted further to the left (with a mean of 1 in 30 000) and a higher PPV. Note that the results of the experimental screening program effectively rule out false-positive rates greater than 1 in 1000. The posterior probability mass in this region is shrunk nearly to zero (because the likelihood is essentially zero), with the mass assigned in other regions increasing in response.

In **Figure 3**, we used our base-case prior distribution in all cases but varied the outcome of the experimental trials; consequently, different likelihood functions resulted. The top row shows the results that would have occurred if we had seen only 100 newborns without a false-positive result, instead of 13 000. The second row shows what would happen if we had tested 100 000 newborns without a false-positive result. Comparing these two cases with the middle row of **Figure 2**, we see that the greater the number of tests without a false-positive result, the lower the posterior estimate of the false-positive rate and the greater the resulting PPV. In the final row, we consider a scenario in which 1 false-positive result occurred in 13 000 newborns tested; in this case, the PPV roughly halves (decreasing from 5.3% to 2.7%). The variations in PPVs in these examples suggest that in problems like these, it is important to keep clinicians up to date on previous test results so that when they compute PPVs for their patients, they have access to the best possible estimate of the false-positive rate. Testing centers might consider, for example, creating a database or World Wide Web site that keeps current counts of false-positive results on new tests and reports current estimates of these rates for use by clinicians.

Once a test has been used for a while and several false-positive results have been obtained, the estimated false-positive rates and PPVs become less sensitive to the assumed prior distribution and will change less in response to observing a single false-positive result. If, for example, we had seen 100 false-positive results in 1.3 million tests, it would make little difference which prior distribution was assumed (as long as it was not too unreasonable), and observing one more false-positive result would have little effect. However, when tests are highly specific and the indicated conditions are rare, it may take a long time to accumulate this much experience.

Conclusions

It is difficult to estimate PPVs intuitively for cases involving rare disorders, and it is a good idea to rely on formulas for calculating PPVs (11). But when we know little about the false-positive rate, formulas for PPVs, coupled with classic estimation procedures, do not give much guidance. In Casey's situation, for example, using a maximum likelihood estimate gives a PPV of 100%, and a 95% confidence interval leads to PPVs ranging from 1.7% to 100%. The Bayesian approach provides a logically correct way to arrive at a specific probability that can be used in clinical settings. This probability reflects the results of the experimental screening program to date as well as outside information about the biochemical or genetic methods used in the test, captured in the prior distribution.

Although we focused on the use of this Bayesian approach in updating false-positive probabilities, we could use the same approach to update beliefs about false-negative rates or the prevalence of a condition. In both of these extensions, we would proceed in the same manner as we did when considering uncertainty about the false-positive rate: We assess prior distributions on the rates and then update these distributions as we observe results. To determine the predictive value of the test at any time, we use the mean of the posterior distribution for the rate as the current estimate and calculate predictive values in the usual way. The Bayesian approach could also be used to describe uncertainty

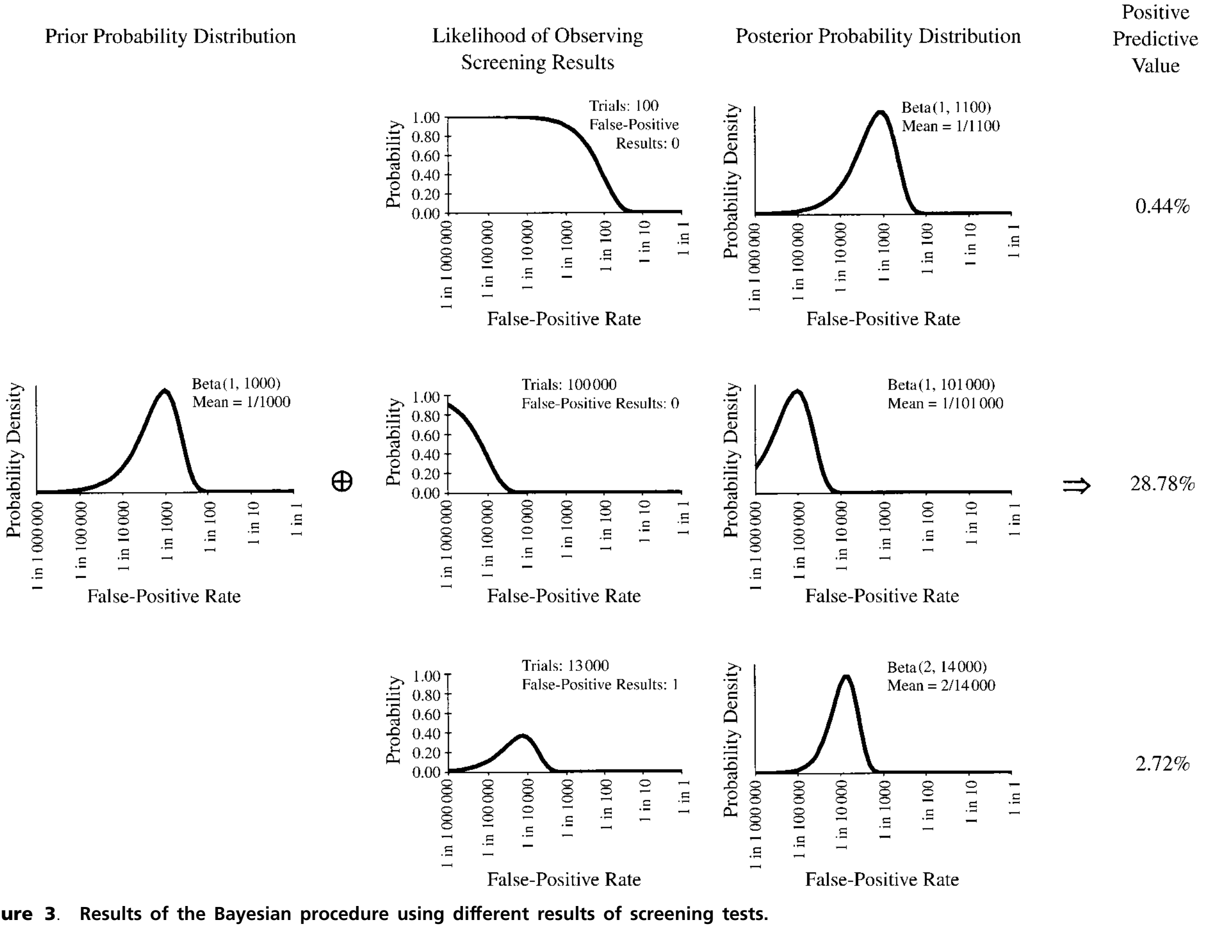

Figure 3. **Results of the Bayesian procedure using different results of screening tests.**

about failure rates more generally (for example, the probability of adverse effects of new drugs or complications of new surgical procedures).

The dependence of this approach on prior beliefs may cause distress for some; we would like to think and are used to thinking that these rates are clearly and objectively deduced from observed results. But when dealing with new tests—highly specific tests for rare disorders in particular and new procedures in general—the data alone may not provide sufficient guidance for patient management until a substantial track record accumulates, which may take a very long time. In these settings, the Bayesian approach is helpful in providing specific estimates of these rates and probabilities. As argued by Davidoff (12), once one gets past the "associations of hazy prior probabilities and abstruse mathematical formulas [that] strike fear into the hearts of most of us," we realize that this is how it should be: Outside information and prior knowledge should play a role in our assessments of accuracy, and failure to take such information into account can easily lead to serious misinterpretation of the evidence. Ultimately, we find that Bayesian reasoning is not only helpful but also necessary in obtaining reasonable and defensible estimates of these rates for use in treatment decisions.

By the way, Casey is fine. Follow-up tests showed that she does not have the suspected disorder—Casey was the first false-positive.

Appendix

Here, we present the formulas used to generate the numeric results summarized in **Figures 2** and **3**. For more detailed discussion of these formulas, see references 13 and 14. The binomial probability formula gives the probability of observing exactly r false-positives in n trials (tests of patients without the condition) and assuming a probability f of seeing a false-positive result in each trial, as

$$P_B(R = r \mid n, f) = \frac{n!}{r!(n-r)!} f^r(1 - f)^n - r.$$

The likelihood function in **Figure 2** is given by taking $n = 13\,000$ and $r = 0$ and varying f. In this context, the Bayes theorem says that if we start with a prior distribution with a probability density $p'(f)$ for the false-positive rate and observe r false-positive results in n trials, the posterior probability density $p''(f)$ is given by:

$$p''(f) = \frac{1}{\displaystyle\int_{f=0}^{1} P_B(R = r \mid n, f)\, p'(f)\, df} P_B(R = r \mid n, f)\, p'(f).$$

The Bayes formula can be applied with any prior distribution but is particularly easy to apply if the prior distribution is a beta distribution. Given parameters r' and n', the beta distribution assumes a probability density of the form

$$p'(f) = K_{r',n'} f^{r'-1}(1-f)^{n'-r'-1}$$

where $K_{r',n'}$ is a scaling constant that depends on r' and n' and is required to make the total probability under $p'(f)$ equal 1. (Note that r' and n' need not be integers.) The mean of the beta distribution with parameters r' and n' is given by r'/n' and probabilities (for example, the probability of a false-positive rate f being less than 1 in 1000) can be calculated by using built-in functions in popular spreadsheet programs (such as Microsoft Excel [15]). If we start with a beta distribution with parameters r' and n' and then observe r false-positive results in n trials, the posterior distribution is given by a new beta distribution with updated parameters $r'' = r + r'$ and $n'' = n + n'$. The posterior mean is $r''/n'' = (r + r')/(n + n')$. Intuitively, if we interpret the prior as having seen r' false-positives in n' trials, the posterior distribution can be interpreted as having seen $r'' = r + r'$ in $n'' = n + n'$ trials.

From Duke University, Durham, North Carolina; and University of Wisconsin-Madison, Madison, Wisconsin.

Grant Support: In part by the National Science Foundation under grants SBR 98-09176 (Dr. Smith) and SES 98-18855 (Dr. Winkler).

Requests for Single Reprints: James E. Smith, PhD, Fuqua School of Business, Duke University, Box 90120, Durham, NC 27708; e-mail, jes9@mail.duke.edu.

Requests To Purchase Bulk Reprints (minimum, 100 copies): Barbara Hudson, Reprints Coordinator; phone, 215-351-2657; e-mail, bhudson@mail.acponline.org.

Current Author Addresses: Drs. Smith and Winkler: Fuqua School of Business, Duke University, Box 90120, Durham, NC 27708.
Dr. Fryback: Department of Preventive Medicine, 785 WARF Building, 610 North Walnut Street, Madison, WI 53705-2397.

References

1. **Smith JE, Winkler RL.** Casey's problem: interpreting and evaluating a new test. Interfaces. 1999;29:63-76.
2. **McNeil BJ, Keller E, Adelstein SJ.** Primer on certain elements of medical decision making. N Engl J Med. 1975;293:211-5.
3. **Griner PF, Mayewski RJ, Mushlin AI, Greenland P.** Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med. 1981;94(4 Pt 2):557-92.
4. **Sox HC, Blatt MA, Higgins MC, Marton KI.** Medical Decision Making. Boston: Butterworth–Heinemann; 1988.
5. **Hanley JA, Lippman-Hand A.** If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983;249:1743-5.
6. **Blyth CR.** Approximate binomial confidence limits. Journal of the American Statistical Association. 1986;81:843-55.
7. **Pauker SG, Kassirer JP.** The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109-17.
8. **Spetzler CS, Stael von Holstein CA.** Probability encoding in decision analysis. Management Science. 1975;39:176-90.
9. **Morgan MG, Henrion M.** Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. New York: Cambridge Univ Pr; 1990.
10. **Clemen RT, Winkler RL.** Combining probability distributions from experts in risk analysis. Risk Analysis. 1999;19:187-203.
11. **Eddy DM.** Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. Judgment under Uncertainty: Heuristics and Biases. New York: Cambridge Univ Pr; 1982:249-67.
12. **Davidoff F.** Standing statistics right side up. Ann Intern Med. 1999;130: 1019-21.
13. **Winkler RL.** Introduction to Bayesian Inference and Decision. New York: Holt, Rinehart and Winston; 1972.
14. **Berry DA.** Statistics: A Bayesian Perspective. Belmont, CA; Duxbury Pr: 1996.
15. Microsoft Excel 2000. Redmond, WA: Microsoft Corp.; 1999.