Large, randomized clinical trials (“megatrials”) are key drivers of modern cardiovascular practice, since they are cited frequently as the authoritative foundation for evidence-based management policies. Nevertheless, fundamental limitations in the conventional approach to statistical hypothesis testing undermine the scientific basis of the conclusions drawn from these trials. This review describes the conventional approach to statistical inference, highlights its limitations, and proposes an alternative approach based on Bayes’ theorem. Despite its inherent subjectivity, the Bayesian approach possesses a number of practical advantages over the conventional approach: 1) it allows the explicit integration of previous knowledge with new empirical data; 2) it avoids the inevitable misinterpretations of p values derived from megatrial populations; and 3) it replaces the misleading p value with a summary statistic having a natural, clinically relevant interpretation—the probability that the study hypothesis is true given the observations. This posterior probability thereby quantifies the likelihood of various magnitudes of therapeutic benefit rather than the single null magnitude to which the p value refers, and it lends itself to graphical sensitivity analyses with respect to its underlying assumptions. Accordingly, the Bayesian approach should be employed more widely in the design, analysis, and interpretation of clinical megatrials. 

What used to be called judgment is now called prejudice, and what used to be called prejudice is now called a null hypothesis . . . . It is dangerous nonsense (dressed up as the “scientific method”) and will cause much trouble before it is widely appreciated as such. — A. W. F. Edwards (Cambridge University Press, 1972) 

The randomized trial is the apotheosis of scientific progress in clinical medicine (1–4). Presently, more and more investigators are employing this tool in larger and larger study populations to identify smaller and smaller differences between treatment groups (5–13). These so-called “megatrials” have thereby become key drivers of modern medical practice, since they are cited frequently as the authoritative foundation for evidence-based management policies. Nevertheless, the published reports of these trials persistently fail to interpret the observations in the context of relevant background information—our prior convictions—relying almost exclusively instead on the conventional p value as the operative standard of scientific inference (14). This lapse is all the more troubling because these very same trials serve to reveal fundamental limitations in the inferential process itself, which, although presaged for some time (15–19), have had little practical consequence until the advent of the megatrial era. Without exaggeration, if this process is undermined, so too is the scientific basis of cardiovascular practice. Yet, this issue has never been addressed in the cardiovascular literature (17–24).

Accordingly, we herein: 1) review the process of scientific inference from a clinician’s perspective—with particular reference to the cardiovascular megatrial—outlining the inherent limitations of the prevailing statistical paradigm and the rationale in support of an alternative Bayesian approach; 2) describe ways to implement this Bayesian approach by integrating the trial data with relevant background information; and 3) suggest actions to encourage the adoption of this new exemplar by clinical investigators, journal editors, and practitioners alike.

From the Division of Cardiology, Cedars-Sinai Medical Center, and the School of Medicine, University of California, Los Angeles, California.

Manuscript received October 1, 2003; revised manuscript received January 2, 2004, accepted January 12, 2004.
reduction is determined to be statistically significant ($\chi^2 = 5.6, p = 0.02$). The investigators thereby concluded that treatment A is superior to treatment B, and that the magnitude of risk reduction is clinically important, because the CI for the OR includes the 0.78 threshold value. Shortly after the study was published, B. A. Zion, Professor of Clinical Epistemology at New Haven University, submitted a letter to the editor—impolitely entitled “FISHy Conclusions”—arguing that the data are consistent instead with about a 10% chance that the observed risk reduction is clinically important, as well as a 25% chance that the two treatments are actually equivalent! What is the basis for these contradictory interpretations?

Just as many questions in cardiology require us to know something of the relevant laws of physics (for instance, the rules governing fluid pressure and flow), this question requires us to know something of the relevant principles of logic (the rules of evidence). As we shall see, the controversy here stems from two rival views of scientific inference—as profoundly different as luminal narrowing and plaque instability in the pathophysiology of atherosclerotic events—and because most of us have never received formal instruction regarding these views, we must begin with a brief synopsis.

Our investigators’ stylized conclusions are grounded on R. A. Fisher’s time-honored theory of statistical inference (25). Fisher recognized that deductive hypotheses, such as if $a$ then $b$, can be refuted with certainty by so much as a single observation of $a$ and not $b$, but that statistical hypotheses, such as if $a$ then $b$ with probability $c$, cannot be refuted by any number of observations. He responded to this difficulty by positing that a statistical conjecture (what he called the “null hypothesis”) should be “rejected,” instead, by an observation that is unlikely, relative to all other possible observations, on the assumption of that conjecture (25). His famous $p$ value (the tail area under a frequency distribution representing the null hypothesis) was the evidentiary measure that provided a quantitative rationale for this judgment. As he expressed it, a small $p$ value means, “Either an exceptionally rare chance has occurred or the [null hypothesis] is not true” (25).

Fisher’s argument is roughly that of a deductive syllogism:

- If the null hypothesis is true, then the observations are unlikely
- The observations occurred.

Therefore, the null hypothesis is unlikely.

But if this argument sounds right to you, consider its parallel:

- If Tom has hypertension, then he is unlikely to have pheochromocytoma.
- Tom has pheochromocytoma.

Therefore, he is unlikely to have hypertension.

This faulty reasoning is identical to that used to characterize a patient as abnormal, just because some diagnostic test result falls outside its putative normal range—a one-dimensional strategy equivalent to relying solely on the specificity (or its complement, the false-positive rate) of the test (17,18). Thus, although Fisher’s approach has been supremely influential, critics charge he never provided it with a fully objective foundation (16,19).

Neyman and Pearson (26) sought to overcome this difficulty by testing the null hypothesis, not in isolation, as did Fisher, but in comparison to one or more alternative hypotheses. To do so, they defined a new test statistic (the ratio of the likelihood of the observations given the null hypothesis to the likelihood of the observations given the alternative hypothesis), and used Fisher’s approach to determine if this “likelihood ratio” exceeded some threshold at predefined false-positive ($\alpha$) and false-negative ($\beta$) levels of error. If so, they argued, then the null hypothesis was to be rejected, not by way of Fisher’s inductive logic, but on pragmatic grounds that “… in the long run of experience, we shall not often be wrong” (27).

This so-called “frequenstist” approach is the same as that used to classify a patient as abnormal whenever the true-positive rate of some diagnostic test result is greater than its false-positive rate (28). Although this two-dimensional strategy did succeed in providing a rationale for some of Fisher’s arbitrary choices, it did not really circumvent the subjectivity inherent in the process of statistical inference.

### Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>FRISC</td>
<td>Fragmin and Fast Revascularization during InStability in Coronary artery disease trial</td>
</tr>
<tr>
<td>HPS</td>
<td>Heart Protection Study</td>
</tr>
<tr>
<td>LIFE</td>
<td>Losartan Intervention for Endpoint reduction in hypertension trial</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio</td>
</tr>
<tr>
<td>PURSUIT</td>
<td>Platelet Glycoprotein IIb/IIIa in Unstable Angina Receptor Suppression Using Integrilin Therapy trial</td>
</tr>
<tr>
<td>RITA</td>
<td>Randomized Intervention Treatment of Angina trial</td>
</tr>
<tr>
<td>TACTICS</td>
<td>Treat Angina with Aggrastat and determine Cost of Therapy with an Invasive or Conservative Strategy-Thrombolysis In Myocardial Infarction-18 trial</td>
</tr>
<tr>
<td>TIMI-IIIB</td>
<td>Thrombolysis In Myocardial Infarction-IIIB trial</td>
</tr>
<tr>
<td>VANQWISH</td>
<td>Veterans Affairs Non–Q-wave Infarction Strategies in Hospital trial</td>
</tr>
</tbody>
</table>

### Table 1. Primary Outcomes in the FISH Trial

<table>
<thead>
<tr>
<th>Event</th>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>430 (8.6%)</td>
<td>500 (10%)</td>
</tr>
<tr>
<td>No event</td>
<td>4,570 (91.4%)</td>
<td>4,500 (90%)</td>
</tr>
</tbody>
</table>
(29) (for example, the 20% relative risk reduction that went into the sample size determination [30] for the FISH trial).

The founding fathers were well aware of such subjective influences. Fisher acknowledged that his calculations were “. . . absurdly academic . . .” and that the prudent scientist “. . . rather gives his mind to each particular case in the light of the evidence and his ideas” (25). Likewise, Pearson freely admitted that he and Neyman (31):

left in our mathematical model a gap for the exercise of a more intuitive process of personal judgement in such matters. . . as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities.

Nonetheless, the frequentist school has since come to sweep these matters under the carpet in its rush to venerate a single metric—the iconic p value—both as Neyman-Pearson’s “long run” error rate and Fisher’s “rare chance” evidentiary measure (never mind that the two interpretations are mutually inconsistent) (23).

LIMITATIONS OF THE CLASSIC APPROACH

This p value is usually computed from some amalgam of the observations (such as z or t or χ²). The z statistic, for example, is formulated as the mean difference in outcome between two groups divided by the standard error of the mean:

\[ z = \frac{x_A - x_B}{\sqrt{s^2_A/n_A} + \sqrt{s^2_B/n_B}} \]

where \(x_A\) and \(x_B\) are the mean values for groups A and B; \(s^2_A\) and \(s^2_B\) are their standard deviations; and \(n_A\) and \(n_B\) are their sample sizes.¹

Frequentist summary statistics such as this behave badly when applied to clinical megatrials. Because the sample size appears as a reciprocal in the denominator of the above equation, for example, the value of \(z\) will increase with the size of the trial for any non-zero numerator. Consequently, the p value (the tail area for \(z\) under the null hypothesis) will become arbitrarily small as the sample size becomes arbitrarily large (15). Eventually, even the smallest difference in outcome cannot escape the pull of a “statistical black hole” fueled by a sufficient mass of patients.² Carried to the extreme, everything becomes “significant” in a trial of infinite size.

This is no idle speculation. Just as a normal heart can fail if the imposed stress is great enough, any difference in outcome, however trivial in magnitude, will become “statistically significant” if the clinical trial is large enough, as with the 1.4% absolute difference among 10,000 subjects in our FISH trial. Smaller p value thresholds (e.g., 0.005 vs. 0.05) will postpone, but not prevent, the problem. In practical terms, then, some trials may have to be large, but never too large.

Even if the p value were numerically well behaved, it would nevertheless remain deeply misleading. Technically, the p value quantifies the probability of having obtained the data (or even more extreme, unobserved data), assuming the null hypothesis is true. However, what we really want to know is the inverse or “posterior” probability that the null hypothesis is true given the data that were observed. Many believe—or act as if they believe—the p value represents this more relevant posterior probability (17). But it does not!

The probability that “Tom is hypertensive given that he has pheochromocytoma” is not the same as the inverse probability that “Tom has pheochromocytoma given that he is hypertensive.” Likewise, the probability of observing a difference in outcome (p < 0.05) given that treatments A and B are equivalent is not the same as the probability that treatments A and B are equivalent given the observed difference in outcome (hence, our fallacious syllogisms). Simply stated, the “bassackward” p value provides the right answer to the wrong question.

The right question is, “What do you know about hypothesis \(b\) after seeing evidence \(e\)?”, and the p value is the wrong answer to this question. The right answer (the posterior probability for \(b\) given \(e\)) clearly cannot be based on \(e\) alone, but must depend also on one’s answer to the more primitive question, “What did you know about \(b\) before seeing \(e\)?” (the prior probability for \(b\)).

As a matter of fact, specific neurons in the parietal cortex physically encode and process such prior probabilities (32) by the time we are four years of age (33). However, the frequentist (like a sentencing judge who overlooks the prior convictions of a habitual criminal) ignores these signals. This “historical blindness” is particularly disabling with regard to megatrials for which prior information is usually abundant.

ADVANTAGES OF A BAYESIAN APPROACH

Bayes’ theorem resolves this spectrum of problems (19,29). It can be expressed succinctly by the following relation:

\[ p(b|e) \propto p(e|b) \times p(b) \]

In words, the probability for the hypothesis given the evidence (the “posterior”) is proportional to the probability for the evidence given the hypothesis (the “likelihood”) times the probability for the hypothesis independent of the evidence (the “prior”). This seminomial relationship—a straightforward consequence of the fundamental axioms of probability theory³—bridges Pearson’s aforementioned “gap,” by connecting the evidentiary observations to the historical context within which they occur. Scientific infer-

¹When \(n_A\) and \(n_B\) are large, swapping their values in this equation provides an expression in which \(z^2 = t^2 = \chi^2\).

²If \(n_A = n_B\), this “mass” is given by \(n = 2\sigma^2 / d^2\), where \(\sigma\) is the pooled variance (\(\sigma^2_A + \sigma^2_B\)) and \(d\) is the difference in outcome (\(x_A - x_B\)).
ence, like common sense, is thereby seen to rely equally on the background information and the empirical data.

However, there is a price to be paid for this gain. To a Bayesian, probabilities represent degrees of belief rather than real-world frequencies (29), even those expressed in terms of ratios (34) or distributions (35) of empirical counts, and because our beliefs are not always based on (objective) data, they often come from the (subjective) mind of the observer. Now, if different observers have different prior beliefs, they will have different posterior beliefs given the same set of data. These subjective prior beliefs are anathema to the frequentist, who relies instead on a series of ad hoc algorithms that maintain the facade of scientific objectivity, even while taking similar liberties apropos Pearson’s “gap” (31).

Thus, the frequentist first calculates the value of one or another test statistic quantifying the degree to which the observations deviate from those expected under the null hypothesis (χ² = 5.6 for FISH, based on Table 1), then estimates the frequency of observing at least this value in numerous imaginary repetitions of the experiment under that hypothesis (p = 0.02 for FISH, analogous to the 4% false-positive rate for ≥1.5 mm exercise-induced electrocardiographic ST-segment depression for diagnosis of coronary artery disease [36]), and “rejects” the hypothesis if this p value fails to reach some arbitrary threshold (e.g., α = 0.05). Harold Jeffreys, a pupil of Fisher’s and the first to develop a fundamental theory of scientific inference based on Bayes’ theorem, summarizes this convoluted reasoning process by noting that (37):

A hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. (italics as in the original)⁴

Instead, the Bayesian calculates the likelihood of the observations with respect to the test hypothesis and then multiplies this likelihood by a prior probability to obtain the posterior probability. In this context, ignoring the prior would be as much a failing as ignoring the data.

Using this approach, clinicians have come to appreciate that a diagnostic hypothesis cannot be properly assessed solely by reference to the one-dimensional specificity or two-dimensional likelihood ratio of some test, but only by a three-dimensional integration of the sensitivity and specificity with the probability of disease in the patient being tested (34).

Likewise, a scientific hypothesis cannot be properly assessed solely by reference to the observational data, but only through the integration of those data with one’s prior beliefs regarding the hypothesis. Bayes’ theorem is the formal means by which we perform this explicit integration—a logically consistent, mathematically valid, and intuitive way to draw inferences about the hypothesis in light of our experience (19,29).

In contrast, pure evidentiary metrics (such as p values, CIs, and likelihood ratios) are no more than compass headings. They tell us only where we are going—toward or away from some hypothesis—but not where we are.

Therefore, the straightforward Bayesian approach has a number of practical advantages over the convoluted conventional approach: 1) it eliminates the frequentist’s “historical blindness,” thereby facilitating the integration of prior knowledge with new empirical data; 2) it replaces the “bassackward” p value with a measure having true clinical relevance—the probability for the study hypothesis given the observations; and 3) it skirts the “statistical black hole” resulting from large samples, thereby forestalling erroneous inferences. Additional advantages are summarized in Table 2.

In summary, the operative standard of scientific inference (the frequentist p value) is undermined by a variety of theoretical and practical shortcomings. Its failings call into question the published conclusions of many highly influential clinical megatrials (38–40), thereby echoing a recent New York Times claim that, “half of what doctors know is wrong” (41). Cynics might well acknowledge these failings, but argue nonetheless that our polemic is directed at a straw man—that no one really relies on p values to the exclusion of other important factors. Indeed, investigators often entertain a number of Bayesian-like assumptions in the course of a clinical trial (such as the 20% threshold for clinical importance [42] in FISH), but they usually do so only to estimate the sample sizes required for calculating the p values expected of them by the statisticians, journal editors, and reviewers. Editorialists similarly enlist a number of Bayesian-like considerations in their post hoc commentaries on these trials, but this is usually done to explain away conflicts between the empirical results and their own preconceived notions (43). Lacking legitimate ways to characterize the truth of their hypotheses, how would any of them ever come to learn which half of what they “know” is wrong?

### Table 2. Frequentist Versus Bayesian Attributes of Randomized Clinical Trials

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior information</td>
<td>Excluded</td>
<td>Included</td>
</tr>
<tr>
<td>Subjective influences</td>
<td>Implicit</td>
<td>Explicit</td>
</tr>
<tr>
<td>Theoretical foundation</td>
<td>Inconsistent</td>
<td>Consistent</td>
</tr>
<tr>
<td>Randomization</td>
<td>Essential</td>
<td>Incidental</td>
</tr>
<tr>
<td>Sample size</td>
<td>Prespecified</td>
<td>Unrestricted</td>
</tr>
<tr>
<td>Alpha and beta errors</td>
<td>Prespecified</td>
<td>Unnecessary</td>
</tr>
<tr>
<td>Stopping rules</td>
<td>Prespecified</td>
<td>Unnecessary</td>
</tr>
<tr>
<td>Test hypothesis</td>
<td>Prespecified</td>
<td>Unrestricted</td>
</tr>
<tr>
<td>Summary measure</td>
<td>p value</td>
<td>Posterior probability</td>
</tr>
<tr>
<td>Clinical interpretation</td>
<td>Misleading</td>
<td>Straightforward</td>
</tr>
</tbody>
</table>

### INTEGRATING PRIOR BELIEFS WITH EMPIRICAL DATA

Bayes’ theorem is the heart of this learning process by which we update our existing beliefs (the prior) with new information (the data). Thus, just as medical diagnosis begins with the clinical history, learning begins with the prior; and just as the history begins from ignorance so too does that prior (29,37,44). Accordingly, when the component risks

---

⁴Recall Tweedledum’s demonstration of logic to Alice: “If it was so, it might be; and if it were so, it would be; but as it isn’t, it aint.
about the null value for clinically unimportant
ing some "mum ignorance. We can mirror these constraints by de
sources, they are rarely initiated under conditions of maxi-
Moreover, because megatrials are very demanding of re-

\( \sigma = 0.03 \)


\( \sigma = 0.07 \)


\( \sigma = 0.40 \)

\( \rho_1 = 1 - q_2 \) and \( \rho_2 = 1 - q_1 \) are proportionately low, we can employ the OR \( \rho_1 / \rho_2 \) or its Gaussian transform—

\( \log \text{OR} \) as an estimate of relative risk \( \rho_1 / \rho_2 \) \([45-48]\), and thereby model our initial state of ignorance with respect
to the typical null hypothesis by a uniform distribution for
log OR (with mean \( \mu = 0 \) and standard deviation \( \sigma = 1 \)).

Few investigators, however, are inclined to believe that
the null effect will be exactly zero, believing instead that
the effect might be so small as to be clinically unimportant.
Moreover, because megatrials are very demanding of re-
sources, they are rarely initiated under conditions of maximum
ignorance. We can mirror these constraints by defining
some “clinically unimportant” interval of equivalence
about the null value for \( x_p \) \((\pm 5\% \text{, for example})\) \([47]\), and varying \( \sigma_p \) so as to adjust the proportion of the distribution
falling within that interval \([24]\), in accordance with our
beliefs (Fig. 1). Alternatively, we can derive the parameters of
the prior distribution \( (x_p, \sigma_p) \) from previously available
data \([46]\), just as we determine the parameters of the
empirical distribution \( (x, \sigma) \) from the current trial
data.\(^5\)

Now that we have independent determinations of a prior
distribution for the log OR based on our beliefs before
consideration of the trial data, \( G(x_p, \sigma_p) \), and an empirical
distribution based on the trial data alone, \( G(x, \sigma) \), we can

\[ \frac{x}{a}, \frac{\sigma}{c} \]

\[ a - \frac{1}{c} \]

\[ a + \frac{1}{c} + 1/d \]

The mean log odds ratio \( \mu \) is \( \ln(ab/cd) \), and its standard deviation \( \sigma \) is \( (1/a + 1/b + 1/c + 1/d)^{1/2} \).

\[ G(x, \sigma) = G(x_p, \sigma_p) \times G(x, \sigma) \]

\[ G(-0.07, 0.05) = G(0, 0.07) \times G(-0.12, 0.06) \]

We can use the resultant posterior distribution to quan-
tify the probability for any interval therapeutic response (the
area under the curve between putative limits of interest) or

\[ x_p \pm \frac{\sigma_p}{\sigma} \]

\[ \sigma = (\frac{1}{\sigma_p} + \frac{1}{\sigma})^{-1/2} \]

EMPIRICAL APPLICATIONS

OF THE BAYESIAN APPROACH

Sensitivity analysis. According to Bayes’ theorem, then,
our belief about the hypothesis after seeing the data depends
on our belief about the hypothesis before seeing the data.
This variable degree of belief stands in sharp contrast to the frequentist’s categorical interpretation of the p value as “significant” or “nonsignificant,” based on the data alone. Obviously, such variability will influence our subsequent inferences in material ways. We can determine the degree of this influence by performing graphical or tabular sensitivity analyses (17,44,46) similar to those employed by economists and decision theorists (49).

Table 3 summarizes representative sensitivity analyses for a spectrum of well-known cardiovascular trials (5–13,50), and Figure 4 illustrates one of these analyses (for the HPS [6]) graphically. Each of these trials—one of which (LIFE [8]) is quantitatively similar to our hypothetical FISH trial—reported a p value or CI for the comparison of some primary outcome in two randomized groups (A vs. B). Hence, the investigators formally entertained the null hypothesis—an implicit representation of clinical equipoise (51)—as the operative basis of their statistical analysis (even if this hypothesis might have conflicted with previously available data or their own personal beliefs). Accordingly, we determined the posterior probability for this null hypothesis given the empirical data, based on an uninformative and moderately skeptical prior.

In each case, the specific magnitude of posterior null probability is highly dependent on our particular choice of prior (the smaller the value of \( \sigma_p \), the more informative is that prior and the greater is its influence relative to the empirical data). With an uninformative prior, the posterior trial...
null is similar to the reported p value, but increases nonlinearly with more informative priors (as in Fig. 4). Using a moderately skeptical prior, the posterior null probabilities range widely (from near zero to over 30%), regardless of the empirical log ORs. As a result, our beliefs concerning these highly influential, statistically significant megatrials appear less confident than implied by the p values alone.

This is not to imply that the published conclusions regarding any of these trials are necessarily wrong (something no programmatic system of induction can do), but rather to highlight the potential for such errors. Bayesian analysis minimizes this potential by reinforcing the empirical evidence with the prior information. It does not guarantee that each of us will look at the same data and come to the same conclusion, but it does assure that we will do so if we begin with the same prior beliefs. It is in just this way that the Bayesian approach can be considered scientifically “objective.”

In the FISH trial, too, the posterior probability is highly dependent on our particular choice of prior. Using a moderately skeptical prior, the posterior probability for the (±5%) interval null hypothesis is 0.23 (recall B. A. Zion’s 25% chance of equivalence), but falls to 0.05 based on a mildly skeptical prior and rises to 0.81 based on a highly skeptical prior. Including such sensitivity analyses in published trial reports would serve to obviate any appearance that the investigators have gerrymandered these subjective parameters in support of a particular point of view.

**Magnitude of therapeutic response.** One of the most important advantages of Bayesian analysis is its ability to assess any magnitude of therapeutic response (i.e., the probability that the risk reduction exceeds some putative “threshold of benefit” given the observations), rather than the precise null magnitude to which the p value refers (i.e., the frequency of obtaining a risk reduction of at least the magnitude observed given that the true magnitude is 0) (47,52). Table 4 summarizes such threshold analyses for the same trials as those in Table 3, using an uninformative prior ($x_0 = 0, \sigma_p = 10$). In each case, the posterior probability for benefit falls as the threshold for benefit increases and is far less than that implied by conventional statistical significance.

Figure 5 illustrates a comparable analysis of therapeutic benefit for our hypothetical FISH trial, again using an uninformative prior. Although the chance of any degree of benefit (>0%) approaches 100% (consistent with the statistically significant p value of 0.02), the chance of >10% benefit is only 77%, and the chance of >20% benefit is no more than 13%. These values are summarized in Table 5, along with those for several more informative, skeptical priors.

This approach provides us with a clinically relevant numerical substitute for p values in the published reports of these trials. Recall that the FISH investigators assumed that the smallest clinically important risk reduction was 20%. If so, then the most relevant representation of the trial results is given by the posterior probability that the relative risk reduction exceeds this putative threshold. As noted earlier, the value of this probability is 0.13, using an uninformative prior (and would be even less for more informative priors, as shown in the bottom row of Table 5). Thus, despite a

**Table 4. Posterior Probability of Benefit for Representative Clinical Trials**

<table>
<thead>
<tr>
<th>Trial</th>
<th>% RRR (95% CI)</th>
<th>RRR &gt;0%</th>
<th>RRR &gt;10%</th>
<th>RRR &gt;20%</th>
<th>RRR &gt;30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PURSUIT (5)</td>
<td>9 (0–18)</td>
<td>0.979</td>
<td>0.581</td>
<td>0.033</td>
<td>0.000</td>
</tr>
<tr>
<td>HPS (6)</td>
<td>12 (6–18)</td>
<td>0.999</td>
<td>0.847</td>
<td>0.030</td>
<td>0.000</td>
</tr>
<tr>
<td>GUSTO (7)</td>
<td>14 (6–21)</td>
<td>0.994</td>
<td>0.858</td>
<td>0.089</td>
<td>0.000</td>
</tr>
<tr>
<td>LIFE (8)</td>
<td>14 (4–23)</td>
<td>0.997</td>
<td>0.876</td>
<td>0.252</td>
<td>0.003</td>
</tr>
<tr>
<td>HOPE (9)</td>
<td>21 (13–28)</td>
<td>0.999</td>
<td>0.999</td>
<td>0.844</td>
<td>0.092</td>
</tr>
<tr>
<td>BHAT (10)</td>
<td>26 (9–41)</td>
<td>0.998</td>
<td>0.974</td>
<td>0.826</td>
<td>0.420</td>
</tr>
<tr>
<td>4S (11)</td>
<td>28 (14–40)</td>
<td>0.999</td>
<td>0.996</td>
<td>0.926</td>
<td>0.547</td>
</tr>
<tr>
<td>MADIT II (12)</td>
<td>28 (8–44)</td>
<td>0.999</td>
<td>0.973</td>
<td>0.879</td>
<td>0.619</td>
</tr>
<tr>
<td>ASCOT-LLA (13)</td>
<td>35 (17–49)</td>
<td>0.999</td>
<td>0.996</td>
<td>0.958</td>
<td>0.757</td>
</tr>
<tr>
<td>Lyon Diet Heart (50)</td>
<td>68 (48–83)</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
</tr>
</tbody>
</table>

*Probability that the percent of relative risk reduction (%RRR) is greater than the putative threshold value for clinical importance (0%, 10%, 20%, 30%), based on an uninformative prior and the empirical data.
CI = confidence interval.
Bayesian Interpretation of Clinical Trials

Table 5. Posterior Probability of Benefit Based on One’s Choice of Prior

<table>
<thead>
<tr>
<th>Threshold of Benefit (%)</th>
<th>Uninformative Prior ($\sigma_p = 10$)</th>
<th>Skeptical Priors ($\sigma_p = 0.4$)</th>
<th>Skeptical Priors ($\sigma_p = 0.07$)</th>
<th>Skeptical Priors ($\sigma_p = 0.03$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0</td>
<td>0.992</td>
<td>0.991</td>
<td>0.957</td>
<td>0.831</td>
</tr>
<tr>
<td>&gt;10</td>
<td>0.769</td>
<td>0.750</td>
<td>0.263</td>
<td>0.001</td>
</tr>
<tr>
<td>&gt;20</td>
<td>0.128</td>
<td>0.110</td>
<td>0.001</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$\sigma_p =$ prior standard deviation.

statistically significant p value of 0.02—and contrary to the conclusion drawn by the investigators using CIs—there is little more than a 10% chance that the observed magnitude of benefit is clinically important (consistent, again, with B. A. Zion’s assessment).

This is just what we should have expected. Even if the observed risk reduction equaled the 20% threshold for clinical importance, this value represents the mean of a symmetrical Gaussian distribution. Thus, there would be only a 50% chance that the risk reduction exceeded this mean value and a 50% chance that it did not. However, because the observed risk reduction was only 14%, the chance of exceeding the 20% threshold is even less than this. In the final analysis, then, despite its impressive sample size and significant p value, FISH turns out to be a quantitative example of the rhetorical “distinction without a difference.”

**Bayesian meta-analysis.** By its nature, Bayesian analysis is particularly suited to the meta-analysis of clinical trials addressing a common hypothesis. The aggressive (“anatomy-driven”) versus conservative (“ischemia-driven”) management of acute coronary syndromes is a case in point. Over the past decade, five large, randomized trials have examined this issue in almost 9,000 patients (53). Results have been inconsistent—with the two older trials supporting a conservative approach (TIMI-IIIB and VANQWISH) and the three more recent trials (FRISC-II, TACTICS TIMI-18, and RITA-3) supporting an aggressive approach—predominantly with respect to surrogate outcomes such as recurrent ischemia and referral for revascularization. The impact on definitive outcomes such as death and myocardial infarction remains controversial; a recent meta-analysis reported a 12% reduction in relative risk for these events ($p = 0.04$), despite significant heterogeneity from study to study ($p = 0.005$) (54).

The top panel of Figure 6 illustrates a Bayesian meta-analysis of these studies, with respect to these definitive outcomes, in a sequence that parallels their dates of publication (54). The first trial (TIMI-IIIB) is analyzed using an uninformative prior given the absence of previous data. Thereafter, the posterior for the preceding trial serves as the prior for the subsequent trial. As illustrated in Figure 6, the second trial (VANQWISH) has a substantial negative impact on the probability of benefit given the limited amount of prior information (TIMI-IIIB) available at the time, but this is offset by subsequent trials (FRISC-II and TACTICS TIMI-18). Consequently, the most recent trial (RITA-3) has little effect on the posterior probability given the large amount of prior information available from the four trials preceding it. This meta-analysis indicates a 70% chance that the risk reduction is more than 10%, but only a 10% chance it is more than 20%. In other words, there is a 30% chance the risk reduction is under 10% and a 90% chance it is under 20%—values far different from that implied by a conventional meta-analysis (54) (summarized in the bottom panel of Fig. 6). Thus, although conventional meta-analysis shows that aggressive management is associated with a statistically significant reduction in death and myocardial infarction, Bayesian meta-analysis suggests that the magnitude of this reduction is unlikely to be clinically important.

**ENCOURAGING THE ADOPTION OF AN INTEGRATED APPROACH**

In the end, statistical inference—whether frequentist or Bayesian—can take us only so far. In fact, our clinical decisions are rarely based on subjective judgments or objective data alone, but rather on something between and beyond the two—the ethical doctrines that ultimately imbue the decisions with meaning and value.

Such valuations typically rely on the utilitarian principle advocating “the greatest happiness for the greatest numbers” (55). This principle is commonly applied to strategic decisions regarding health care policy. The current emphasis on clinical outcomes and prescriptive guidelines is a clear reflection of both its influence on modern medical practice and the importance of probabilistic reasoning to clinical decision-making. In this context, good decisions succeed in balancing the objective scientific data against our subjective ethical values; they are evidence-based, but not evidence-bound. This is more than metaphor. Our brains are actually hardwired to compute probabilities and utilities using the very same principles of game theory and decision analysis that describe rational economic behavior (32,56,57).

Several journals have taken a leadership position in the clinical application of these principles (58). The *Journal of the American Medical Association’s* decade-long series of “Users’ Guides to the Medical Literature” provides physicians with strategies and tools to interpret (59) and apply (60) such evidence in the care of their patients, and the *Annals of Internal Medicine’s* “Information for Authors” now includes specific recommendations that contributors (61):

... use Bayesian methods as an adjunct to frequentist approaches, ... state the process by which they obtained the
Despite this enlightened editorial endorsement, however, there are only 322 citations for the search string "Bayes*" among 374,747 "clinical trial" citations in the National Library of Medicine’s PubMed data base since the publication of Cornfield’s seminal 1969 paper proposing the application of Bayes’ theorem to clinical trial assessment (62) (as of January 12, 2004). In the last analysis, then, we would be well advised to develop academic, political, and economic incentives to encourage the diffusion of these recommendations into common practice.

We do not champion a particular means to this end. Instead, we advocate agencies such as the National Institutes of Health, Food and Drug Administration, Center for Medicare and Medicaid Services (formerly the Health Care Financing Administration), and Institute of Medicine to empanel a task force of experts along the lines of the Consolidated Standards of Reporting Trials (CONSORT) group (63) to perform this function. The task force—comprising clinicians, trialists, health outcomes researchers, epidemiologists, statisticians, journal editors, and policy makers—should be mandated to standardize the representations and choice of prior probability, as well as methods to integrate the posterior probability with the observed magnitude of treatment effect (e.g., absolute and relative risk reductions). The standards should be supported by scientific comparisons of previously published empirical data and by suitable computer simulations. Appropriately vetted statistical software instantiating these standards should be developed and disseminated via the Internet (64).

Large, randomized trials, as well as their subsequent meta-analyses, are highly demanding of resources and possess an aura of scientific respectability that almost ensures their publication in influential medical journals, even in the face of methodological deficiencies (39,65–67). For just these reasons, greater attention must be paid to explicitly quantifying the probability for the hypotheses being tested by these trials and the degree of credibility that their conclusions are to be accorded. Until then, evidence-based medicine will continue to rest more on the limitations of statistical inference than on the strength of the evidence itself.

None of this will happen overnight. Giants from Bayes and Laplace to Fisher and Jeffreys have debated the foundations of inductive logic for over 200 years without resolution, and our recondite comments are unlikely to change anyone’s prior convictions regarding these matters. More than a century ago, the eminent nineteenth century physicist James Clerk Maxwell suggested the real way such change comes about, in noting that, ‘‘change anyway,’’ perception, and our recondite comments are unlikely to change anyone’s prior convictions regarding these matters.

He was probably right (p < 0.05).

Acknowledgments
The authors gratefully appreciate the encouragement and constructive comments of three anonymous reviewers and several journal editors in our efforts to present these technical issues in a way that is comprehensible and relevant to the typical thoughtful clinician.


REFERENCES

Reprint requests and correspondence: Dr. George A. Diamond, 2408 Wild Oak Drive, Los Angeles, California 90068. E-mail: gadiamond@pol.net.

APPENDIX

A Microsoft Excel 2000 spreadsheet used to perform the analyses in this paper is available on the June 2, 2004, issue of JACC at www.cardiosource.com/jacc.html.