

A Good P -value is Hard to Find:
Why I'm a Bayesian When Time
Allows

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine

`f.harrell@vanderbilt.edu`

WESTERN TENNESSEE ASA CHAPTER 15 JUNE 2007

Outline

- Disadvantages and Controversies in Frequentist Inference
- What's Wrong with Hypothesis Testing?
- What's Wrong with Confidence Intervals?
- Bayesian Approach
- Controversy: Choice of Prior
- Sequential Testing
- Subgroup Analysis
- Multiple Endpoints
- Software
- Examples from Clinical Trials
- Suggested Design Criteria
- Implications for Design/Evaluation

Frequentist Approach

- Demonstrate S by assuming \bar{S} and showing it's unlikely
- Compute $\Pr[T \text{ as or more impressive as one observed} | H_0]$
- Is a measure of how embarrassing the data are to the null hypothesis^a
- Probabilities “refer to the frequency with which different values of statistics (arising from sets of data other than those which have actually happened) could occur for some fixed but unknown values of the parameters” (Box and Tiao¹¹)
- Simple to think of unknown parameter as a constant
- P -values relatively easy to compute
- Accepted by most of the world

^aNicholas Maxwell, *Data Matters: Conceptual Statistics for a Random World*, 2004.

- Prior beliefs not needed at computation time

Disadvantages and Controversies

- “Have to decide which ‘reference set’ of groups of data which have not actually occurred we are going to contemplate”¹¹; what is “impressive”?
- Conditions on what is unknowable (parameters) and does not condition on what is *already* known (the data)
- Does not address clinical significance
- If real effect is mean decrease in BP by 0.2 mmHg, large enough n will yield $P < 0.05$
- Controversy surrounding 1-tailed vs. 2-tailed tests
- No method for trading off type I and type II error
- No uniquely accepted P -value for 2×2 table! What is “extreme”: of all possible tables or all tables with same total no. of deaths?

No consensus on the optimum procedure for obtaining a P -value (e.g., Pearson χ^2 vs. Fisher's so-called exact test, continuity correction, likelihood ratio test, new unconditional tests).

- For ECMO trial, 13 P -values have been computed for the same 2×2 table, ranging from 0.001 to 1.0
- P -values very often misinterpreted
- Since P -values are probabilities of obtaining a result as or more extreme than the study's result under repeated experimentation, frequentists interpret results by inferring "what would have occurred following results that were not observed at analyses that were never performed"¹⁹.
- Berger and Berry: $n = 17$ matched pairs, $P = 0.049$, the **maximum** $\Pr[\text{treatment effective}] = 0.79$
- Very hard to directly answer interesting questions such as $\Pr[\text{similarity}]$
- Much controversy about need for/how to ad-

just for multiple comparisons

- Do you want $\Pr[\text{Reject} \mid \text{this } H_0 \text{ true}] = 0.05$, or $\Pr[\text{Reject} \mid \text{this and other } H_0\text{s true}] = 0.05$?
- If the latter, C.L.s must use e.g. $1 - \frac{\alpha}{k}$ conf. level \rightarrow precision of a parameter estimate depends on what other parameters were estimated
- Cook and Farewell¹⁷: If results are intended to be interpreted marginally, there may be no need for controlling experimentwise error rate
- Many conflicting alternative adjustment methods
- Bonferroni adjustment is consistent with a Bayesian prior distribution which specifies that the probability that all null hypotheses is true is a constant (say 0.5) no matter how many hypotheses are tested⁵²
- Much controversy about need for adjusting for sequential testing. Frequentist approach is complicated.

Example: 5 looks at data as trial proceeds

Looks had no effect, trial proceeded to end
Usual $P = 0.04$, need to adjust upwards for
having looked

Two studies with identical experiments and
data but with investigators with different in-
tentions → one might claim “significance”, the
other not (Berry⁶)

Example: one investigator may treat an in-
terim analysis as a final analysis, another may
intend to wait.

- It gets worse — need to adjust “final” point estimates for having done interim analyses
- Freedman et al.²⁴ give example where such adjustment yields 0.95 CI that includes 0.0 even for data indicating that study should be stopped at the first interim analysis
- As frequentist methods use intentions, they are not fully objective⁵
- P -values can only be used to provide evidence

against a hypothesis, not to give evidence in favor of a hypothesis. Schervish⁴⁴ gives examples where P -values are incoherent: if one uses a P -value to gauge the evidence in favor of an interval hypothesis for a certain dataset, the P -value based on the same dataset but for a *more restrictive* sub-hypothesis (i.e., one specifying a subset of the interval) actually gives more support (larger P).

- Equal P -values do not provide equal evidence about a hypothesis⁴²
- Rejecting H_0 just suggests that something is wrong with the model, but we may not know what¹⁶ (e.g., non-normality, non-independence, unequal variances)

- Why are P -values still used?

Feinstein²² believes their status "... is a lamentable demonstration of the credulity with which modern scientists will abandon biologic wisdom in favor of any quantitative ideology that offers the specious allure of a mathematical replacement for sensible thought."

What's Wrong with Hypothesis Testing?

- Hypotheses are often “straw men” that are imagined by the investigator just to fit into the traditional statistical framework
- Most phenomena of interest are not all-or-nothing but represent a continuum

The Applied Statistician's Creed

- Nester³⁷:
 - (a) TREATMENTS — all treatments differ;
 - (b) FACTORS — all factors interact;
 - (c) CORRELATIONS — all variables are correlated;
 - (d) POPULATIONS — no two populations are identical in any respect;
 - (e) NORMALITY — no data are normally distributed;
 - (f) VARIANCES — variances are never equal;
 - (g) MODELS — all models are wrong;
 - (h) EQUALITY — no two numbers are the same;
 - (i) SIZE — many numbers are very small.

- → no two treatments actually yield identical patient outcomes
- → Most hypotheses are irrelevant

Has Hypothesis Testing Hurt Science?

- Many studies are powered to be able to detect a huge treatment effect
- → sample size too small → confidence interval too wide to be able to reliably estimate treatment effects
- “Positive” study can have C.L. of [.1, .99] for effect ratio
- “Negative” study can have C.L. of [.1, 10]
- Physicians, patients, payers need to know the magnitude of a therapeutic effect more than whether or not it is zero
- “It is incomparably more useful to have a plausible range for the value of a parameter than to know, with whatever degree of certitude, what single value is untenable.” — Oakes³⁸
- Hypothesis testing usually entails fixing n ; many studies stop with $P = 0.06$ when adding 20

more patients could have resulted in a conclusive study

- Many “positive” studies are due to large n and not to clinically meaningful treatment effects
- Hypothesis testing usually implies inflexibility⁴⁵

What's Wrong with Confidence Intervals?

- Misinterpreted twice as often as P -values
- Consumers act as if “degree of confidence” is uniform within the interval —
- C.I. for OR of [.35, 1.01] misinterpreted as indicating that a 1% increase in mortality is as likely as a 10% decrease

Methods

- Attempt to answer question by computing probability of the truth of a statement
- Let S denote a statement about the drug effect, e.g., patients on drug live longer than patients on placebo
- Want something like $\Pr[S \mid \text{data}]$
- If θ is a parameter of interest (e.g., log odds ratio or difference in mean blood pressure), need a probability distribution of $\theta \mid \text{data}$
- $\Pr[\theta \mid \text{data}] \propto \Pr[\text{data} \mid \theta] \Pr[\theta]$
- $\Pr[\theta]$ is the *prior* distribution for θ
- Assuming θ is an unknown random *variable*
- $p(\theta \mid y) \propto l(\theta \mid y)p(\theta)$
- $l(\theta \mid y) =$ likelihood function
- Function through which data y modifies the prior knowledge of θ ¹¹

- Has the information about θ that comes from the data

Advantages

- “intended for measuring support for hypotheses when the data are fixed (the true state of affairs after the data are observed)” (Schervish⁴⁴)
- Results in a probability most clinicians think they’re getting
- Can compute (posterior) probability of interesting events, e.g.
Pr[drug is beneficial]
Pr[drug A clinically similar to drug B]
Pr[drug A is $> 5\%$ better than drug B]¹⁵
- Provides formal mechanism for using prior information/bias — $\Pr[\theta]$
- Places emphasis on estimation and graphical presentation rather than hypothesis testing
- Avoids 1-tailed/2-tailed issue
- If $\Pr[\text{drug B is better than drug A}] = 0.92$, this is true whether drug C was compared to drug D or not

- Avoids many of complexities of sequential monitoring —
P-value adjustment is needed for frequentist methods because repeatedly computed test statistics no longer have a χ^2 or normal distribution;
A posterior probability is still a probability →
Can monitor continuously
- Allows accumulating information (from this as well as other trials) to be used as trial proceeds
- No need for sufficient statistics

Controversy: Choice of Prior

- Biased prior – expert opinion
difficult, can be manipulated, medical experts often wrong, whose opinion do you use?²³
- Skeptical prior (often useful in sequential monitoring)
- Unbiased (flat, non-informative) prior
- Truncated prior — allows one to pre-specify e.g. there is no chance the odds ratio could be outside $[\frac{1}{10}, 10]$
- For monitoring, Spiegelhalter et al.⁴⁸ suggest using “community of priors”:
 - Skeptical prior with mean 0 against which judge early stopping for efficacy
 - Enthusiastic prior with mean δ_A (hypothesized effect) against which judge early stopping for no difference
- Heitjan³⁰ uses a “theoretical skeptical expert”; Stylized or “automatic” priors^{23, 34}

- Data quickly overwhelm all but the most skeptical priors, especially in clinical applications
- In scientific inference, let data speak for themselves (non-informative prior)
- \rightarrow *A priori* relative ignorance, draw inference appropriate for an unprejudiced observer (Box and Tiao¹¹)
- Scientific studies usually not undertaken if precise estimates already known. Also, problems with informed consent.

Invalid Bayesian Analyses

- Choosing an improper model for the data (can be remedied by adding e.g. non-normality parameter with its own prior¹¹)
- Sampling to a foregone conclusion if a continuous prior is used but the investigators and the consumers were convinced that prob. of treatment effect is *exactly* zero $> 0^a$
- Suppression of the latest data by an unscrupulous investigator:
 Current results using 200 patients nearly conclusive in favor of drug
 Decide to accrue 50 more patients to draw firm conclusion
 Results of 50 less favorable to drug
 Based final analysis on 200 patients^b

^aThis is easily solved by using a prior with a lump of probability at zero.

^bNote the martingale property of posterior probs.:
 $E[\Pr(\theta_1 > \theta_2 | \text{data}, \text{data}')] = \Pr(\theta_1 > \theta_2 | \text{data})$.

Two-Sample Binomial Example

- Advantageous to specify prior for OR instead of for the two probabilities of response θ_1, θ_2 ⁴⁸
Consider this later
- For now consider priors for θ_1, θ_2 :
 - Flat
 - $\propto [\theta(1 - \theta)]^{-\frac{1}{2}}$
- Data: Treatment A $\frac{30}{200}$
Treatment B $\frac{18}{200}$
- OR = 0.56; $2P = 0.064$ (LR), 0.068 (Wald);
 $1P = 0.034$ (Wald)
0.95 C.L. [.304, 1.042] (Wald based on normality of log OR)

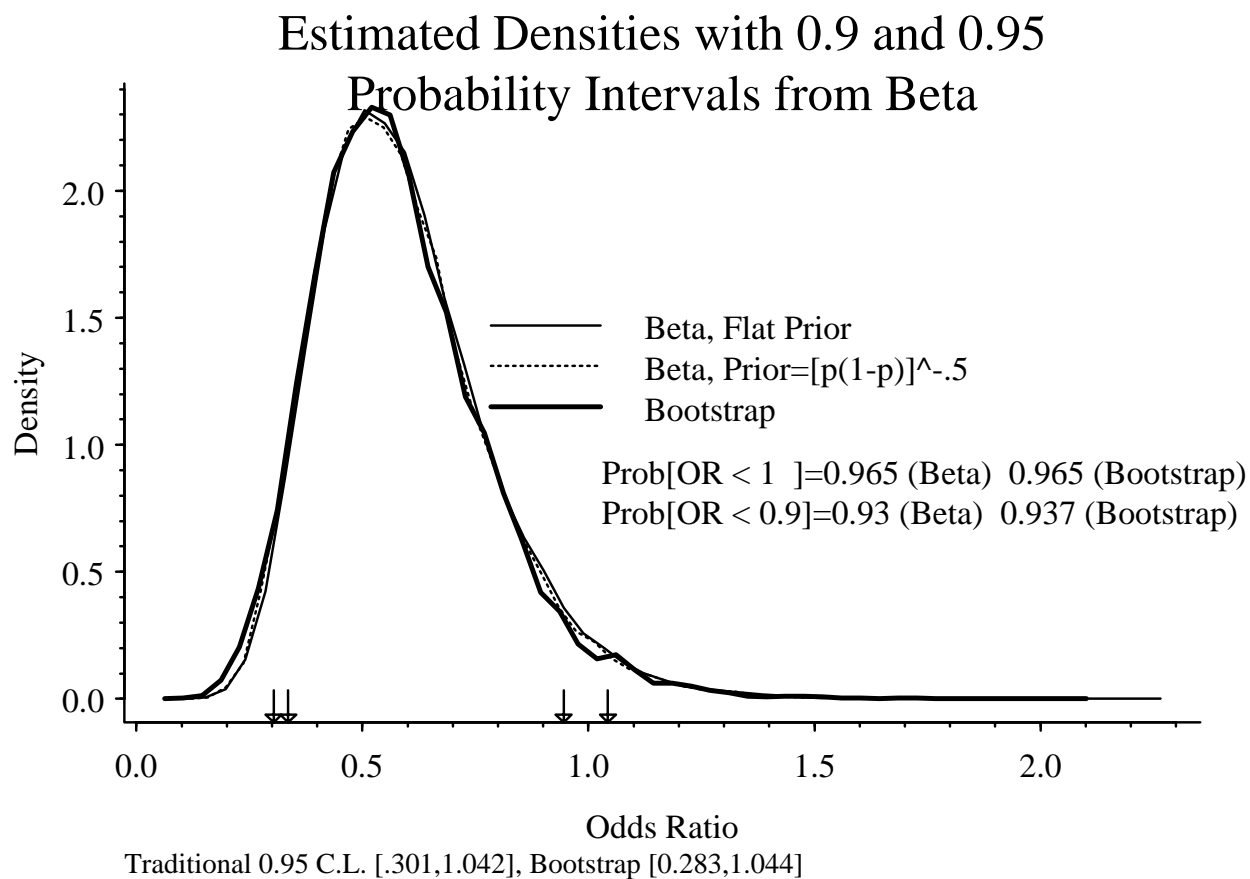


Figure 1: *Posterior density of OR from a kernel estimator. The posterior were derived using the bootstrap and using a Bayesian approach with 2 prior densities.*

Sequential Testing

- Frequentist approach to deciding when to quit watching a football game:
Of all games which ended in a tie or with your team losing, what proportion had your team leading by 10 points with 12m to go in 4th quarter?
Must consider sample space
- Bayesian approach: at each moment can estimate the probability that your team will ultimately win based on the time left and the point spread
- No problem with estimating this probability every second
- Distribution of unknown parameters updated at any time⁶
- No need for independent increments
- No need for equal information time

- No scheduling
- No adjustment of point estimates, C.L. for monitoring strategy
- Determining number of “looks” (k) that minimizes expected sample size — frequentist: plot of avg. sample size vs. k is U-shaped^a; Bayesian: the larger k the better⁶

^aBecause of α -adjustment

Subgroup Analysis

- Even with “significant” treatment effect in subgroup, point estimates of effects will be greatly exaggerated
- → Need to get away from hypothesis testing within subgroups
- Shrinkage methods needed
- Example: Represent differential treatment effects as random effects, shrinking them down to achieve optimal prediction⁴³
- If prior distribution for each parameter of interest is well-calibrated, posterior probabilities need no adjustment for the number of subgroups tested⁵²

Multiple Endpoints

- Example: 3 endpoints
- Target Z_1 : Population mean blood pressure $\downarrow \geq 5$ mmHg
- Target Z_2 : Population exercise time $\uparrow \geq 1$ min.
- Target Z_3 : Population mean angina score $\downarrow \geq 1$ point
- Posterior $\Pr[Z_1] = 0.97$
- Posterior $\Pr[Z_2] = 0.94$
- Posterior $\Pr[Z_3] = 0.6$
- $\Pr[Z_1 \cup Z_2 \cup Z_3] \geq 0.97$
- $\Pr[\bar{Z}_1 \cap \bar{Z}_2 \cap \bar{Z}_3] \leq 0.03$
- $\Pr[\#Z_i \geq 2] = 0.95$ for example
- To demonstrate that a drug improves at least one endpoint, study many endpoints!
- May want to show that at least $\frac{1}{2}$ of the endpoints are improved with high probability

- Alternative: Panel of experts rate importance of outcomes, e.g., $Z_1 = 1, Z_2 = 2, Z_3 = 3$
- Target could be ≥ 3 points
- Here $\Pr[Z_3 \cup (Z_1 \cap Z_2)] \geq 0.95$
- Simply count number of samples from posterior satisfying $Z_3 \cup (Z_1 \cap Z_2)$
- Another way to summarize results: Estimate $E[\#Z_i] = 0.97 + 0.94 + 0.6 = 2.51$ out of 3
- If all endpoints are binary, a kind of random effects model for the endpoints may be useful³⁵
- If prior distribution for each parameter of interest is well-calibrated, posterior probabilities need no adjustment for the number of responses tested⁵²
- See Berry⁷ for a Bayesian perspective on data-generated hypothesis testing

Software

- BUGS (Bayesian Inference using Gibbs Sampling) package (public domain, Cambridge)^{50, 27}
- Available for variety of computer systems
- <http://www.mrc-bsu.cam.ac.uk/bugs>
- <http://muskie.biostat.umn.edu/mirror/methodology/bugs/>.
- Works in conjunction with any version of S-PLUS using BUGS' CODA S-PLUS functions
- BUGS has a general modeling language
- Two-volume Examples Guide is must reading!

GUSTO I

- Four thrombolytic strategies for acute MI, $n = 41,021$ ⁴⁹
- SK=streptokinase, Combo=SK+t-PA
- Here consider only death \cup disabling stroke

Treatment	N	Events	Fraction
<i>t</i> -PA	10393	712	0.068
Combo	10370	783	0.076
SK+IV	10409	811	0.078
SK+SQ	9837	752	0.076
SK	20246	1563	0.077

- *t*-PA:SK OR=0.879, $2P = 0.006$
- Bayesian analysis using 3 different priors
 - Flat (log OR Gaussian with variance 10^6)
 - log OR truncated Gaussian with
 - $\Pr[OR > 4 \cup OR < \frac{1}{4}] = 0$
 - * $\Pr[OR > 2 \cup OR < \frac{1}{2}] = 0.05$

$$* \Pr[OR > 1\frac{1}{3} \cup OR < \frac{3}{4}] = 0.05$$

- Similarity region: $OR \in [0.9, \frac{1}{0.9}]$

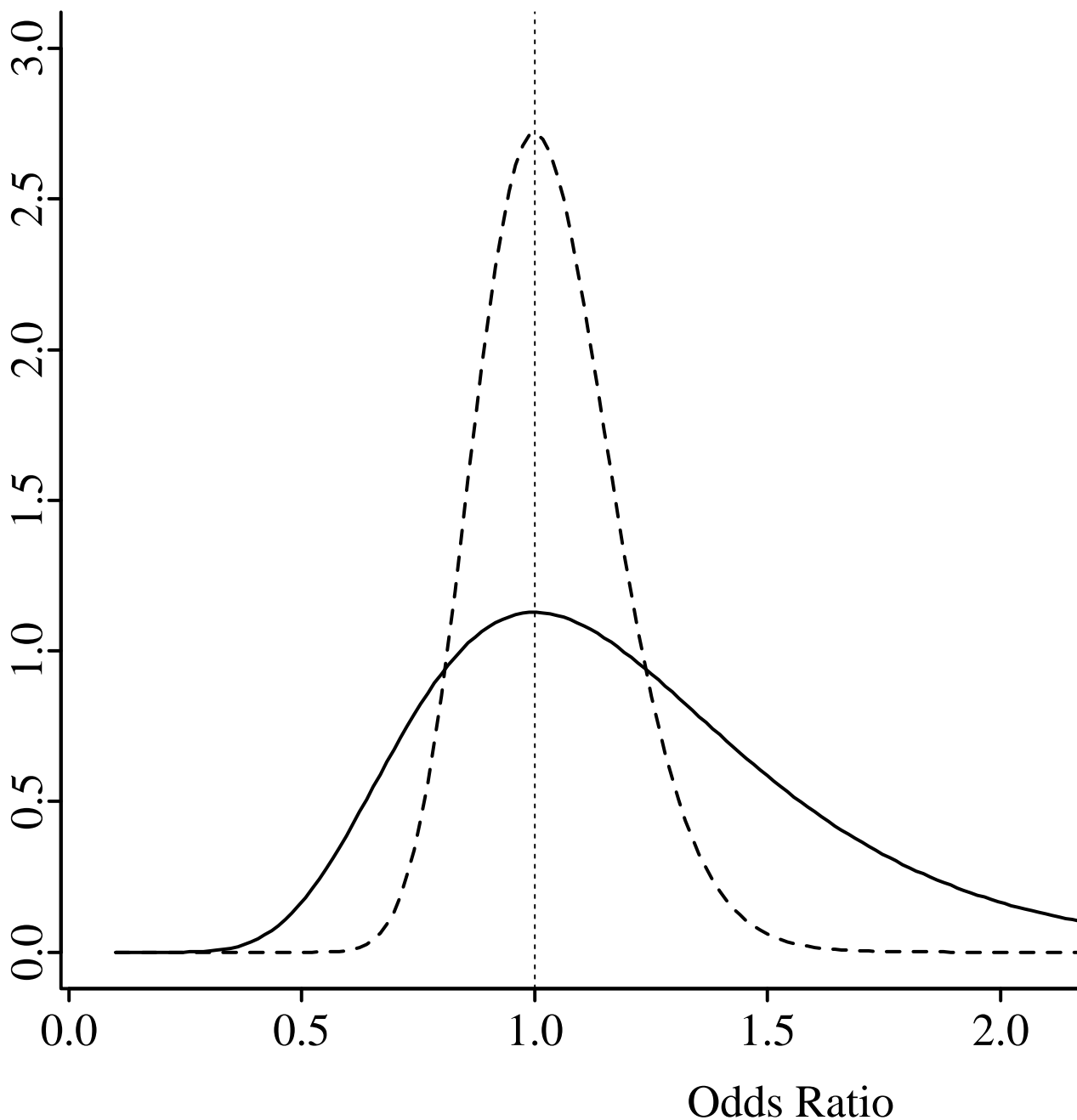
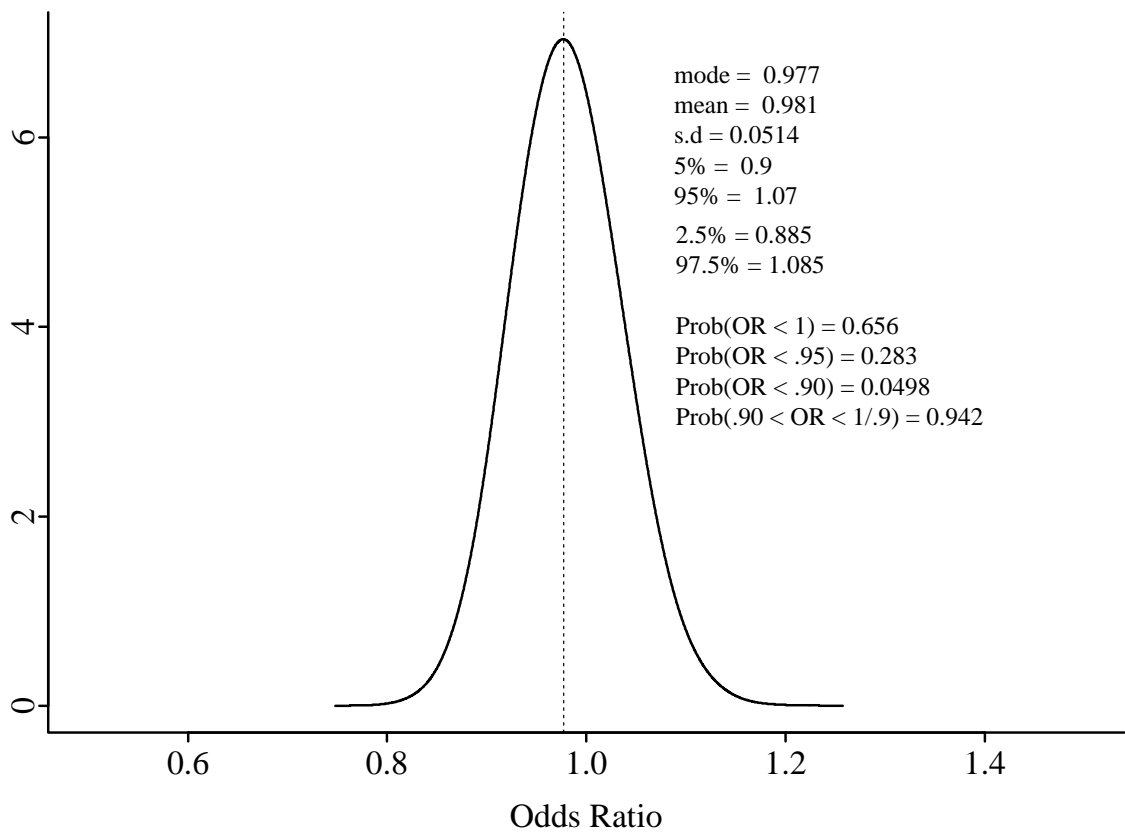


Figure 2: *Prior probability densities for $OR = e^\beta$. Both distributions assume that $OR = 1$ (no effect) is the most likely value, and that ORs outside the interval $[\frac{1}{4}, 4]$ are impossible. The solid curve corresponds to a truncated normal distribution for $\log OR$ having a standard deviation of 0.354. The dashed curve corresponds to a more skeptical prior distribution with a standard deviation of 0.147.*



SK+SQ Heparin vs. SK+IV Heparin 11Jan96 10:52

Figure 3: *Posterior probability density for the ratio of the odds of a clinical endpoint for SK+SQ heparin divided by the odds for SK+IV heparin, using a flat prior distribution for log OR.*

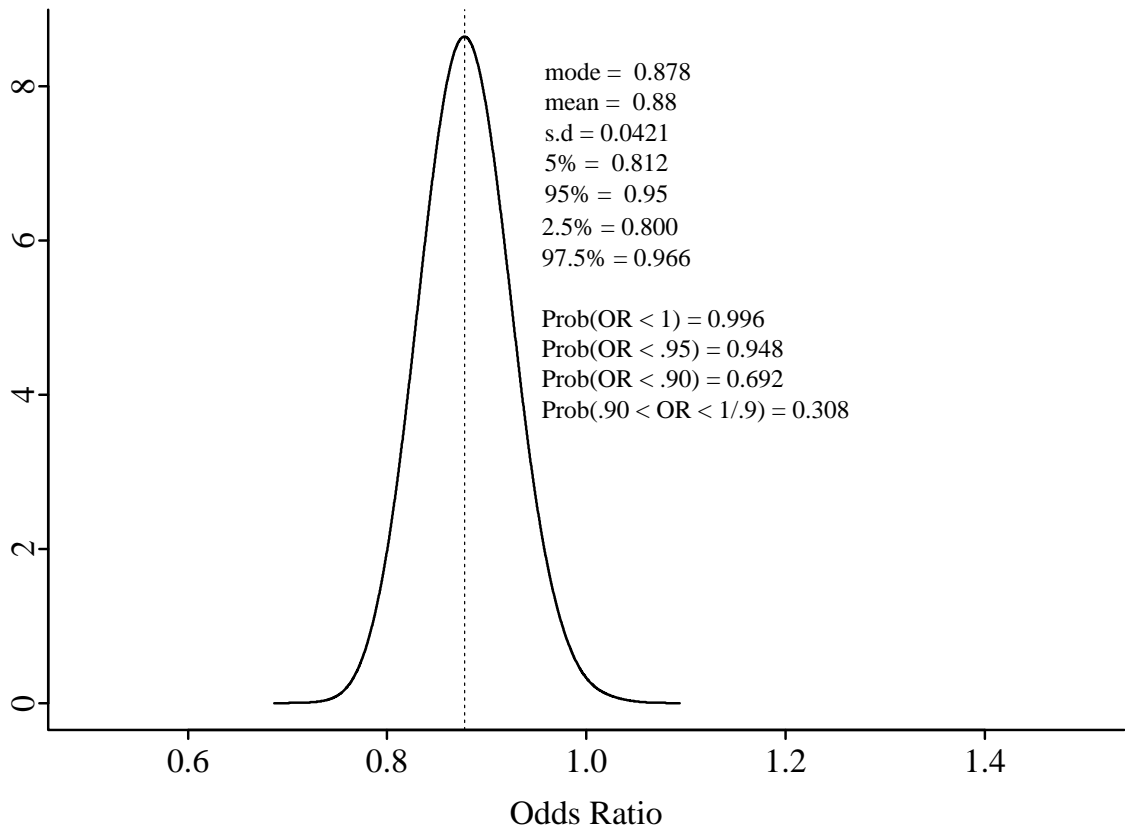
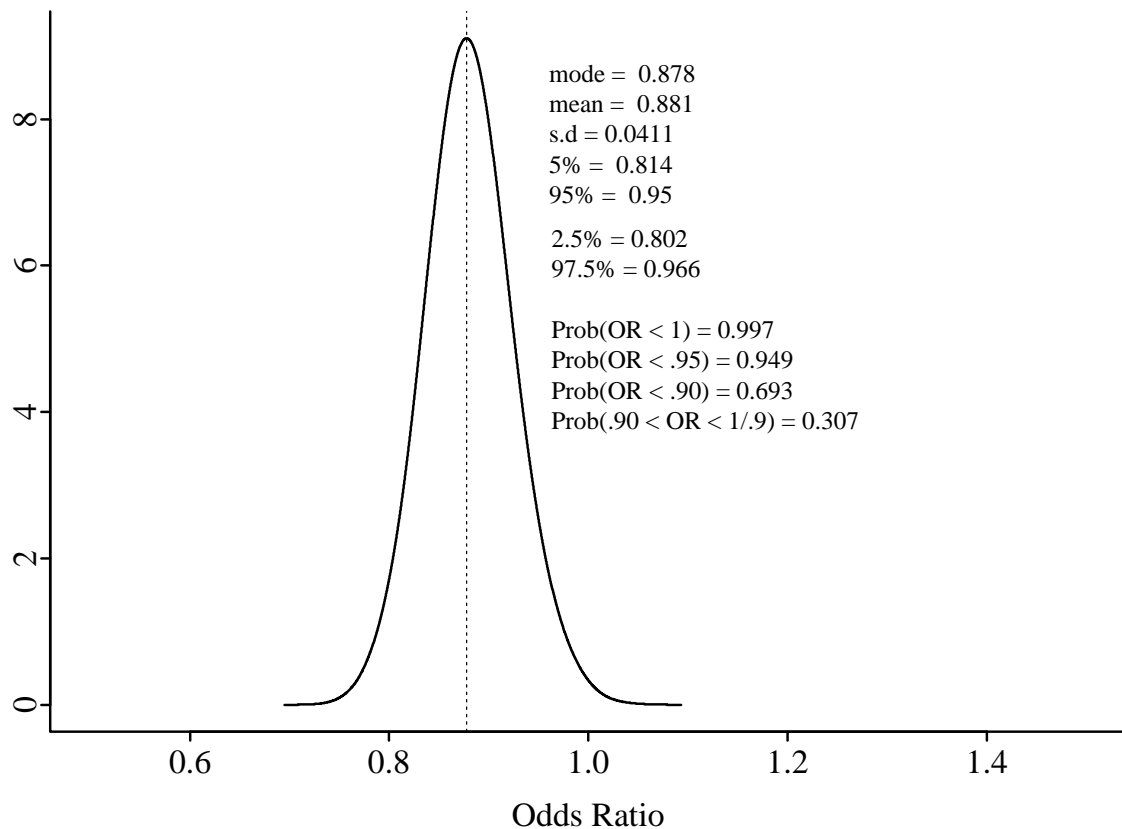
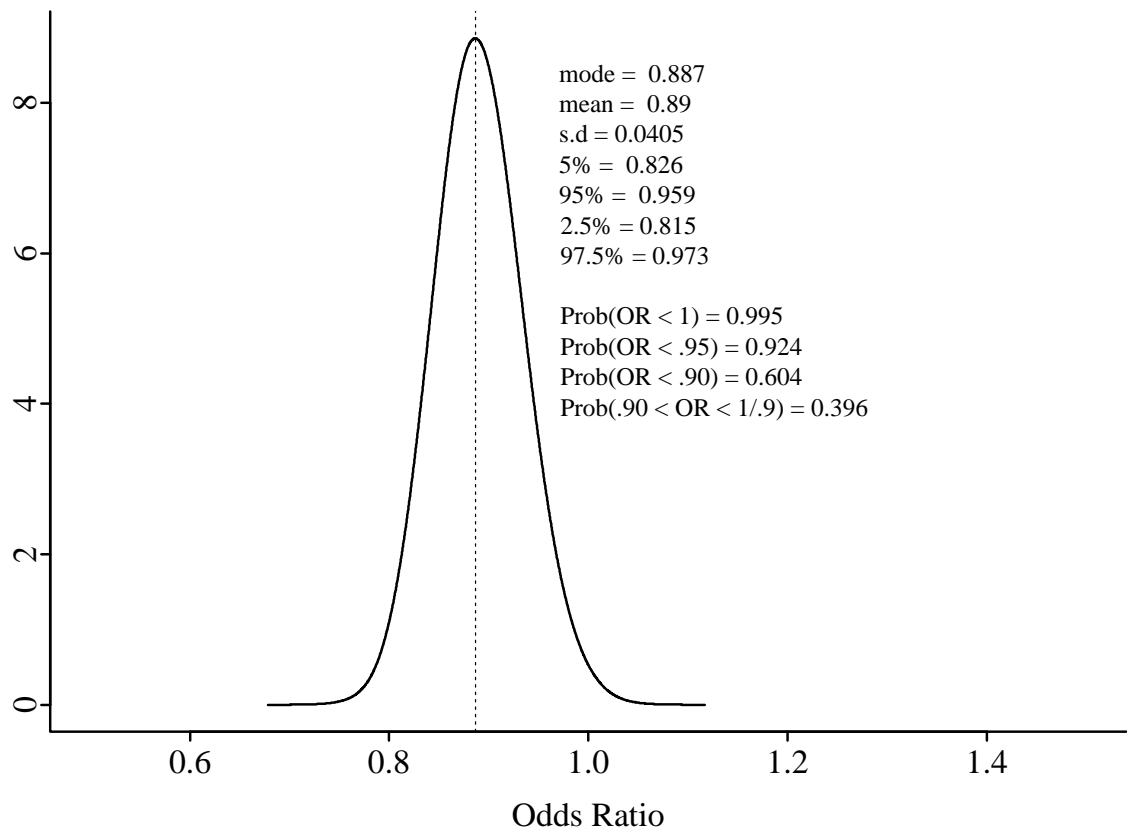


Figure 4: *Posterior probability density for accelerated t-PA compared with SK, using a flat prior for log OR.*



Combined SK vs. Accelerated t-PA Skeptical prior 11Jan96 11:36

Figure 5: *Posterior probability density for accelerated t-PA compared with SK, using a prior distribution which assumed that $\Pr(\text{OR} > 2) = \Pr(\text{OR} < \frac{1}{2}) = 0.025$.*



Combined SK vs. Accelerated t-PA Very skeptical prior 11Jan96 13:59

Figure 6: *Posterior probability density for accelerated t-PA compared with SK, using a prior distribution which assumed that $\Pr(\text{OR} > 1\frac{1}{3}) = \Pr(\text{OR} < \frac{3}{4}) = 0.025$.*

Meta-analysis of Short-Acting Nifedipine

- From meta-analysis of 16 randomized trials by Furberg et al.^{25a}
- Individual subjects' data not available
- Used dead/alive; studies had varying follow-up and dose
- Model: $\text{logit } p_{ij} = \alpha + \text{study}_i + \beta \times \text{dose}$
- Fixed effects for β
- Random effects for studies^b: Gaussian, σ^2 unknown but finite, has its own prior distribution $(\Gamma(10^{-4}, 10^{-4}))^c$
- Quantity of interest: 100mg : placebo odds ratio for all-cause mortality
- **Model Code** (File `bugs.bug`)

^aWith changes for the two Muller studies³⁹

^bFor a single study, sites could be treated as random effects in the same way

^cUse of a hyperprior to estimate σ^2 makes this similar to Empirical Bayes

```

model logistic;

const
    S=16,                # no. studies
    M=32;                # no. records (2 * # studies)

var
    dead[M],
    dose[M],
    study[M],
    N[M],
    p[M],
    int,b.dose,b.study[S],c.study[S],tau,sigma,or;

data in "nifbugs.dat";
inits in "bugs.in";

{
    for(k in 1:S) { # make random effects sum to zero
        c.study[k] <- b.study[k] - mean(b.study[])
    }

    or <- exp(100*b.dose);

    for(i in 1:M) {
        logit(p[i]) <- int+b.dose*dose[i]+ c.study[study[i]];
        dead[i] ~ dbin(p[i],N[i]);
    }

    for(k in 1:S) {
        b.study[k] ~ dnorm(0.0, tau);
    }

    #Prior distributions

    int ~ dnorm(0.0, 1.0E-6);
    b.dose ~ dnorm(0.0, 7.989E4) I(-0.01386,0.01386);
    # trunc at or=4, .025 prob>2

    tau ~ dgamma(0.0001, 0.0001);
    sigma <- 1/sqrt(tau);
    # s.d. of random effects
}

```

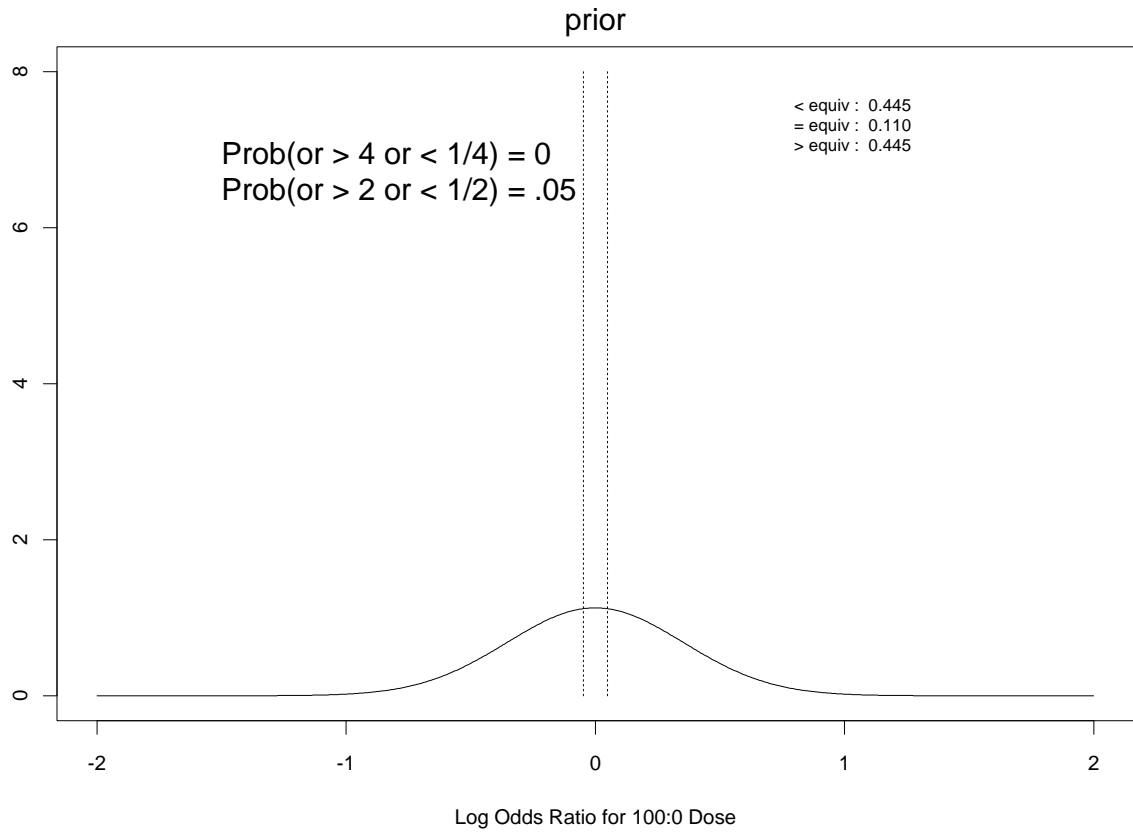


Figure 7: *Skeptical prior density for log OR; similiary (“equivalence”) zone is log odds $\in [-0.05, 0.05]$*

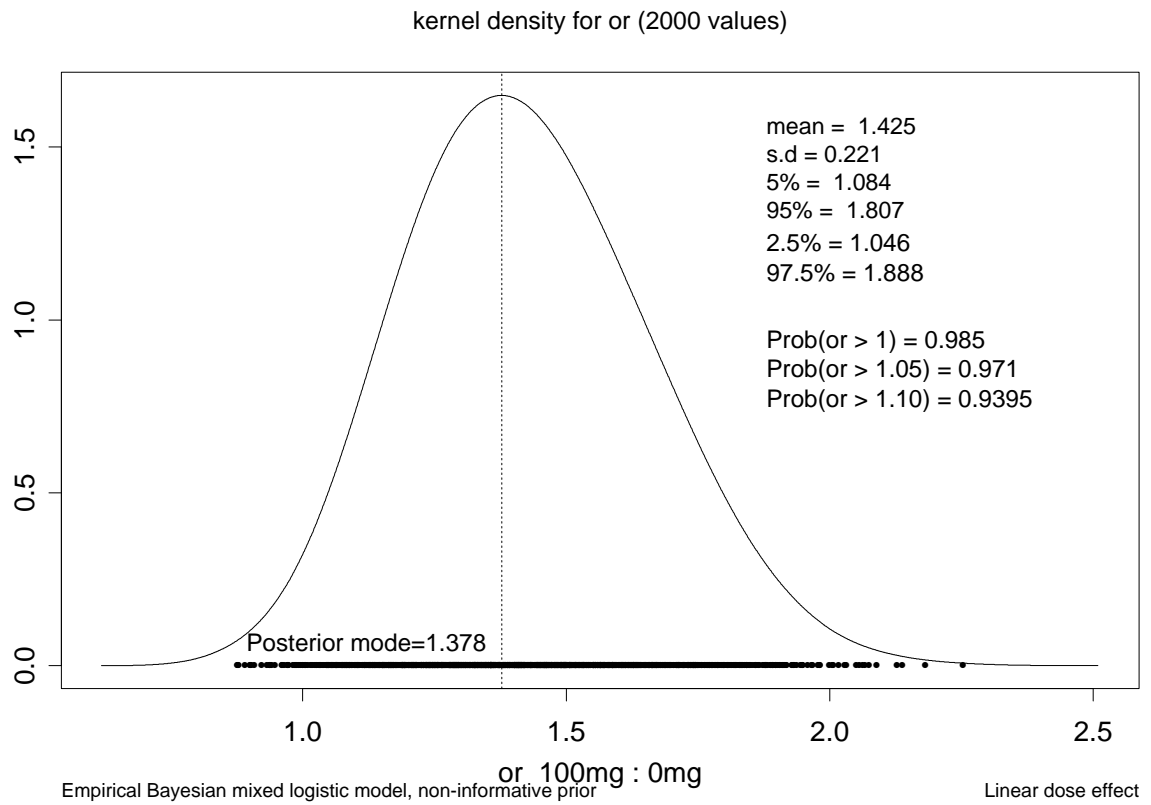


Figure 8: *Posterior density for pooled 100mg:0mg Nifedipine OR using a flat prior (Gaussian with variance 10^6) for β*

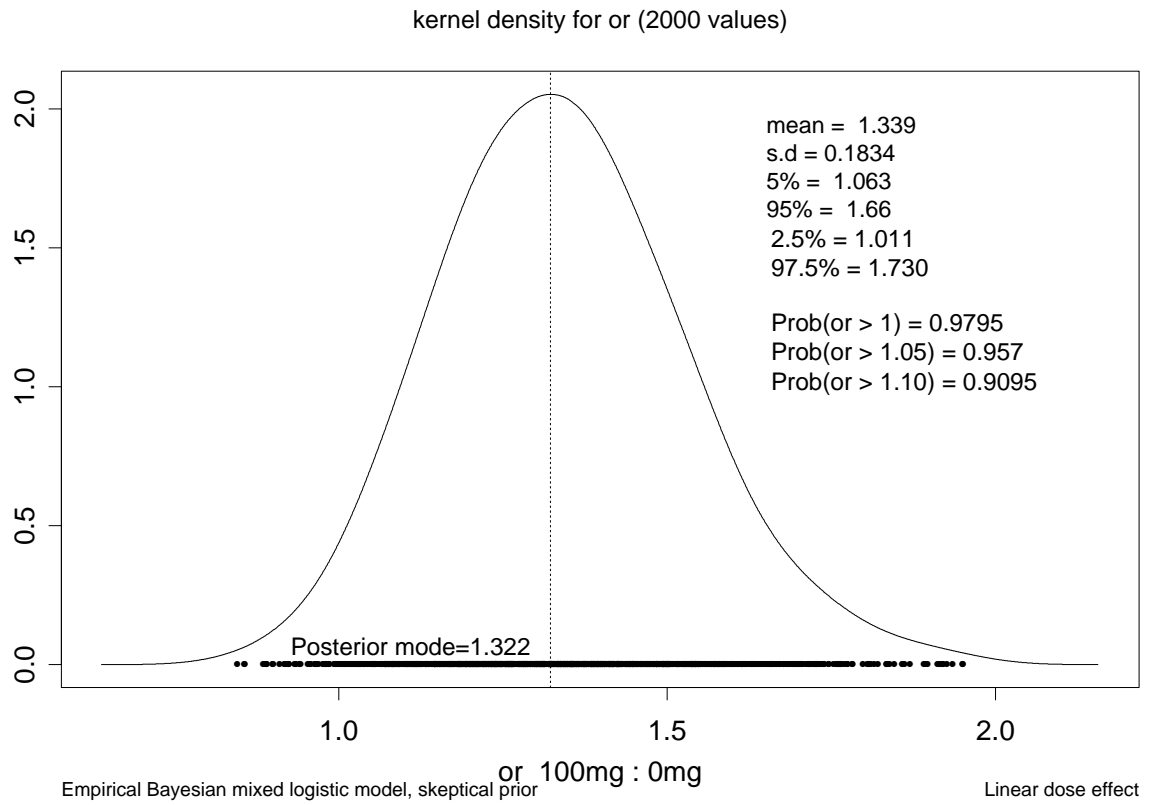


Figure 9: *Posterior density for pooled 100mg:0mg Nifedipine OR using a skeptical prior tilted toward no mortality effect*

Suggested Design Criteria

- Zone of clinical similarity is most important to pre-specify
- Mortality efficacy:
 $\Pr[OR < 1] \geq 0.95$
 $\Pr[OR < 0.9] \geq 0.9$
- Mortality safety: $\Pr[OR > 1] \geq 0.9$
- Similarity in an efficacy study:
 $\Pr[0.9 \leq OR \leq \frac{1}{0.9}] \geq 0.8$
- Similarity study: $\Pr[OR \leq \frac{1}{0.9}] \geq 0.9$
- Can accommodate relative and absolute effects simultaneously:
 $\Pr[OR < 0.9 \cup \theta_2 < \theta_1 - 0.05] > 0.9$
- Can base sample sizes on standard C.L.-based formulas^{13, 10}, Bayesian confidence interval widths³³, or use predictive distributions⁴⁶

Implications for Design/Evaluation

- Many studies overoptimistically designed
 - Tried to detect a huge effect (one much larger than clinically useful) → n too small
 - Power calculation based on variances from small pilot studies
- Some studies can have lower sample sizes, e.g., more aggressive monitoring/termination, one-tailed evaluation, no need to worry about spending α
- Some studies will need to be larger because we are more interested in estimation than point-hypothesis testing or because we want to be able to conclude that a clinically significant difference exists
- Studies can be much more flexible
 - Adapt treatment during study
 - Unplanned analyses

- With continuous monitoring, studies can be better designed — bailout still possible
- Can extend a promising study
- Reduce number of small, poorly designed studies
- Reduce distinction between Phase II and III studies
- Most scientific approach is to experiment until you have the answer
- Allow for aggressive, efficient, better designs

Summary

- Bayesian analysis actually reduces time spent in arguing about statistical tests/designs!
- Substitutes an argument about the choice of a prior for the following arguments:
 - Which treatment effect to use for sample size calculations
 - One-tailed vs. two-tailed test
 - “Exact” vs. approximate P -values (conditional vs. unconditional analyses)
 - How to test for similarity
 - Multiplicity adjustments for multiple endpoints
 - Scheduling, adjustments for sequential monitoring
 - How to penalize for extending a study
 - How to translate results to clinical significance
 - How to prevent the audience from misinter-

preting a small or large P -value

References

- [1] K. Abrams, D. Ashby, and D. Errington. Simple Bayesian analysis in clinical trials: A tutorial. *Controlled Clin Trials*, 15:349–359, 1994.
- [2] V. Barnett. *Comparative Statistical Inference*. Wiley, second edition, 1982.
- [3] E. J. Bedrick, R. Christensen, and W. Johnson. A new perspective on priors for generalized linear models. *J Am Stat Assoc*, 91:1450–1460, 1996.
- [4] E. J. Bedrick, R. Christensen, and W. Johnson. Bayesian binomial regression: Predicting survival at a trauma center. *Ann Math Stat*, 51:211–218, 1997.
- [5] J. O. Berger and D. A. Berry. Statistical analysis and the illusion of objectivity (letters to editor p. 430-433). *American Scientist*, 76:159–165, 1988.
- [6] D. A. Berry. Interim analysis in clinical trials: The role of the likelihood principle. *Ann Math Stat*, 41:117–122, 1987.
- [7] D. A. Berry. Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 3, pages 79–94. Oxford University Press, 1988.
- [8] D. A. Berry, M. C. Wolff, and D. Sack. Decision making during a phase III randomized controlled trial. *Controlled Clin Trials*, 15:360–378, 1994.
- [9] M. Borenstein. The case for confidence intervals in controlled clinical trials. *Controlled Clin Trials*, 15:411–428, 1994.
- [10] M. Borenstein. Planning for precision in survival studies. *J Clin Epi*, 47:1277–1285, 1994.

- [11] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.
- [12] N. Breslow. Biostatistics and Bayes (with discussion). *Statistical Science*, 5:269–298, 1990.
- [13] D. R. Bristol. Sample sizes for constructing confidence intervals and testing hypotheses. *Stat Med*, 8:803–811, 1989.
- [14] J. M. Brophy and L. Joseph. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA*, 273:871–875, 1995.
- [15] P. R. Burton. Helping doctors to draw appropriate inferences from the analysis of medical studies. *Stat Med*, 1994:1699–1713, 1994.
- [16] R. Christensen. Testing Fisher, Neyman, Pearson, and Bayes. *Ann Math Stat*, 59:121–126, 2005.
- [17] R. J. Cook and V. T. Farewell. Multiplicity considerations in the design and analysis of clinical trials. *J Roy Stat Soc A*, 159:93–110, 1996.
- [18] G. A. Diamond and J. S. Forrester. Clinical trials and statistical verdicts: Probable grounds for appeal (*note: this article contains some serious statistical errors*). *Ann Int Med*, 98:385–394, 1983.
- [19] S. S. Emerson. Stopping a clinical trial very early based on unplanned interim analysis: A group sequential approach. *Biometrics*, 51:1152–1162, 1995.
- [20] R. D. Etzioni and J. B. Kadane. Bayesian statistical methods in public health and medicine. *Annual Review of Public Health*, 16:23–41, 1995.
- [21] P. M. Fayers, D. Ashby, and M. Parmar. Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Stat Med*, 16:1413–1430, 1997.

- [22] A. R. Feinstein. *Clinical Biostatistics*. C. V. Mosby, St. Louis, 1977.
- [23] L. D. Fisher. Comments on Bayesian and frequentist analysis and interpretation of clinical trials. *Controlled Clin Trials*, 17:423–434, 1996.
- [24] L. S. Freedman, D. J. Spiegelhalter, and M. K. B. Parmar. The what, why and how of Bayesian clinical trials monitoring. *Stat Med*, 13:1371–1383, 1994.
- [25] C. D. Furberg, B. M. Psaty, and J. V. Meyer. Nifedipine: Dose-related increase in mortality in patients with coronary heart disease. *Circulation*, 92:1326–1331, 1995.
- [26] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.
- [27] W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A language and program for complex Bayesian modeling. *The Statistician*, 43:169–177, 1994. <http://www.mrc-bsu.cam.ac.uk/bugs>.
- [28] J. B. Greenhouse and L. Wasserman. Robust Bayesian methods for monitoring clinical trials. *Stat Med*, 14:1379–1391, 1995.
- [29] J. Grossman, M. K. B. Parmar, and D. J. Spiegelhalter. A unified method for monitoring and analysing controlled trials. *Stat Med*, 13:1815–1826, 1994.
- [30] D. F. Heitjan. Bayesian interim analysis of phase II cancer clinical trials. *Stat Med*, 16:1791–1802, 1997.
- [31] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, IL, 1989.
- [32] M. D. Hughes. Reporting Bayesian analyses of clinical trials. *Stat Med*, 12:1651–1663, 1993.

- [33] L. Joseph and P. Bélisle. Bayesian sample size determination for normal means and differences between normal means. *The Statistician*, 46:209–226, 1997.
- [34] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *J Am Stat Assoc*, 91:1343–1370, 1996.
- [35] J. M. Legler and L. M. Ryan. Latent variable models for teratogenesis using multiple binary outcomes. *J Am Stat Assoc*, 92:13–20, 1997.
- [36] R. J. Little. Calibrated Bayes: A Bayes/frequentist roadmap. *Appl Stat*, 60:213–223, 2006.
- [37] M. R. Nester. An applied statistician’s creed. *Appl Stat*, 45:401–410, 1996.
- [38] M. Oakes. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. Wiley, New York, 1986.
- [39] L. H. Opie and F. H. Messerli. Nifedipine and mortality: Grave defects in the dossier. *Circulation*, 92:1068–1073, 1995.
- [40] G. L. Rosner and D. A. Berry. A Bayesian group sequential design for a multiple arm randomized clinical trial. *Stat Med*, 14:381–394, 1995.
- [41] K. J. Rothman. A show of confidence (editorial). *NEJM*, 299:1362–3, 1978.
- [42] R. M. Royall. The effect of sample size on the meaning of significance tests. *Ann Math Stat*, 40:313–315, 1986.
- [43] D. J. Sargent and J. S. Hodges. A hierarchical model method for subgroup analysis of time-to-event data in the Cox regression setting. Presented at the Joint Statistical Meetings, Chicago, 1996.

- [44] M. J. Schervish. p values: What they are and what they are not. *Ann Math Stat*, 50:203–206, 1996.
- [45] L. B. Sheiner. The intellectual health of clinical drug evaluation. *Clin Pharm Ther*, 50:4–9, 1991.
- [46] D. J. Spiegelhalter and L. S. Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med*, 5:1–13, 1986.
- [47] D. J. Spiegelhalter, L. S. Freedman, and M. K. B. Parmar. Applying Bayesian ideas in drug development and clinical trials. *Stat Med*, 12:1501–1511, 1993.
- [48] D. J. Spiegelhalter, L. S. Freedman, and M. K. B. Parmar. Bayesian approaches to randomized trials. *J Roy Stat Soc A*, 157:357–416, 1994.
- [49] The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *NEJM*, 329:673–682, 1993.
- [50] A. Thomas, D. J. Spiegelhalter, and W. R. Gilks. Bugs: A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 4, pages 837–842. Clarendon Press, Oxford, UK, 1992. <http://www.mrc-bsu.cam.ac.uk/bugs>.
- [51] R. Tibshirani. Noninformative priors for one parameter of many. *Biometrika*, 76:604–608, 1989.
- [52] P. H. Westfall, W. O. Johnson, and J. M. Utts. A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, 84:419–427, 1997.