Placing Trials in Context Using Bayesian Analysis

GUSTO Revisited by Reverend Bayes

James M. Brophy, MD, Lawrence Joseph, PhD

Standard statistical analyses of randomized clinical trials fail to provide a direct assessment of which treatment is superior or the probability of a clinically meaningful difference. A Bayesian analysis permits the calculation of the probability that a treatment is superior based on the observed data and prior beliefs. The subjectivity of prior beliefs in the Bayesian approach is not a liability, but rather explicitly allows different opinions to be formally expressed and evaluated. The usefulness of this approach is demonstrated using the results of the recent GUSTO study of various thrombolytic strategies in acute myocardial infarction. This analysis suggests that the clinical superiority of tissue-type plasminogen activator over streptokinase remains uncertain.

(*JAMA*. 1995;273:871-875)

BEFORE any clinical trial results are available, different clinicians will have different opinions regarding the relative benefits of the therapies under study. These opinions will usually range from skepticism to enthusiasm for a new therapy compared with a standard therapy. Regardless of how well it is conducted, no single clinical trial can provide absolutely definitive conclusions. Thus, even after trial results are reported, it is reasonable to expect that a diversity of opinions will persist, although perhaps with some convergence toward the observed trial results. The degree of convergence will depend on the strength of the trial in terms of sample size and scientific rigor in its execution. Therefore, in any medical experiment, clinical researchers must give careful consideration to issues of both design and analysis. Randomized clinical trials are almost universally accepted as the gold standard design for comparative clinical research, since bias and confounding are minimized. Much attention has been directed to the scientific reasoning behind statistical analysis in the medical and statistical literature.[1-3] However, while most clinicians are aware of the importance of good experimental designs, few are aware of the full array of statistical methods available. Some of these methods allow for the reporting of a range of conclusions corresponding to the diversity of prior opinions. They can also answer directly questions of interest to clinicians.

Classical (frequentist) analysis is the most prevalent statistical method used, leading to the ubiquitous P values and confidence intervals. P values from research trials may be viewed as analogs of false-positive (1−specificity) diagnostic tests. If neither the disease nor the treatment is malignant, we may well accept test specificity of 95% (P=.05). However, before accepting a limb amputation for osteosarcoma, we would rightly demand a false-positive value much less than .05. Generally, we are more interested in knowing what is the probability of disease given the test result (analogous to predictive value), and this cannot be supplied from classical statistical considerations alone. Clinicians routinely interpret diagnostic test results in the "clinical context," that is, by considering the background rate of the disease in a given population. In a similar manner, the interpretation of clinical trials should be considered in the light of preexisting knowledge.[1] (The analogy between hypothesis testing and diagnostic testing is completed by noting that statistical power corresponds to the sensitivity of a diagnostic test.)

In the classical approach, model parameters such as population means are fixed (nonrandom) quantities and probability distributions are considered only for test statistics (such as the t statistic in a t test). The randomness of test statistics arises because frequentists must consider not only the observed data in a given experiment, but also other data that might have occurred had the experiment been repeated. Each of these hypothetical repetitions leads to a different value of the test statistic, and the collection of these form a distribution. It is this distribution that is used to calculate P values and confidence intervals.

Rather than directly addressing desired clinical questions, such as "Which treatment is superior?" or "What is the probability of a clinically meaningful treatment difference?," classical analysis usually examines the null hypothesis of no difference between the competing strategies. P values denote the probability that a statistic as extreme as or more extreme than the observed test statistic would occur on hypothetical re-

From the Department of Medicine, Centre Hospitalier de Verdun (Quebec) (Dr Brophy); Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec (Drs Brophy and Joseph); and Division of Clinical Epidemiology and Centre for the Analysis of Cost Effective Care, Department of Medicine, Montreal (Quebec) General Hospital (Dr Joseph).

Reprint requests to Department of Medicine, Centre Hospitalier de Verdun, 4000, Boul Lasalle, Verdun, Quebec, Canada H4G 2A3 (Dr Brophy).

Table 1.—Data From GUSTO, GISSI-2, and ISIS-3*

Trial	Agent	No. of Patients	No. (%) of Deaths	No. (%) of Nonfatal Strokes	Combined Deaths or Strokes
GUSTO†	SK	20 173	1473 (7.3)	101 (0.5)	1574 (7.8)
	t-PA	10 343	652 (6.3)	62 (0.6)	714 (6.9)
GISSI-2	SK	10 396	929 (8.9)	56 (0.5)	985 (9.5)
	t-PA	10 372	993 (9.6)	74 (0.7)	1067 (10.3)
ISIS-3	SK	13 780	1455 (10.6)	75 (0.5)	1596 (11.6)
	t-PA	13 746	1418 (10.3)	95 (0.7)	1513 (11.0)

*SK indicates streptokinase; and t-PA, tissue-type plasminogen activator.
†The 10 374 patients who received both SK and t-PA are not included here.

peated trials if the null hypothesis is exactly true. This raises two problems. First, it seems counterintuitive to base statistical inferences on events more extreme than those observed, since these events did not actually occur.[3] Second, one almost never believes that the null hypothesis of exact equivalence is true, and it is consequently usually more relevant to test for a range of equivalence. Such a test is very rarely carried out in practice. P values do not measure the true quantity of interest, namely, the probability that the null or alternative hypothesis is true. This contributes to the confusion between the information P values provide and the information that is more naturally desired. Therefore, it is not surprising that P values are often misinterpreted as the probability that the null hypothesis is true or that $1-P$ represents the probability that the alternative hypothesis is true. Classical statistical analysis does not directly or indirectly provide these probabilities.

Another inherent limitation of P values derives from their dependence on sample size. Basically, any difference, no matter how small, can reach statistical significance if the sample size is large enough. For example, an observed difference of only one tenth of a standard deviation will become statistically significant at the .05 level if each group in the trial includes at least 768 subjects and will be nonsignificant otherwise. On the other hand, it is well known that the low power accompanying small trials may lead to P values greater than .05 even when clinically meaningful effects are observed in the trial.[4]

All of these limitations of P values have prompted an increased use of confidence intervals. Many clinicians do not appreciate that a 95% confidence interval only means that with unlimited repeated experiments, 95% of all the confidence interval limits derived using similar procedures in different studies would contain the true parameter. While this may provide some comfort in the long run, little can be said about the likelihood that, for example, a given treatment is superior or that the true value of the parameter under current study lies in any particular interval.

The shortcomings of classical statistics may obscure the interpretation of even a well-designed and well-executed trial. For example, the recent GUSTO trial (Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Arteries) was a multicenter, randomized study comparing different thrombolytic regimens for the treatment of acute myocardial infarction.[5] This trial is of particular interest since there continues to be controversy over the clinical importance of any treatment differences. In addition, there have been other randomized trials involving large numbers of patients that examine the same question, namely, is tissue-type plasminogen activator (t-PA) superior to streptokinase (SK) in the treatment of acute myocardial infarction.[6,7] The question of therapeutic superiority is of considerable public health importance, since myocardial infarction is a frequent occurrence and t-PA is approximately 10 times more expensive than SK. While many critiques of the GUSTO trial have been published,[8-11] these have mostly centered on design issues and the interpretation of the clinical relevance of the observed mortality differences. This article raises further questions while highlighting some advantages of an alternative (Bayesian) statistical approach. Bayesian analysis has often been dismissed due to its "subjectivity" and because of computational difficulties. While Bayesian analysis can be computationally complex, computer algorithms now exist that make this hurdle more historical than contemporary. As will be seen, Bayesian subjectivity is an asset that can provide an ideal forum for debate, since prior beliefs, including clinical experience, must be formally specified, and one can directly observe how the beliefs are updated in the light of new data. This procedure permits the appreciation of the logic for various a posteriori opinions, which should tend to converge as data accumulate. This process is different from classical meta-analysis, which suffers from all the problems associated with P values and confidence intervals mentioned above and furthermore does not permit the incorporation of prior beliefs.[12]

METHODS

Model parameters such as the success rate of a given medical treatment are generally unknown, and therefore experiments are designed to provide information about their values. In virtually any well-designed experiment, more is known about these values after the experiment than before, although at least some information usually exists preexperimentally. A Bayesian statistical analysis is designed to represent this learning process.

The first step in any Bayesian analysis is to obtain a prior distribution over all model parameters. The prior distribution summarizes the preexperimental beliefs about the parameter values. This can be accomplished by using past data, if available, by drawing on expert knowledge, or by a combination of both. This step is nontrivial and can take considerable time and effort. Furthermore, many prior distributions are not unique; clinicians are free to summarize their beliefs into their own prior distribution. Because Bayesian methods can incorporate clinical opinion, they are often labeled "subjective." The experimental data are then used to update the prior distribution to a posterior distribution using Bayes' theorem. This is done through the likelihood function, which provides the probability of obtaining the observed data as a function of the unknown model parameter. This is analogous to using a likelihood ratio (sensitivity/[1−specificity]) to update background probabilities after observing results from a diagnostic test. The posterior distribution represents the postexperimental beliefs about the parameter values, given the new data and the previously stated prior distribution. The two main quantities of interest, namely, the probability that a given treatment is superior and the probability of a clinically meaningful effect, are both directly available from the posterior distribution. Unlike the standard approach, no references to data sets other than those observed are required, since all of the information contained in the data is summarized by the likelihood function.

No one prior distribution is likely to be sufficient to represent the diversity of clinical opinions that exists before a trial is carried out. Indeed, this diversity is usually a prerequisite for ethical randomization. Therefore, trial results should usually be reported starting from a range of prior distributions.[13] The corresponding set of posterior distributions

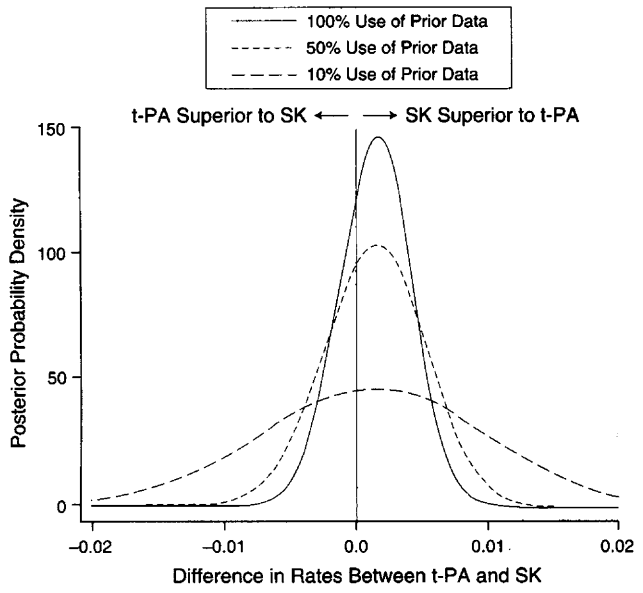

Figure 1.—Plot of the prior distributions for the difference in mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK) using weights of 100%, 50%, and 10% of the GISSI-2 and ISIS-3 data, representing a range in prior beliefs in the relevance of these trials to the GUSTO trial. The area under the curve between any two points on the x-axis is the posterior probability that the difference in mortality rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

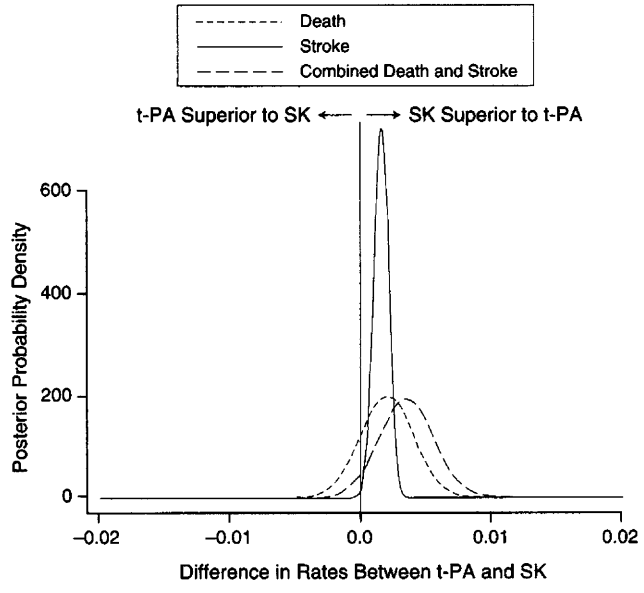

Figure 2.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial, with full prior use of data from the GISSI-2 and ISIS-3 trials. The area under the curve between any two points on the x-axis is the posterior probability that the difference in rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

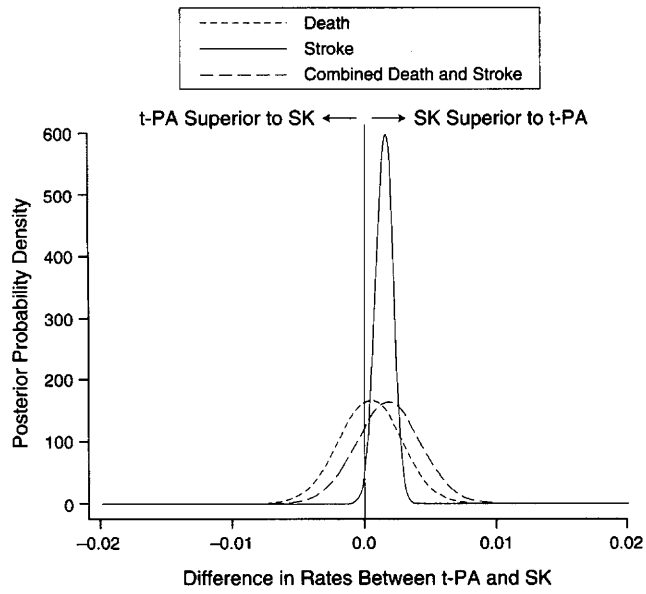

Figure 3.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial, with 50% prior use of data from the GISSI-2 and ISIS-3 trials. The area under the curve between any two points on the x-axis is the posterior probability that the difference in rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

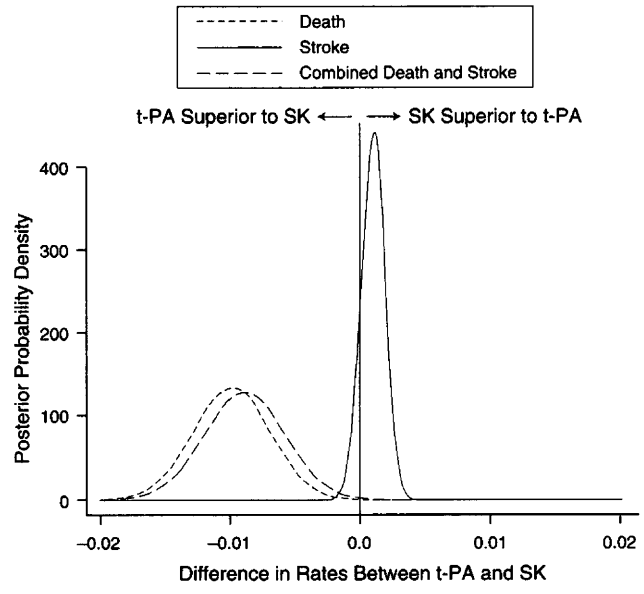

Figure 4.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial only. The area under the curve between any two points on the x-axis is the posterior probability that the difference in rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

then summarizes the range of posttrial beliefs. If this latter set of distributions includes only a sufficiently narrow range of possible effects, conclusions could be drawn with which most clinicians should agree regardless of their initial opinions. Otherwise, the debate continues and further research is indicated.

These methods and their interpretation are illustrated below. Other studies[1,3,13,14] provide fuller descriptions of the use of Bayesian analysis in the con-

Table 2.—Probability of t-PA Superiority as a Function of Prior Belief in GISSI-2 and ISIS-3 Data After Consideration of the GUSTO Data*

Prior Belief in GISSI-2 and ISIS-3, %	Probability of t-PA Mortality Higher Than SK Mortality	Probability of t-PA Net Clinical Benefit Greater Than SK Benefit	Probability of t-PA Net Clinical Benefit Greater Than SK Benefit by at Least 1%
100	.17	.05	<.001
50	.44	.24	<.001
10	.98	.94	.03
0	.999	.998	.36

*See footnote to Table 1 for expansions of abbreviations. Net clinical benefit is the combined death and stroke rate.

text of clinical trials. In this study, posterior distributions for the difference in survival rates between groups of patients receiving two different thrombolytic regimens following acute myocardial infarction are derived and graphically displayed. (Mathematical equations used to derive the Figures are available from the authors on request.)

The GUSTO trial randomized 41 021 patients to four different thrombolytic strategies involving SK, t-PA, or a combination of the two for the treatment of acute myocardial infarction. Compared with SK, the strategy of "front-loaded" or "accelerated" t-PA showed a statistically significant lowered mortality (6.3% vs 7.3%, respectively; P=.001) and combined end point of 30-day mortality or disabling stroke (6.9% vs 7.8%, respectively; P<.006) (Table 1). The interpretation of a P value of .001 is that if the two agents had exactly equivalent mortality rates, then data as extreme as or more extreme than the observed mortality rates would occur once in every 1000 hypothetical repeated trials.

This well-executed clinical trial possesses many of the desirable attributes of a well-done study. The sample size was very large and was designed to have at least 80% power to detect a 15% reduction in mortality or an absolute decrease of 1% between experimental groups. This value has been (somewhat arbitrarily) defined by the GUSTO investigators as the clinically important difference between the two agents. Economic analyses that incorporate patient utilities and health care expenditures may be required to further investigate what difference is clinically meaningful. In this article, we will accept a 1% decrease as the clinically meaningful difference. Potential confounding and bias were minimized by the randomization process. Most clinicians would accept the frequentist analysis of this study as being conclusive (or almost conclusive) proof of the superiority of t-PA, that is, the mortality rate for t-PA was less than that for SK. But is this an adequate summary of the available evidence?

Two previous randomized clinical trials have directly compared SK with t-PA in 48 000 patients. The GISSI-2[6] trial (Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico) compared t-PA (alteplase) and SK both with and without subcutaneous heparin beginning 12 hours after the start of therapy. The 35-day total mortality and nonfatal stroke data are summarized in Table 1. The ISIS-3[7] trial (Third International Study of Infarct Survival) compared t-PA (duteplase) and SK both with and without subcutaneous heparin in a similar factorial design but began heparin 4 hours after the start of therapy. The 35-day mortality and morbidity data are also shown in Table 1.

Although all the trials were randomized with uniform entry criteria and drug dosages, reservations have been expressed about the relevance of any comparisons between these studies. The major sources of controversy are as follows:

- The t-PA used in ISIS-3 was of a slightly different form (although the clinical difference is not believed to be significant).
- Adjunctive therapy accompanying t-PA in GUSTO included more aggressive use of intravenous heparin.
- In GUSTO t-PA was administered in an accelerated fashion.

While there is an abundance of prior information comparing these two agents, there is little consensus as to which agent is superior. Clinicians may vary in their weighting of the importance of the similarities and differences between the trials. This only enhances the utility of a Bayesian analysis, because their uncertainty can be explicitly considered by employing a range of prior beliefs.[13,14]

Figure 1 shows the probability density for the difference in mortality between t-PA and SK as determined from the data of GISSI-2 and ISIS-3. (The area under the probability density curve between two given points on the x-axis represents the probability that a value will fall between the two points.) The difference in mortality rates between t-PA and SK appears along the x-axis (0.01=1% and so forth), and the height of the probability density for this difference is given by the y-axis. The mean of these curves is close to zero (0.0013), suggesting no difference between the two agents. Fully accepting the results of these two trials would suggest almost no possibility of t-PA's being clinically superior to SK (a decrease in the mortality rate with t-PA ≥1% is represented by the area to the left of −0.01, and this area is essentially zero in the case using 100% of the prior data). This leads to a very skeptical prior distribution as to the superiority of t-PA. On the other hand, a clinician who believes that the difference in trial protocols cannot be ignored might elect to only partially consider the earlier results. For example, one could arbitrarily treat the value of each observation in the previous trials as worth only 50% or even 10% of each observation in the GUSTO data. Prior distributions based on these weights also appear in Figure 1. A more extreme position would be that the trials are too dissimilar to be combined and that consequently all previous research should be ignored, thereby assuming that nothing is known about the potential difference in mortality between the two agents (in statistical parlance, this implies a noninformative or uniform prior distribution). Other prior distributions are also possible and are not necessarily derived by a weighting of previous data. Most of these would fall in between the above-mentioned extremes. As the belief in the utility of the prior studies decreases, so increases the possibility that t-PA is a clinically superior agent (widening of the curves and increasing area to the left of −0.01).

RESULTS

The data from Table 1 may be used to derive posterior distributions for stroke, death, and net clinical benefit (death and nonfatal stroke) using Bayes' theorem (the solved equation is available from the authors on request). Figure 2 considers the skeptical prior belief that assigns equal weight to each observation from GISSI-2, ISIS-3, and GUSTO and shows that the mean difference in mortality between t-PA and SK is 0.20% (0.002 in favor of SK), and the final (posterior) probability of t-PA's being superior to SK is only about 17% (area under the curve to the left of 0). Figure 2 also demonstrates that there are 0.15% more nonfatal strokes with t-PA and that the probability that the rate of nonfatal stroke is greater with t-PA exceeds 99.5% (the area to the left of the curve <.005). A similar interpretation of the combined curve suggests that the probability that t-PA is superior to SK is 5.1% with an almost zero probability of exceeding the clinically signficant difference of 1% (area to the left, on the

combined curve of 0 and −0.01, respectively).

Figure 3, which considers observations from the previous randomized clinical trials to have 50% the value of each observation in GUSTO (a more intermediate prior belief), shows that the probability that t-PA is superior to SK for mortality alone is about 44% (again refer to the area to the left of 0 for the appropriate curve). Further, accepting that a difference of 1% mortality is the minimum clinically significant value, the probability that t-PA is clinically superior remains negligible. The probability of increased stroke with t-PA remains high at almost 98%.

Finally, Figure 4 shows the scenario where all prior data from GISSI-2 and ISIS-3 are considered irrelevant and are ignored. In this case, t-PA is virtually certain to have a lower death rate than SK (99.95%), but the probability that t-PA exceeds the defined clinical superiority is only 48%. The probability of a net clinical benefit exceeding 1% is only 36%, and the probability of increased stroke with t-PA is 86%. The salient elements of Figures 2 through 4 are displayed in Table 2.

COMMENT

The current study demonstrates several advantages of a Bayesian analysis. The most apparent is that the analysis permits the direct answer as to the probability that t-PA is superior to SK. It also permits the calculation of the probability of clinical superiority. The answers, however, can vary since readers must each draw their own conclusions by selecting the posterior distribution that belongs to the prior distribution most closely matching their own initial personal beliefs. The GUSTO investigators suggested a minimum clinical superiority based on economic factors of one life saved per 100 patients treated, but Table 2 could be expanded to include any personalized prior distribution and clinical superiority cut point.

The Bayesian analysis presented herein suggests that restraint in accepting t-PA into routine clinical practice would be appropriate. The same conclusion was reached by Dr Diamond and colleagues,[15] who used a Bayesian point null hypothesis test. When one accepts only partial recognition (50%) of previous randomized clinical trials, the probability that t-PA is superior to SK for mortality or net clinical benefit is only 44% and 24%, respectively. The probability that either mortality or net clinical benefit would exceed clinical importance with the 50% assumption is much less than 1%. Even if one totally ignores all prior studies, the chance that t-PA would exceed the clinical superiority cut point for mortality and net clinical benefit is only 48% and 36%, respectively.

Neither *P* values nor the Bayesian analysis presented herein measures potential bias or confounding. The GUSTO trial was unblinded, which may lead to some degree of confounding. For example, while not reported in the original article, it appears that 9.5% of the t-PA group underwent coronary artery bypass surgery compared with 8.5% in the SK group. This difference may have contributed to the observed mortality differences. A Bayesian approach to adjustments for a wide variety of biases is described by Eddy et al.[12]

In assessing the public health impact of choosing a thrombolytic agent, the following seems clear. *P* values or confidence intervals from conventional statistical analysis are poor tools for formulating public health policy, even when there is a considerable amount of data from the best-designed randomized clinical trials. This is due to the shortcomings of standard significance tests in addressing clinically relevant questions and to the problems in their interpretation, especially across different sample sizes. Furthermore, classical analysis of clinical trials does not easily permit the synthesis of trial results with the range of clinicians' prior beliefs. This makes it difficult to evaluate the coherency of the conclusions and what clinical impact the conclusions should have. Bayesian analyses along the lines presented herein may help to overcome these problems, thereby raising the level of debate following publication of a clinical trial.

References

1. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Intern Med.* 1983;98:385-394.
2. Browner WS, Newman TB. Are all significant *P*-values created equal? the analogy between diagnostic tests and clinical research. *JAMA.* 1987; 257:2459-2463.
3. Berger J, Berry D. Statistical analysis and the illusion of objectivity. *Am Scientist.* 1988;76:159-165.
4. Frieman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, type II error and sample size in the randomized control trial: survey of 71 'negative' trials. *N Engl J Med.* 1978;299:690-694.
5. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med.* 1993;329:673-682.
6. The International Study Group. In-hospital mortality and clinical course of 20,891 patients with suspected acute myocardial infarction randomised between alteplase and streptokinase with or without heparin. *Lancet.* 1990;336:71-75.
7. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group. ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 4,299 cases of suspected acute myocardial infarction. *Lancet.* 1993;339:753-770.
8. Rapaport E. GUSTO: assessment of the preliminary results. *J Myocard Ischemia.* 1993;5:15-24.
9. Sleight P. Thrombolysis after GUSTO: a European perspective. *J Myocard Ischemia.* 1993;5:25-30.
10. Ridker PM, O'Donnell C, Marder VJ, Hennekens CH. Large-scale trials of thrombolytic therapy for acute myocardial infarction: GISSI-2, ISIS-3, and GUSTO-1. *Ann Intern Med.* 1993;119:530-532.
11. Ridker PM, O'Donnell C, Marder VJ, Hennekens CH. A response to 'holding GUSTO up to the light.' *Ann Intern Med.* 1994;120:882-884.
12. Eddy DM, Hasselblad V, Shachter R. *Meta-analysis by the Confidence Profile Method.* New York, NY: Academic Press; 1992.
13. Hughes M. Reporting Bayesian analyses of clinical trials. *Stat Med.* 1993;12:1651-1663.
14. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *J R Stat Soc A.* 1994;157:357-416.
15. Diamond GA, Denton TA, Forrester JS, Shah PK. Is tissue plasminogen really superior to streptokinase? *Circulation.* 1993;88:I-452. Abstract.