What Is the Value of a p Value?

Gary L. Grunkemeier, PhD, YingXing Wu, MD, MS, and Anthony P. Furnary, MD

Medical Data Research Center, Providence Health & Services, Portland, Oregon

Successful publication of a research study usually requires a small p value, typically $p < 0.05$. Many clinicians believe that a p value represents the probability that the null hypothesis is true, so that a small p value means the null hypothesis must be false. In fact, the p value provides very weak evidence against the null hypothesis, and the probability that the null hypothesis is true is usually much greater than the p value would suggest. Moreover, even considering "the probability that the null hypothesis is true" is not possible with the usual statistical setup and requires a different (Bayesian) statistical approach. We describe the Bayesian approach using a well-established diagnostic testing analogy. Then, as a practical example, we compare the p-value result of a study of aprotinin-associated operative mortality with the more illuminative interpretation of the same study data using a Bayesian approach.

(Ann Thorac Surg 2009;87:1337–43)

"P values are a practical success but a critical failure. Scientists the world over use them, but scarcely a statistician can be found to defend them. Bayesians in particular find them ridiculous, but even the modern frequentist has little time for them [1]."

A value of $p < 0.05$ is usually considered essential for the success of scientific studies, ensuring the publication of research reports and enabling the advancement of academic careers. But p values do not provide a good measure of evidence against the null hypothesis. You would be forgiven if you think that they do, because p values as similar as $p = 0.06$ and $p = 0.04$ can make grown men weep, or academic careers flourish, respectively. A recent article arguing that p values exaggerate the evidence against the null hypothesis [2] begins with these two attention-getting quotations: "The most important task before us in developing statistical science is to demolish the p-value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology" [3], and "My personal view is that p values should be relegated to the scrap heap and not considered by those who wish to think and act coherently" [4]. To complete the process of getting your attention on this important and widely misunderstood issue, here is another: "The null hypothesis significance test should not even exist, much less thrive as the dominant method for presenting statistical evidence . . . It is intellectually bankrupt and deeply flawed on logical and practical grounds" [5]. This last author cited 33 references to support his statements.

A p-Value Primer

Statistics is not a unified science [6]. There are fundamentally different approaches whose advocates argue, on philosophical and epistemological grounds, about their relative merits [5, 7, 8]. The typical study collects data to investigate a possible difference in an outcome variable that is caused by a risk factor or intervention. The statistical conclusions are reached indirectly—using inductive reasoning—by "disproving" a null hypothesis. The null hypothesis usually states that there is no difference in outcomes induced by the risk factor or intervention tested, whereas the purpose of the study is usually to establish that a difference does exist by rejecting the null hypothesis.

The classical statistical approach (called "frequentist") produces a p value, putatively to measure the evidence that the study data provides against the null hypothesis, with the smaller the p value, the more evidence for rejection. The p value is the probability of observing (1) the study data plus (2) data even more extreme than that actually observed, given that the null hypothesis is true; and, in most cases, it also includes the probability of (3) observations equally extreme that are in the other (opposite) direction (two-sided test).

A True Story

We are among those fortunate biostatisticians who regularly interact with scientifically sophisticated research surgeons. So, when one of them makes a statistical misstatement, we assume that many other clinicians would make the same mistake and consider it an opportunity for an educational effort. Recently, it happened again: on a conference call in which we participated, one of our senior surgeons was discussing a statistical test of significance that resulted in a p value of 0.08, and he commented that "it is not significant, but, still, there is only an 8% chance that the null hypothesis is true." The vice president for medical science of a large pharmaceutical company, to whom he was talking, agreed.

A p value is *not* the probability that the null hypothesis is true, although many other clinicians apparently believe this interpretation [2, 9, 10]. The p value is not nearly that informative. Only a Bayesian type of analysis, the statistical paradigm arising from Bayes' theorem [11], can

Address correspondence to Dr Wu, 9205 SW Barnes Rd, Ste 33, Portland, OR 97225; e-mail: yingxing.wu@providence.org.

Published by Elsevier Inc

0003-4975/09/$36.00
doi:10.1016/j.athoracsur.2009.03.027

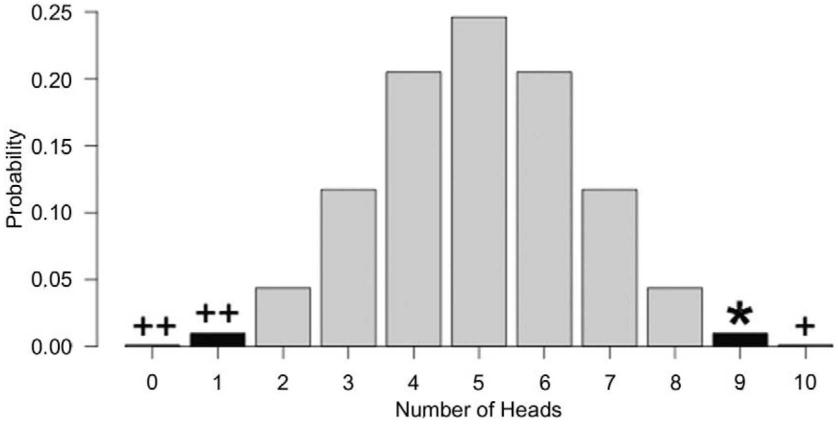

Fig 1. Probability distribution for a coin toss experiment: the probabilities of the number of heads observed in 10 tosses, resulting in 1024 possible (ordered) patterns. Our fictional story observed 9 heads (*), and the p *value for the experiment was the sum of the probabilities of 0, 1, 9, and 10 heads (black bars). The + signifies a probability of 0.001 that 1 of 1024 patterns would have all 10 heads. The ++ signifies 9 tails and 10 tails.*

provide a probability statement about the null hypothesis. But the commonality of the above misinterpretation shows that clinicians desire to know this probability, and hence are "Bayesians" at heart. To facilitate a compact explanation, we must begin with a small bit of technical jargon and related notation.

Conditional Probability

A conditional probability is one that is modified by an "if . . ." or a "given that . . ." condition. The *p* value is a conditional probability: It is the probability of observing the observed data (plus other data that is at least as extreme as that observed) given that the null hypothesis (Ho) is true. This can be written in a compact notation: *p* value = *Prob*(data | Ho), where *Prob* means probability, and the vertical line means *given.* In words, this equation says that "the *p* value equals the probability of observing the data if the null hypothesis is true." This is not the same as the inverse probability: *Prob*(Ho | data), the probability that the null hypothesis is true given the observed data, as our surgeon proclaimed it to be.

To easily appreciate that the quantities on the opposite sides of the "given" symbol (vertical bar) in a conditional probability cannot be reversed, consider this simple example. The probability that a given cardiothoracic (CT) surgeon is female, *Prob*(female | CT surgeon), is about 2% [12]. But the inverse probability, that a given female is a CT surgeon, *Prob*(CT surgeon | female), is certainly much smaller than 2%. An even more elementary example is the Dormouse's comment at the Mad Hatter's tea party in *Alice's Adventures in Wonderland* [13]*:* "I breathe when I sleep" is not the same as "I sleep when I breathe."

A Fictional Story

Probability concepts are often illustrated with examples from games of chance—not inappropriately, because the desire for gambling success sponsored the birth of probability theory [14]. In the familiar coin toss experiment, a fair coin is defined as one with a 50% probability of heads (Ho—the null hypothesis). Suppose you undertook an experiment to determine whether a particular coin was fair by tossing it 10 times, and the result was 9 heads. Is that enough evidence to reject this hypothesis and declare the coin biased? If so, you would be willing to pay a high price for it to make money by using it to win future bets. The *p* value from this experiment—the probability of getting 9 or more heads or tails, from a fair coin—is $p = 0.02$ and is easy to compute by deductive or direct reasoning using simple combinatory principles (Appendix 1, Fig 1). Thus, because $p < 0.05$, the difference is statistically significant, and the null hypothesis of a fair coin is rejected. Given this, our surgeon friend would say, "The probability that the coin is fair is only 2%. It must be biased, so, yes, let's buy it."

But as we have seen, the probability of Ho (that the coin is fair) is not equal to the *p* value; it is a bit more difficult to compute, because its determination requires using Bayes' theorem and an estimate of the prior probability. To show how this is done, though, we must abandon this coin toss example, and switch to another story. Why? Well, it turns out that it is not physically possible to make a biased coin if the coin toss is done properly [15]. So, even though the *p* value in our coin toss experiment was 0.02, which provides putative "significant" evidence against the null hypothesis, the null hypothesis is in fact true. The coin is not unfair (biased), because biased coins do not exist in nature. So, a rare event occurred, that is all: to infer that this coin is unfair/biased would be deception.

Diagnostic Testing for Coronary Artery Disease

Because we must abandon the coin toss example, where shall we turn to continue our evaluation (devaluation) of the *p* value? We have intimated that Bayesian analysis is required to produce the desired inverse probability. So let us move to a well-accepted clinical application of Bayesian reasoning—diagnostic testing for coronary artery disease (CAD)—and take advantage of its close connection with hypothesis testing [16, 17]. We will examine a series of patient scenarios to determine the information required to reach an appropriate conclusion from a diagnostic test about the probability that a patient has CAD. This will give us insight into the information—and methodology—required to reach an appropriate conclusion from a hypothesis test regarding the probability of the truth of the null hypothesis.

Patient A

Suppose a patient tests positive for CAD with a new diagnostic test that has a 95% specificity. Specificity is the (conditional) probability that a healthy person tests negative: *Prob*(Negative | Healthy). On the basis of this positive test, knowing only that the specificity is 95%, would you assume that your patient has CAD, reject the null hypothesis that the patient is healthy, and perform a revascularization surgery? Of course you would not.

First, you do not know the sensitivity of the test, the probability that a diseased person will test positive. To see why this matters, take the extreme case in which the sensitivity is zero; that is, the test never produces a positive result when the patient has CAD. This would mean that your patient has to be healthy because the only positives are false-positives (there is never a true-positive because the sensitivity is zero). Such a crippled test is not realistic, but the thought experiment should convince you that the sensitivity of the test matters.

Let us see how this relates to interpreting a *p* value by invoking the parallel relationship between diagnostic testing and hypothesis testing [16, 17]. The probability of a false-positive result (in terms of diagnostic) equals 1-specificity (0.05 in this case of 95% specificity). This is the analogue of (in hypothesis-testing terms) the probability of a type 1 error (Appendix 2). So, judging the evidence against the null hypothesis (declaring statistical significance) based on $p = 0.05$ alone is analogous to accepting a positive diagnostic test result based on its 95% specificity alone.

Patient B

Suppose a patient tests positive for CAD with a new diagnostic test that has 95% specificity and 90% sensitivity. Sensitivity is the analogue of power in a hypothesis testing setup, the probability of finding a significant difference when it exists. So you now have a test with a *p* value of 0.05 and a power of 90%. Are you going to reject the null hypothesis, declare the patient diseased, and perform the surgery?

No, not yet. It is well known that the proper interpretation of a diagnostic test must incorporate the prevalence of the disease, the patient's pretest probability of disease [18, 19]. Sensitivity and specificity are attributes of the diagnostic test. But more important for patient management is the positive predictive value (PPV) of the test, the probability that a patient who has tested positive has the disease: PPV = *Prob*(Disease | Positive). And the PPV depends on the prevalence of disease in the tested individual (Appendix 2). Thus, to guide proper decision making after a diagnostic test, it is essential to know the patient's pretest probability of disease as well as the sensitivity and specificity of the test [20]. These can be combined, using Bayes' theorem, to provide the PPV (Appendix 2). In terms of testing the hypothesis, disease prevalence is analogous to the prior probability that the null hypothesis is true.

The probability that the null hypothesis is true after statistical significance has been declared is given by the analogue of 1-PPV: *Prob*(Healthy | Positive), and not by the *p* value—*Prob*(Positive | Healthy) [21]. Let us see how the prior probability (prevalence) of disease affects the assessment of a positive diagnostic test.

Patient C1

Suppose your patient tests positive for CAD with a new diagnostic test that has 95% specificity and 90% sensitivity, and that CAD has 50% prevalence in the population to which this patient belongs. Then, using the formula in Appendix 2, PPV = 95%, so that *Prob*(Healthy | Positive) = 1–PPV =.05, and the probability that the null hypothesis is true does, in fact, equal the *p* value (1–specificity). But this unlikely scenario is the only one in which this equivalence exists.

Patient C2

Suppose your patient tests positive for CAD with a new diagnostic test that has 95% specificity and 90% sensitivity, and that CAD has a more realistic prevalence of 20% in the population to which this patient belongs. Then, 1–PPV = 0.18, which is almost four times larger than the *p* value (1–specificity) of 0.05.

Patient C3

Finally, suppose your patient tests positive for CAD with a new diagnostic test that has 95% specificity and 90% sensitivity, and that CAD has 5% prevalence in the population to which this patient belongs. Then 1–PPV = 0.51; that is, the patient is slightly more likely to be healthy than diseased, and the null hypothesis is more likely to be true than false, even though the *p* value is 0.05.

So, the rarer the disease, the greater the probability that your patient is healthy (the null hypothesis is true), despite a positive result from a very good diagnostic test, with a 95% specificity ($p = 0.05$) and 90% sensitivity.

Bayes to the Rescue

Besides offering the proper paradigm for interpreting diagnostic tests, the Bayesian approach is equally essential in evaluating the results of clinical studies. The main objection is that there is a subjective element. A prior distribution needs to be specified (prevalence) for the variable of interest, before the study begins—and one study's prior prevalence might be different than another's—yet, good science should be completely objective. But Bayes' methods are more practical, more logically sound, and are becoming more widely used in many disciplines.

The very fact that our surgeon asks the question about the probability of the null hypothesis being true, after the data are collected, indicates he must therefore be prepared to consider its probability before the study is undertaken, which is what the prior distribution is and which can only be considered by invoking the Bayesian statistical approach. Therefore, our surgeon must be a Bayesian by the very fact that he wants to know the probability of the null hypothesis being true.

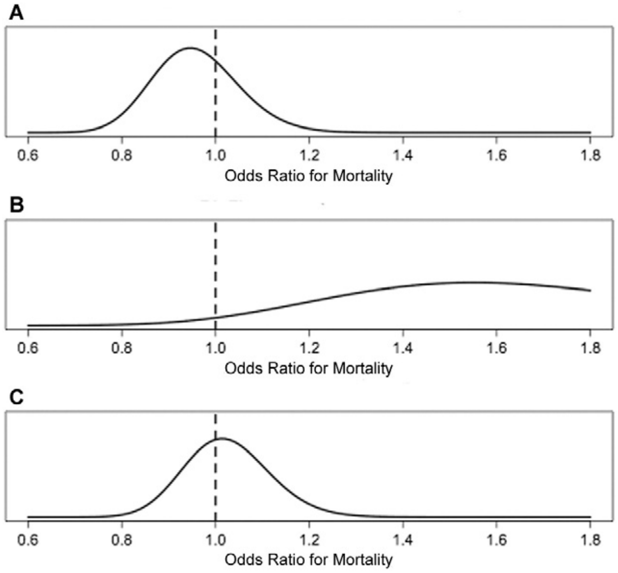

Fig 2. The three elements of a Bayesian analysis of the BART (Blood conservation using Antifibrinolytics in a Randomized Trial) data. Values to the left of the vertical dashed line (odds ratio = 1) favor lower mortality for aprotinin, and values to the right of the vertical dashed line favor lower mortality for aminocaproic acid. The three elements are (A) an initial (prior) estimate of the distribution of the variable of interest; (B) a summary of the study data, called the likelihood (the probability of the data, as a function of the study parameter), and (C) combining (integrating) A and B to produce a final (posterior) estimate of the distribution of the study parameter.

Operative Mortality With Aprotinin

The essence of the Bayesian approach is that the purpose of an experimental study is to modify current beliefs rather than to be interpreted in complete isolation of preexisting knowledge and experience. We are allowed (obligated) to interpret current study findings in light of previous knowledge, just like we all do in everyday life when we interpret new evidence in light of prior experience.

To exemplify, we will reexamine the results of a recent study of the antifibrolytic drug aprotinin (Trasylol, Bayer, Germany) used in cardiac operations to control bleeding. We believe that the recent attacks against aprotinin, which have been successful in removing it from the market, may have been statistically unsound. The final nail in the coffin was the BART (Blood conservation using Antifibrinolytics in a Randomized Trial) study, which claimed that aprotinin was associated with increased operative death [22].

The Statistical Analysis section of that article states that, "For the secondary outcomes of death and serious adverse events, we conducted pairwise chi-square tests to ascertain the relation between aprotinin and tranexamic acid and between aprotinin and aminocaproic acid" [22]. Yet, surprisingly, the results of these tests were not given: nowhere in the Results section or tables or supplementary online material are any χ^2 tests found.

So, we performed these tests, using mortality data from Table 3 in the BART Supplementary Appendix. When the differences between the mortality rates were tested using the Pearson χ^2 test, with continuity correction, no significant differences were found ($p < 0.05$). The p value for aprotinin vs aminocaproic acid was 0.08, based on 1559 patients, and it was this p value that our surgeon was discussing in the conversation mentioned earlier, that motivated this article.

Bayesian analysis proceeds through three steps:

A. determining an initial (prior) estimate of the distribution of the parameter of interest;
B. producing a summary of the study data, called the likelihood (the probability of the data, as a function of the study parameter), and
C. combining (integrating) A and B to produce a final (posterior) estimate of the distribution of the study parameter [23–25]. (Note: the curves in Fig 2 were derived using the log odds ratio [OR] scale because it more closely follows the normal distribution [25], but are plotted on the [untransformed] OR scale for easier interpretation.)

A. Prior Distribution: Merged Cardiac Registry Study

Our estimate of the prior probability of mortality risk comes from the Merged Cardiac Registry (MCR) data that we used [26] to refute the claims of an earlier study concerning renal failure [27]. Of the 23,105 MCR patients, 22% received aprotinin and 42% received aminocaproic acid (patients who received both drugs were not included). We now reuse this data set to investigate the mortality risk associated with aprotinin compared with aminocaproic acid in the 14,887 patients who received just one of these two drugs. The preoperative expected mortality risk of the aprotinin patients was almost double that of the aminocaproic acid patients (Table 1), indicating that the nonrandom allocation of these drugs was heavily skewed towards the use of aprotinin in the riskiest patients.

Because the MCR patients were not randomized to these two drugs, we used the MCR risk model for operative mortality (http://www.healthdataresearch.com/cardiacrisk.htm) to risk-adjust the mortality comparison. This model had good discrimination, with a c-index (area under the receiver operating characteristic curve) of 0.80 for these patients. The logarithm of the odds of the observed/expected (O/E) risk was used as an offset term [28] to derive the risk-adjusted OR for each drug (compared with expected), and for their comparison to each other. Compared with expected mortality, both the O/E ratio and the OR were slightly less than 1 for both drugs, with lower values for the aprotinin patients (Table 1). And aprotinin was slightly protective for death (OR, 0.93) compared with aminocaproic acid (Table 1; Fig 2A).

B. Likelihood: BART Data

The OR for aprotinin in the BART study is 1.55, and the 95% confidence interval (barely) includes the value 1 (Table 2). Because this was a randomized study, no risk-adjustment was done. The likelihood function (probability of the BART data, as a function of the OR) exhibits wider spread than does the MCR data because there are only about 10% as many BART patients as MCR patients (Fig 2B).

Table 1. Merged Cardiac Registry Operative Mortality, 2000 to 2006

	Aprotinin	Aminocaproic Acid
Operations, No.	5193	9694
Observed mortality, %	5.2	2.8
Expected mortality, %	5.6	2.9
O/E mortality ratio	0.93	0.97
OR[a]	0.91	0.96
OR (95% CI)[b]	0.95 (0.78–1.14)	

[a] Odds ratio, for each drug, of the observed mortality vs expected mortality. Computed separately for each drug group, by exponentiationing the intercept term from a simple logistic regression with the logit of the expected mortality as an offset term, and no other risk factors [28].

[b] Odds ratio of aprotinin vs aminocaproic acid after being risk adjusted for varying expected mortality in the two groups. Computed using logistic regression of the combined group, with an indicator variable for aprotinin, and with the logit of expected mortality as an offset term.

CI = confidence interval; O/E = observed/expected; OR = odds ratio.

C. Posterior Distribution (Integrate A and B)

Using Bayes' theorem, we derived the posterior distribution of the OR by multiplying the prior distribution with the likelihood, and the result (Fig 2C) shows that there is no evidence for a mortality difference between aprotinin and aminocaproic acid: the posterior distribution has a mean OR of 1.01. It is this distribution that (finally) allows us to make probability statements about the variable of interest, in this case the OR. For example, the probability is 95% that the true OR is between 0.85 and 1.21. This 95% probability interval is called a credible interval in Bayes' terminology. It looks just like a confidence interval (CI), but whereas the (frequentist) CI has a convoluted and unappealing definition, "A 95% CI will contain the true value on 95% of occasions if a study were repeated many times using samples from the same population," the definition of a Bayesian credible interval is simply that "the probability is 95% that the true value is in the interval." The latter is what we originally hoped to obtain as a result of our study and, again, lobbies for the conclusion that researchers are Bayesians at heart.

Comment

Even before you started reading this expose of the overrated p value, you must have wondered about some of its readily apparent shortcomings:

1. Any small difference, no matter how clinically unimportant, will be statistically significant ($p < 0.05$) if the sample size is large enough.
2. Any large difference, no matter how clinically important, will be not be statistically significant ($p > 0.05$) if the sample size is too small.
3. Because of 1 and 2, a low p value in a small study is more evidential than the same p value in a large study [29]. Moreover, the effect of publication bias may be greater for small studies.
4. The p values add to the probability of the outcome that was observed (eg, 9 heads) the probability of all outcomes more extreme (eg, 10 heads), even though they were not observed.
5. The p values are usually two-tailed, meaning that they also include the probabilities of other outcomes that were not observed, in the opposite direction (eg, 9 tails and 10 tails), that are at least as extreme as the outcome that was actually observed [30].
6. The arbitrary, yet entrenched, threshold level of 0.05 creates a false dichotomy between significant and not significant. The significance value should not be fixed, but should depend on the consequences of the resulting decisions.
7. Some journals no longer accept p values [31], and many insist on estimation (CIs) rather than hypothesis testing [32].
8. For a continuous end point, the probability of the usual simple null hypothesis, that the difference between treatment means is zero, is actually zero; there is virtually no chance that the treatments would be exactly the same.
9. If one generates multiple hypothesis tests, the resulting p values are anticonservative and the p value for significance must be adjusted upward. This is controversial and can be shown to be somewhat silly [33]. Should a statistician adjust for the number of p values she produced for this study, or during this month, or in her lifetime?
10. The p value is not the probability that the null hypothesis is true, although it is often interpreted this way. Nor is it the probability that the alternative hypothesis is false—or any other such desirable information. The purpose of this article is to emphasize this point and to show how such (desired) probabilities can be obtained, using another statistical paradigm—Bayesian analysis.

By demonstrating the effect that disease prevalence has on the interpretation of a positive diagnostic test, we aimed to convince the reader that without information about the (prior) probability of the null hypothesis, a significant ($p < 0.05$) hypothesis test has little value in disproving the null hypothesis. It has the same value as the result of a positive diagnostic test whose only known property is 95% specificity. For practical use with a given patient, the PPV is of more importance than the specificity of a diagnostic test. The significance-testing analogue to this statement would be that the posterior probability of the null hypothesis is more important than the p value. The probability that the null

Table 2. BART Operative Mortality, 2002 to 2007

	Aprotinin	Aminocaproic Acid
Operations, No.	779	780
Observed mortality, %	6.0	4.0
OR (95% CI)[a]	1.55 (0.97–2.47)	

[a] Odds ratio of aprotinin vs aminocaproic acid. CI computed using a formula for the log OR [25], and transforming back to the OR metric.

BART = Blood conservation using Antifibrinolytics in a Randomized Trial; CI = confidence interval; OR = odds ratio.

hypothesis is true is usually much larger than the *p* value. This is a well-described, but perhaps not widely known, phenomenon. A classic, 50-year-old article [34] begins by stating that a test of a null hypothesis can result in $p < 0.05$, while the posterior probability that the null hypothesis is true is as high as 95%.

What about a rejoinder? The fact that *p* values have become so ubiquitous in reporting research results implies that there must be another side to the argument, something good to say in their favor. The case in favor of *p* values is difficult to make. A fairly balanced short article, but with 107 supporting references, concludes that, "Thus, despite the fact that *p* values are dead and buried (by some journals), we would agree . . . that significance tests are 'alive and well'." [35]. A longer, thoughtful and balanced assessment, with commentaries, gives "heavily qualified" support, and uses a nice phrase that sums up the situation: "the test of significance gives significance too easily" [1]. (This article's title is "Two cheers for *P*-values?," but the author's rejoinder to the commentaries suggests that "One cheer" is perhaps more appropriate.)

For the substantive data result of this report, we did not accept the *p* value ($p = 0.08$) of the BART study at face value, but rather used the Bayesian formulation to modify our previous beliefs. We derived the posterior probability for the OR of aprotinin by evaluating 1559 patients from the Bart study in the context of prior knowledge from a study of 14,887 patients. A potential limitation of this analysis is that the risk model used to obtain the OR (Table 1) may not have controlled for all confounders. A more complete analysis would consider the comparability of the patients in the MCR and BART studies, and perhaps partially discount the former in comparison with the latter, so that a patient in the MCR study would contribute less influence to the posterior OR distribution than a patient in the BART study; that is, more variability would be introduced into the prior distribution. The conclusion, in this case, however, should not be dramatically altered.

We started with a quote, and will end with another one:

> When writing for Epidemiology, you can also enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals discourages them, and every worthwhile journal will accept papers that omit them entirely. In Epidemiology, we do not publish them at all [31].

References

1. Senn S. Two cheers for P-values? J Epidemiol Biostat 2001; 6:193–204; discussion 205–10.
2. Hubbard R, Lindsay RM. Why P values are not a useful measure of evidence in statistical significance testing. Theor Psychol 2008;18:69–88.
3. Nelder JA. From statistics to statistical science. Statistician 1999;48:257–69.
4. Lindley DV. Comment on Bayarri and Berger. Bayesian Stat 1999;6:75.
5. Gill J. The insignificance of null hypothesis significance testing. Political Research Quarterly 1999;52:647–74.
6. Berger JO. Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science 2003;18:1–12.
7. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. American Scientist 1988;76:159–65.
8. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995–1004.
9. Carver RP. The case against statistical significance testing. Harvard Educational Review 1978;48:378–99.
10. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. Journal of the American Statistical Association 1987;82:112–22.
11. Bayes T. An essay towards solving a problem in the doctrine of chances. [Reprint of the original article which appeared in Philos Trans Roy Soc London 1763:53;370–418.]. Biometrika 1958;45:295–315.
12. Roberts SR, Kells AF, Cosgrove DM 3rd. Collective contributions of women to cardiothoracic surgery: a perspective review. Ann Thorac Surg 2001;71:S19–21.
13. Carroll L. The complete, fully illustrated works. New York: Gramercy Books, 1995.
14. Ore O. Pascal and the invention of probability theory. American Mathematical Monthly 1960;67:409–19.
15. Gelman A, Nolan D. You can load a die, but you can't bias a coin. American Statistician 2002;56:308–11.
16. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med 1983;98:385–94.
17. Lauer MS. Believability of clinical trials: a diagnostic testing perspective. J Thorac Cardiovasc Surg 2006;132:249–51.
18. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med 1979;300:1350–8.
19. Diamond GA, Forrester JS, Hirsch M, et al. Application of conditional probability analysis to the clinical diagnosis of coronary artery disease. J Clin Invest 1980;65:1210–21.
20. Moons KG, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003;10:670–2.
21. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
22. Fergusson DA, Hebert PC, Mazer CD, et al. A comparison of aprotinin and lysine analogues in high-risk cardiac surgery. N Engl J Med 2008;358:2319–31.
23. Pocock SJ, Spiegelhalter DJ. Domiciliary thrombolysis by general practitioners. BMJ 1992;305:1015.
24. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to bayesian methods in health technology assessment. BMJ 1999;319:508–12.
25. Spiegelhalter DJ, Abrams K, Myles J. Bayesian approaches to clinical trials and health-care evaluation. In: Senn S, ed. Statistics in practice. West Sussex, UK: Wiley; 2004:391.
26. Furnary AP, Wu Y, Hiratzka LF, Grunkemeier GL, Page US 3rd. Aprotinin does not increase the risk of renal failure in cardiac surgery patients. Circulation 2007;116:I127–33.
27. Mangano DT, Tudor IC, Dietzel C. The risk associated with aprotinin in cardiac surgery. N Engl J Med 2006;354:353–65.
28. Grunkemeier GL, Wu Y. What are the odds? Ann Thorac Surg 2007;83:1240–4.
29. Royall RM. The effect of sample size on the meaning of signficance tests. American Statistician 1986;40:313–5.
30. Peace KE. The alternative hypothesis: one-sided or two-sided? J Clin Epidemiol 1989;42:473–6.
31. Rothman KJ. Writing for epidemiology. Epidemiology 1998; 9:333–7.
32. Uniform requirements for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. N Engl J Med 1997;336:309–15.
33. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6.
34. Lindley DV. A statistical paradox. Biometrika 1957;44:187–92.
35. Moran JL, Solomon PJ. A farewell to P-values. Crit Care Resusc 2004;6:130–7.

Appendix 1

Deduction: Coin Toss Experiment

There are two possible results—heads (H) or tails (T)—for the first coin; 2 × 2 = 2^2 = 4 possibilities for the first 2 coins (HH, HT, TH, TT); 2 × 2 × 2 = 2^3 = 8 possibilities for the first 3 coins, and so on; up to 2^{10} = 1024 possible patterns for all 10 coins. Only 10 of these patterns would consist of exactly 1 tails (1 in each position, if you think of the tossed coins as lined up in a row) and thus 9 heads. So the probability of 9 heads is 10/1024 = 0.010 (marked by the asterisk in Fig 1).

But the p value also includes probabilities of events more extreme than that observed, in this example, 10 heads. Only 1 of the 1024 patterns would have all 10 heads, with probability 0.001 (the + in Fig 1). So the probability of 9 or more heads is 0.011. And the (usual) two-tailed p value also includes the probabilities of events in the other direction—in this case, 9 tails and 10 tails (++ in Fig 1). By symmetry, this doubles the p value to 0.022.

Appendix 2

Induction: Bayes' Theorem in Diagnostic Testing

1. *Similarity of diagnostic and hypothesis tests:*

	True Condition	
Diagnostic test	Healthy	Diseased
Positive	False positive (FP)	True positive (TP)
Negative	True negative (TN)	False negative (FN)

	True Condition	
Hypothesis test	Null hypothesis	Alt hypothesis
Significant	Type 1 error	
Not significant		Type 2 error

2. *Definition of conditional probability:*
Probability of A given B = *Prob*(A | B) = *Prob*(A and B)/*Prob*(B)
3. *Definitions of diagnostic test attributes:*
Sensitivity (true-positive rate) is the probability of a positive test result given the presence of disease: Sensitivity = *Prob*(Positive | Diseased). Using the definition of conditional probability, we can derive sensitivity from the values in the table above as TP/(TP + FN).
Specificity (true negative rate) is the probability of a negative test results given the absence of disease: Specificity = *Prob*(Negative | Healthy). Specificity can be derived from the values in the Appendix Table as TN/(TN + FP).
4. *Positive predictive value:*
The PPV = *Prob*(Diseased | Positive) is more important for patient management but cannot be derived from the values in the Appendix 2 Table; it requires knowledge of the prevalence of disease. To appreciate this, consider the Appendix 2 Table and think of it filled in with the 4 numbers resulting from, say, 100 healthy patients (in the first column) and 100 diseased patients (in the second column). Using these 4 numbers, PPV = TP/(TP + FP). But suppose we use the same test on another population, with 1000 healthy patients and 100 diseased patients. Then, the numbers in the cells of the first column would be 10 times larger than before, so the above equation for PPV would give a totally different answer for this same diagnostic test. To determine the PPV, then, requires use of Bayes' theorem.
5. *Bayes' theorem:*
Prob(A | B) = *Prob*(B | A) × *Prob*(A)/*Prob*(B)
6. *Derivation of PPV using Bayes' theorem:*
PPV = *Prob*(Diseased | Positive) = *Prob*(Positive | Diseased) × *Prob*(Diseased)/*Prob*(Positive), and
Prob(Positive) = *Prob*(Positive | Disease) × *Prob*(Disease) + *Prob*(Positive | Healthy) × *Prob*(Healthy).
Thus, in terms of the test attributes plus disease prevalence, PPV = Sensitivity × Prevalence/[(Sensitivity × Prevalence) + (1–Specificity) × (1–Prevalence)]