11. Bayesian Inference

Readings: Rosner: Ch. 3.7-3.8, 7.8
Chihara and Hesterberg: Ch. 10

Homework: Homework 5 due by noon on October 8
Homework 6 due by noon on October 15

Overview

A) Frequentist vs Bayesian
B) The Bayes Factor: Examples
C) GUSTO Trial

A) Frequentist vs Bayesian Statistics

Bayes' Theorem provides the fundamental groundwork for Bayesian inference, but also can draw connections to frequentist inference. Let H represent a hypothesis to be tested and D be the data which may give evidence for or against H, then Bayes' theorem is:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Each of these terms has a different role to play:

- $P(H)$ is the **prior** (probability) that H is true *before* the data is considered
- $P(D|H)$ is the **likelihood** and represents the evidence for H provided by the observed data D
- $P(D)$ is the **total probability** of the data which takes into account all possible hypotheses
- $P(H|D)$ is the **posterior** (probability) that H is true after the data has been considered

If we know the likelihood and prior value for all hypotheses, then we can calculate the posterior exactly. For example, with a known (or estimated) prevalence we can calculate the predictive value (aka posterior probability) for having a disease given a positive test screen (Lecture 9, Section D). In fact, Bayesians and frequentists both use Bayes' Theorem in this case.

However, when the prior is not known, Bayesians and frequentists take different routes:

- Frequentists abandon Bayes' Theorem and conduct inference based on just the likelihood function: $L(H; D) = P(D|H)$
- Bayesians make up some prior distribution, hopefully based on some evidence or expert opinion for something reasonable.

Frequentists take this approach because they believe probabilities represent long-term frequencies of repeatable random experiments. For example, if the probability is 1/6 to roll a 1 on a six-sided die, the relative frequency of rolling 1's should approach 1/6 as the number of die rolls approaches infinity. Frequentists are putting a probability distribution on data from random, repeatable experimental given a hypothesis.

Bayesians, however, put probability distributions on both the hypotheses and data.

Bayesian Inference	**Frequentist Inference**
Uses probabilities for both hypotheses and data	Never uses or gives the probability of a hypothesis (no prior or posterior)
Depends on the prior and likelihood of the observed data	Depends on the likelihood P(D|H) for both observed and unobserved data
Requires one to specify a (subjective) prior	Does not require a prior

A.1) Frequentist and Bayesian Interpretations/Summaries

Intervals of uncertainty:

Consider a situation where we have calculated a 95% interval of (25,35) around our mean estimate of 30 from a sample. Depending on our approach, this interval has different interpretations:

Confidence Intervals (Frequentist): In the long run, 95% of the confidence intervals will include the population mean. Specifically, we are 95% confident that the population mean is between 25 and 35. (Note: we don’t use 95% probability, but use 95% confident!)

Credible Intervals (Bayesian): The range of the middle 95% of a posterior distribution. Specifically, given our data, there is a 95% probability that the true population mean is contained between 25 and 35. (Note: here we do use 95% probability.)

For frequentist confidence intervals the parameter (i.e., mean) is treated as a fixed value. For Bayesian credible intervals the parameter is a random variable.

P-values vs. Bayes Factors/Posterior Probabilities:

Frequentists have evolved to consider a Fisher-Neyman-Pearson hybrid approach involving p-values (Lecture 7, Section B2b). Bayesians don't have p-values but can use some other summary measures.

P-values (Frequentist): The probability of obtaining a test statistic as extreme or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

Bayes Factors (Bayesian): The ratio of the likelihood probability of two competing hypotheses, such as H_0 and H_1. It quantifies the evidence of our data for H_0 vs. H_1.

Posterior Probability (Bayesian): The product of our prior odds and the Bayes Factor results in the posterior odds, which we can convert to the posterior probability. This is the probability that the hypothesis is true based on the data we have observed.

In the discrete case, the Bayes Factor is the ratio of likelihoods. For the continuous case we have the ratio of marginal likelihoods.

B) Bayes Factors

For the Bayes Factor $\frac{P(D|H_0)}{P(D|H_1)}$, Harold Jeffreys proposed classifying strength as

BF	**Strength of Evidence in Favor of H_0**
<1	Negative (supports H_1)
1 to 3	Barely worth mentioning
3 to 20	Substantial
20 to 150	Strong
>150	Very strong

Example: Suppose we want to determine if a coin is "fair" and we design an experiment where we flip the coin 100 times and record the outcome. Before conducting the test, we propose H_1 as P(heads)=0.65. Suppose we observe 60 heads and 40 tails. What is our BF?

$$P(X = 60|H_0) = \binom{100}{60}\left(\frac{1}{2}\right)^{60}\left(1 - \frac{1}{2}\right)^{40} = 0.01084387.$$

$$P(X = 60|H_1) = \binom{100}{60}(0.65)^{60}(1 - 0.65)^{40} = 0.0473922.$$

$BF = \frac{0.01084387}{0.0473922} = 0.2288$ (or we can consider $\frac{1}{BF} = 4.37$ with respect to favor of H_1)

Using the exact binomial test (a frequentist test), we would find p=0.004.

What if we weren't sure about specifying a single H_1, but wanted to equally consider all possible probabilities of heads? Then the probability of heads is uniform on [0,1] and our estimated likelihood for $P(X=60|H_1)$ is:

$$P(X=60|H_{1b}) = \int_0^1 \binom{100}{60} (p)^{60} (1-p)^{40}\, dp = \frac{1}{101}$$

This results in a Bayes Factor of $BF = \frac{0.01084387}{1/101} = 1.095$, which is barely worth mentioning!

What if we instead believed, a priori, that the coin was biased towards tails and wanted to equally restrict to hypotheses that favored this? Then the probability of heads is uniform on [0,0.5]:

$$P(X=60|H_{1c}) = \int_0^{0.5} \binom{100}{60} (p)^{60} (1-p)^{40}\, dp = 0.000227941$$

$BF = \frac{0.01084387}{0.000227941} = 47.6$, strong evidence in favor of H_0 compared to H_{1c}.

We can see that depending on our "prior", the resulting BF's could be drastically different! This doesn't mean we should use Bayesian inference, just that we should thoughtfully consider our priors.

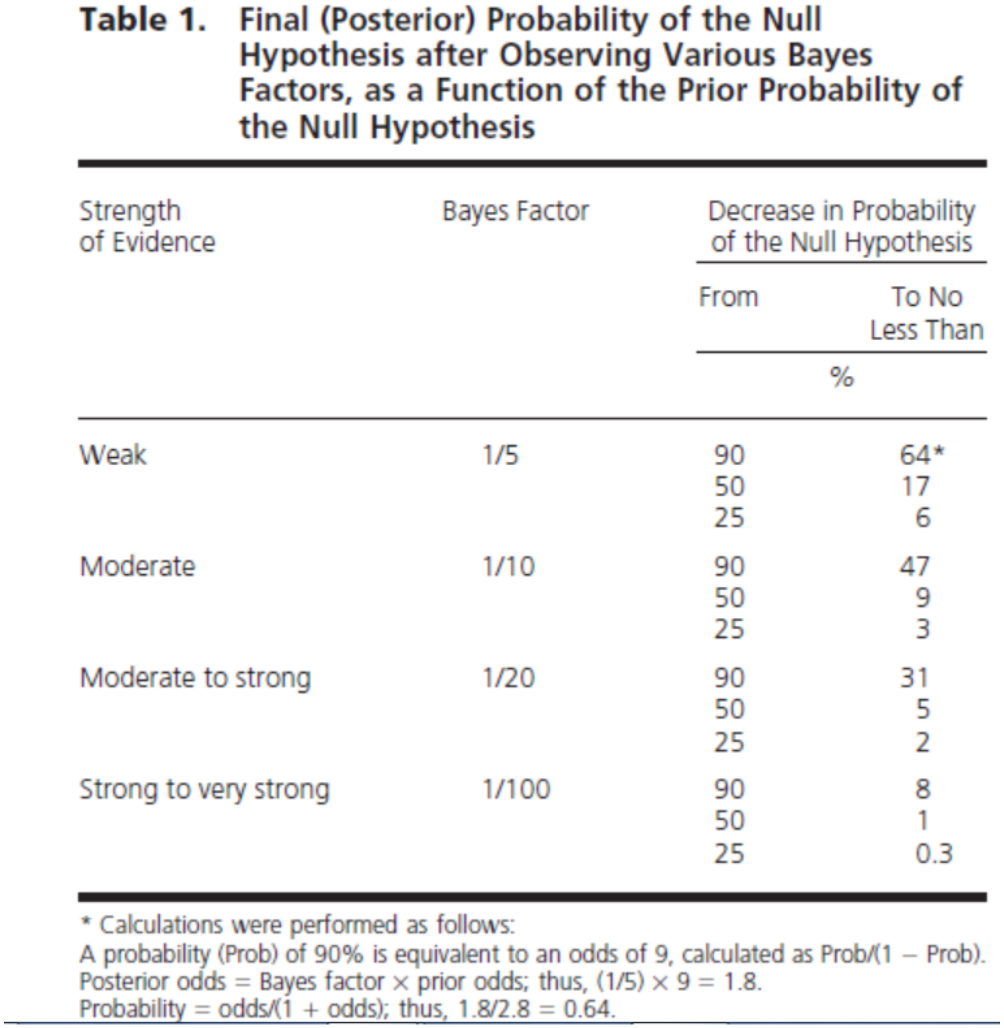

Table 1. Final (Posterior) Probability of the Null Hypothesis after Observing Various Bayes Factors, as a Function of the Prior Probability of the Null Hypothesis

Strength of Evidence	Bayes Factor	Decrease in Probability of the Null Hypothesis	
		From	To No Less Than
		%	
Weak	1/5	90	64*
		50	17
		25	6
Moderate	1/10	90	47
		50	9
		25	3
Moderate to strong	1/20	90	31
		50	5
		25	2
Strong to very strong	1/100	90	8
		50	1
		25	0.3

* Calculations were performed as follows:
A probability (Prob) of 90% is equivalent to an odds of 9, calculated as Prob/(1 − Prob).
Posterior odds = Bayes factor × prior odds; thus, (1/5) × 9 = 1.8.
Probability = odds/(1 + odds); thus, 1.8/2.8 = 0.64.

We can see a similar message from the Goodman (1999) paper on the Bayes Factor. Where stronger prior belief in the null hypothesis (the “From” column) leads to different posterior probabilities that H_0 is true (the “To No Less Than” column).

C) GUSTO Trial

What was the study design of GUSTO? Groups compared?

What was the primary outcome of GUSTO? Other measures?

Did they identify a needed sample size/power?

What conclusions did the NEJM article include?

Brophy and Joseph's thoughts on reanalyzing GUSTO from a Bayesian perspective

Other trial results (GISSI-2 and ISIS-3) did not come to the same conclusion of GUSTO:

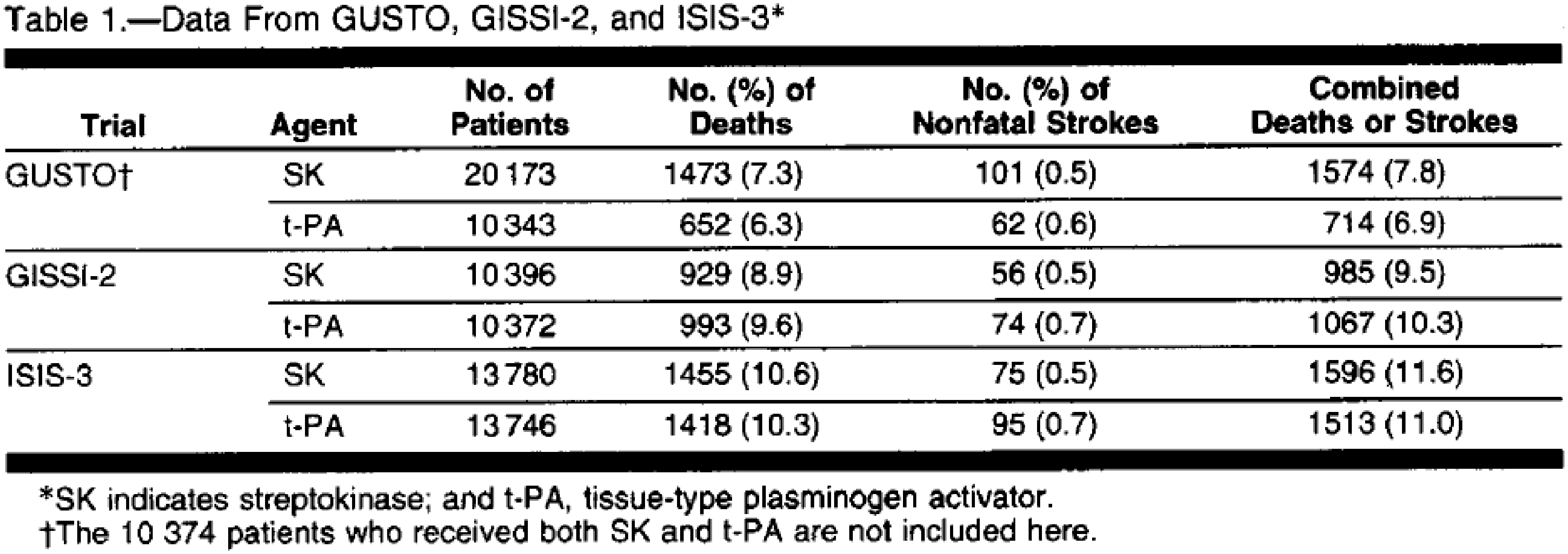

Table 1.—Data From GUSTO, GISSI-2, and ISIS-3*

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.

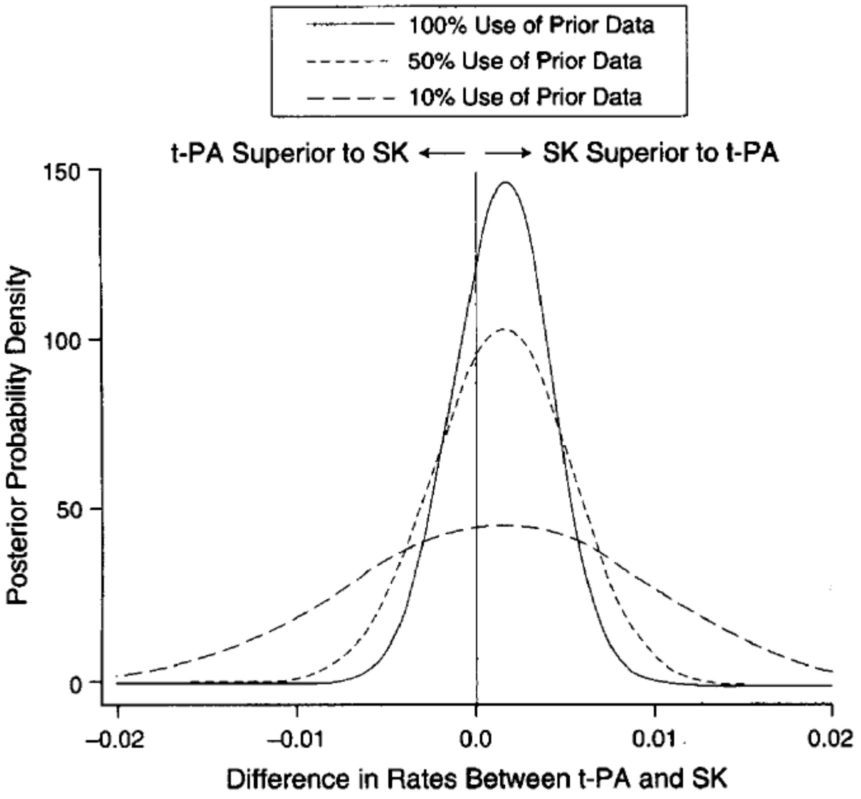

Figure 1.—Plot of the prior distributions for the difference in mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK) using weights of 100%, 50%, and 10% of the GISSI-2 and ISIS-3 data, representing a range in prior beliefs in the relevance of these trials to the GUSTO 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.

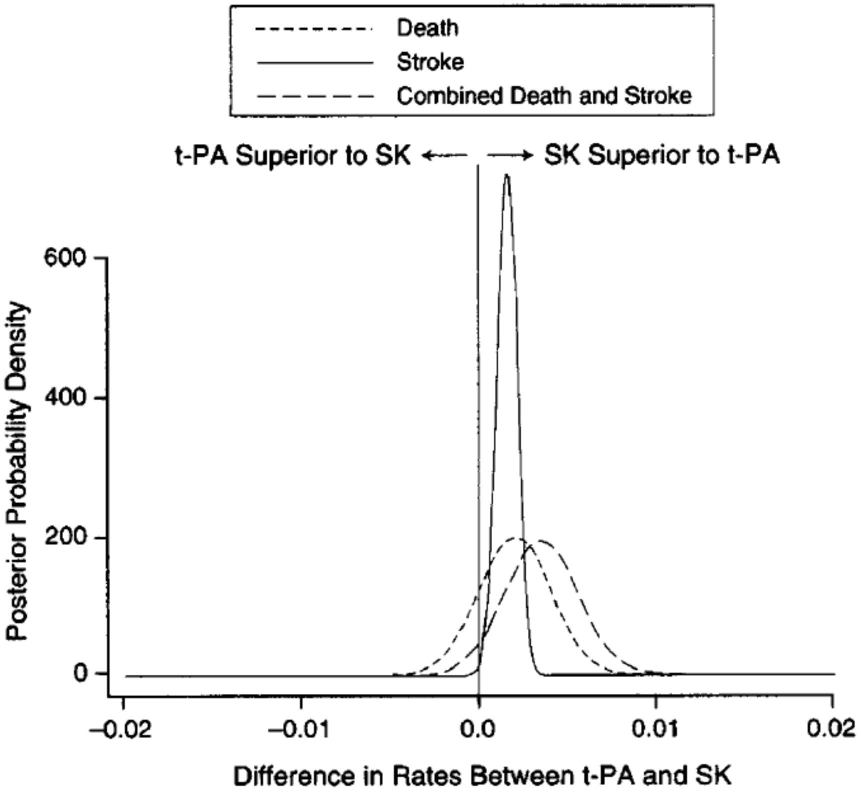

Figure 2.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial, with full prior use of data from the GISSI-2 and ISIS-3 trials. 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.

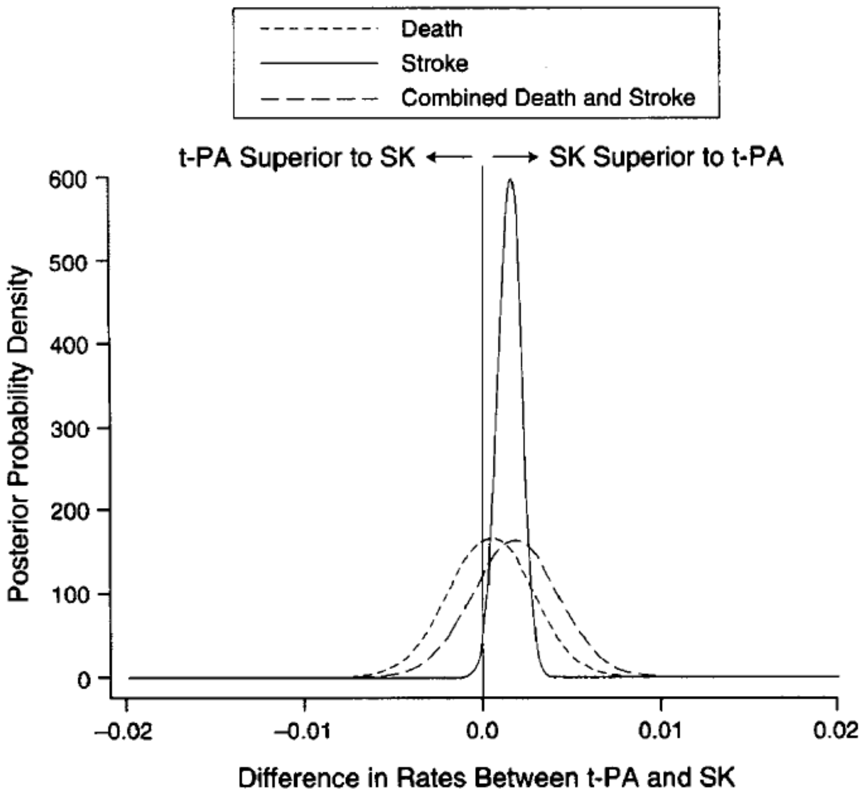

Figure 3.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial, with 50% prior use of data from the GISSI-2 and ISIS-3 trials. 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.

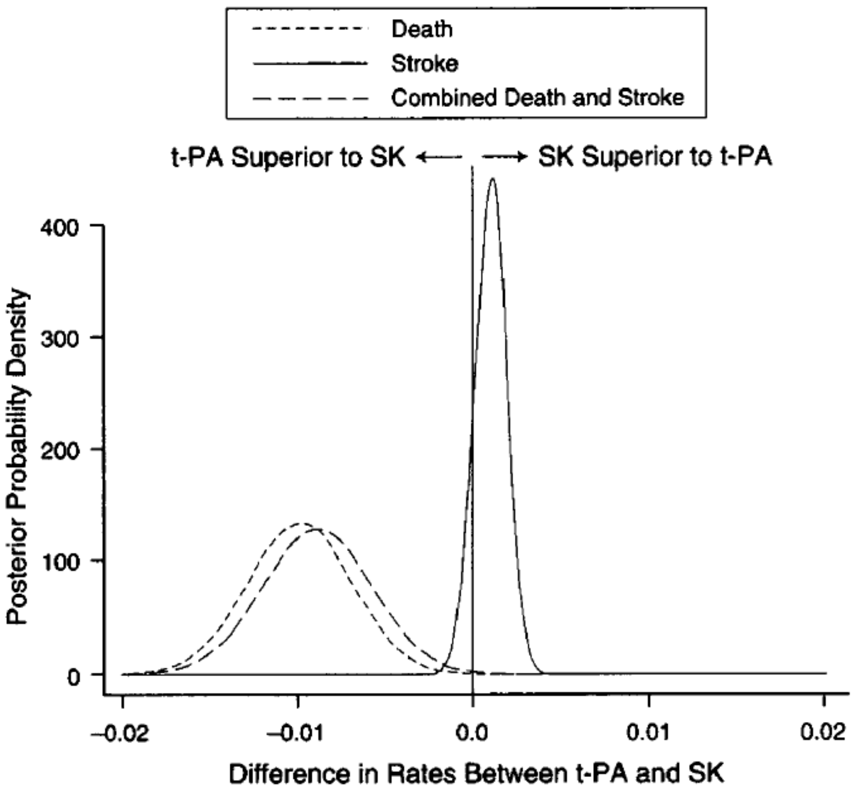

Figure 4.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial only. The area under the curve between any two points on the 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.

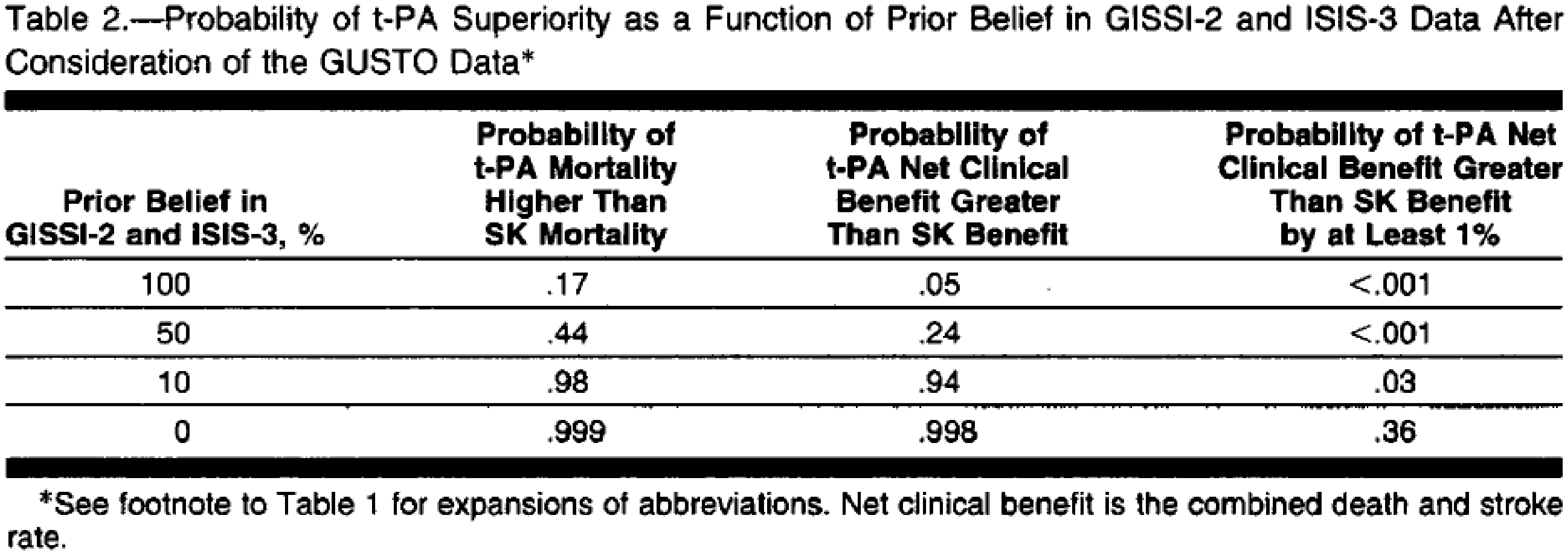

Table 2.—Probability of t-PA Superiority as a Function of Prior Belief in GISSI-2 and ISIS-3 Data After Consideration of the GUSTO Data*

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.

(Note: It appears their column label for mortality is reversed…it should be probability that t-PA is *lower* than SK)

Brophy and Joseph note that even if we borrow no information from GISSI-2 and ISIS-3 (prior belief of 0%), we still only have a probability of net clinical benefit >1% of 36%, which is not very conclusive (however, the posterior probability is strongly in favor of t-PA reducing mortality).

They point out that these results don’t definitively answer the question of t-PA vs. SK, but say that “restraint in accepting t-PA into routine clinical practice would be appropriate.”

Baisy the Bayesian Pup