BAYESIAN CLINICAL TRIALS: WHY BOTHER?

Thomas A. Louis, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
www.biostat.jhsph.edu/˜tlouis/
tlouis@jhsph.edu

BAYESIAN ANALYSIS

1. Design a study (possibly using a Bayesian approach)
2. Specify a (hyper) Prior (possibly using the design information)
3. Collect data and compute a likelihood
4. Bayes' theorem ⇒ Posterior Distribution
5. Do something with it, possibly structured by a loss function
    - **$(\dots)^2$:** Posterior Mean
    - **$|\dots|$:** Posterior median
    - **$0/1 + c \times$ volume:** Tolerance Interval (CI)
    - **$0/1$:** Hypothesis Test/Model Choice

- Steps 1-3 should depend on goals
- Steps 4 & 5 obey the rules of probability
- Step 4 doesn't know what you are going to do in Step 5

Evidence, then decisions

Bother when you want

- Excellent Bayesian performance
- Excellent Frequentist performance
    - use priors and loss functions as tuning parameters
- To strike an effective Variance/Bias trade-off
- Full uncertainty propagation
- To design, conduct and analyze complex studies
- **Sometimes it isn't worth the bother**
- **Sometimes you are (almost) forced into it**

Design

- Everyone is a Bayesian in the design phase
- All evaluations are "preposterior," integrating over both the data (a frequentist act) and the parameters (a Bayesian act)
    - Rubin (1984), "A Bayesianly justifiable frequentist calculation"
- A frequentist designs to control frequentist risk over a range of parameter values
- A Bayesian designs to control preposterior (Bayes) risk
- Bayesian design is effective
  for both Bayesian and frequentist goals

Bayesian Design to Control Frequentist CI Length

- Variance of a single observation: σ^2
- L is the maximal total length of the CI length
- For two-sided coverage probability $(1 - \alpha)$:

$$\mathbf{n}(\boldsymbol{\sigma}, \mathbf{L}, \boldsymbol{\alpha}) = \mathbf{4Z^2} \left(\frac{\boldsymbol{\sigma}}{\mathbf{L}}\right)^{\mathbf{2}}$$

- If we don't know σ^2, then CI length is a RV
- Can do a series of "what ifs" or a "worst case"
- Can use a probability distribution (Bayes): **$[\sigma^2$ | prior $]$**
- Can also adapt: $[\sigma^2 \mid \mathbf{Y}_{\text{available}}\text{ , prior}]$

Frequentist CI Length: The Bayesian approach

- Background data or prior elicitation provide,

$$\begin{aligned}[\sigma^2|\text{data/opinion}] &\sim G \; \{\text{e.g., log-normal}\} \\ E(\sigma^2|\text{data/opinion}) &= \bar{\sigma}^2 \\ \mathit{CoefVar}(\sigma^2|\text{data/opinion}) &= \eta\end{aligned}$$

- Goals:

$$E_G(\text{CI length}|\text{design}_n) < L$$

$$pr_G(\text{CI length} > L|\text{design}_n) \leq \gamma$$

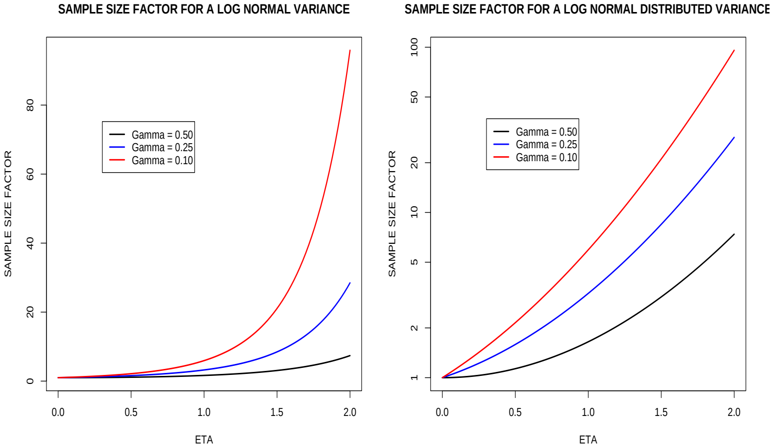
- Similarly, for testing:

$$pr_G(\text{Power} < 0.84|\text{design}_n) \leq \gamma)$$

- More generally,

$$pr_G(\text{Bayes risk} > R^*|\text{design}_n) \leq \gamma$$

CI Length: Sample size factors relative to knowing σ



- Monitor to adjust sample size in the context of accruing information on σ^2

The Basic, Hierarchical Model

$$\begin{aligned}
[\boldsymbol{\theta} \mid \boldsymbol{\eta}] &\sim g(\cdot|\boldsymbol{\eta}) \quad \textbf{Prior} \\
[\mathbf{Y}|\boldsymbol{\theta}] &\sim f(\mathbf{y}|\boldsymbol{\theta}) \quad \textbf{Likelihood} \\
g(\boldsymbol{\theta}|\mathbf{y},\, \boldsymbol{\eta}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})}{f_G(\mathbf{y}|\boldsymbol{\eta})} \quad \textbf{Posterior} \\
f_G(\mathbf{y}|\boldsymbol{\eta}) &= \int f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta} \quad \textbf{Marginal}
\end{aligned}$$

Or, Bayes empirical Bayes via a hyper-prior (H),

$$g(\boldsymbol{\theta}|\mathbf{y}) = \int g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta})h(\boldsymbol{\eta}|\mathbf{y})d\boldsymbol{\eta}$$

Compound Sampling, the Objectivity Enabler

Shrinkage, Variance Reduction, Borrowing Information

Multiple draws from the prior: Gaussian Case

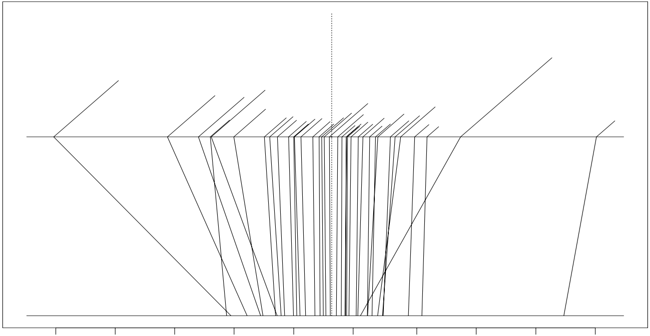
$$\begin{aligned}
\theta_1, \ldots, \theta_K \quad & iid \quad N(\mu, \tau^2) \\
[Y_k \mid \theta_k] \quad & ind \quad N(\theta_k, \sigma_k^2) \\
[\theta_k \mid Y_k] \quad & \sim \quad N\left(\mu + (1 - B_k)(Y_k - \mu), (1 - B_k)\sigma_k^2\right) \\
B_k \quad & = \quad \frac{\sigma_k^2}{\sigma_k^2 + \tau^2}
\end{aligned}$$

EB when $\sigma_k^2 \equiv \sigma^2$ (column means with equal n):

$$\begin{aligned}
\hat{\mu} \;&=\; \mathbf{Y}_\bullet \\
\hat{\tau}^2 \;&=\; (\mathbf{S}^2 - \sigma^2)^+ = \sigma^2(\mathbf{F} - 1)^+
\end{aligned}$$

Toxoplasmosis Rates in Guatemala and Honduras

top(MLEs), whiskers(SEs), bottom(Posterior Means)



- The relatively high-SE estimates are pulled in more, reducing MSE by striking an effective variance/bias trade-off

Historical Controls

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	50	50	100

- Fisher's exact one-sided P = 0.121
- But, scientists get excited:
  - "The 3 tumors are **Biologically Significant"**
- Statisticians protest:
  - "But, they aren't **Statistically Significant"**

Include Historical Data

- Same species/strain, same Lab, recently
- 0 tumors in 450 control rodents

Pooled Analysis

	C	E	Total
Tumor	0	3	3
No Tumor	500	47	547
	500	50	550

- Fisher's exact one-sided $P \doteq .0075$
- **Biological and Statistical significance!**

Bringing In History

- Control rates are drawn from a Beta(μ, M)
- Use all of the data to estimate μ and M
- Give the historical data weight equivalent to a sample size of $\widehat{M}$ with rate $\hat{\mu}$
- Female, Fisher F344 Male Rats, 70 historical experiments (Tarone 1982)

Tumor	N	$\widehat{M}$	$\hat{\mu}$	$\frac{\widehat{M}}{N}$
Lung	1805	513	.022	28.4%
Stromal Polyp	1725	16	.147	0.9%

- **Adaptive down-weighting of history**

Design and Analysis for Cluster Randomized Studies

Setting

- Compare two weight loss interventions
- Randomize clinics in pairs, one to A and one to B
- Compute clinic-pair-specific comparisons combine over pairs
- How to design and how to analyze,
  especially with a small number of clinics?

The equal sample size, unpaired case

- There are K clusters
- Within-cluster sample sizes are $n_k \equiv n$
- The V(treatment comparison), when computed under the assumption of independence is V_{ind}
- Adjust this by the among-clinic variance component

$$
\begin{aligned}
V_{icc} &= V_{ind} \times \mathbf{[1 + \rho\,(n - 1)]} = V_{ind} \times \textbf{[design effect]} \\
\rho &= \tau^{\mathbf{2}}/\sigma^{\mathbf{2}} + \tau^{\mathbf{2}} \text{ (the ICC)} \\
\tau^2 &= \left(\frac{\rho}{1-\rho}\right)\sigma^2 \quad \text{(the among-clinic variance)} \\
\sigma^2 &= \text{single-observation variance}
\end{aligned}
$$

Design and Analysis Considerations

- In the paired-clinic case, to compute

$$V_{icc} = V(\text{treatment comparison}),$$

  need to account for the following variances:
- Individual measurement (σ^2)
    - The trial will provide sufficient information
- Among-clusters: within (τ_w^2) and between (τ_b^2) cluster pairs with ($\tau^2 = \tau_w^2 + \tau_b^2$)

The need for an informative prior

- With a small number of clusters, the trial will provide little information on τ^2 and even less information on $\gamma = \tau_b^2/(\tau_w^2 + \tau_b^2)$
- Without informative priors, an “honest” computation of posterior uncertainty (one that integrates over uncertainty in τ^2 and γ) will be so large as to be useless
- Therefore, either don’t do the study or use informative priors to “bring in” outside information
- Fortunately, other weight loss studies provide credible and informative prior information on τ^2, but not so for γ
  - For γ, we need to rely primarily on expert opinion and sensitivity analysis

A Bayesian Model

- Use an informative, data-based prior for τ^2 and a small-mean, small-variance prior for γ

$$\begin{aligned} \tau^2 &\sim \text{IG:} = \tau^2_{50} \text{ with } \tau^2_{95} = 2 \times \tau^2_{50} \\ [\gamma \mid \epsilon, M] &\sim \text{Beta}(\epsilon, M) \\ E(\gamma) &= \epsilon, V(\gamma) = \epsilon(1-\epsilon)/M \end{aligned}$$

- Take the "best estimates" of (σ^2, ρ) from other cluster-randomized studies of weight change and obtain $\sigma^2 \approx (0.34)^2$, likely $\hat{\rho}$: (0.006, 0.010, 0.050)
- $\Rightarrow 10^4 \times \tau^2 = (7.0, 11.7, 60.8)$, $\mathbf{10^4\tau^2_{50} = 11.7, 10^4\tau^2_{95} = 23.4}$
- Use $\epsilon \approx 0.10$ and a relatively large $M = 15$
  - The 90^{th} percentile is approximately 0.20
  - Conservative in that there is little gain from pairing

Addressing non-standard and otherwise challenging goals
Bayesians have a corner on the market

- Ranks and Histograms
- Complicated, non-linear models
- Complicated goals like adaptive design
- Regions
    - Bioequivalence & non-Inferiority
    - Inherently bivariate treatment comparisons
    - Adaptive design based on relations among parameters

Bioequivalence & Non-inferiority

- Δ is the treatment difference
- $(-\Delta_*, \Delta^*)$ is the interval of equivalence (determined by clinical/biologic/policy considerations)

  Bio-equivalence: $-\Delta_* \leq \Delta \leq \Delta^*$

  Non-inferiority: $-\Delta_* \leq \Delta$ (negative Δ is inferior)

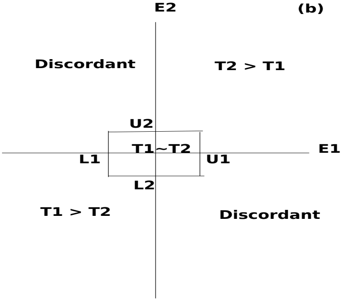
- Compute relevant posterior probabilities and design so that these will be sufficiently extreme under parameter scenarios of interest
- Can use this formalism to produce desired frequentist properties

Inherently bivariate treatment comparisons

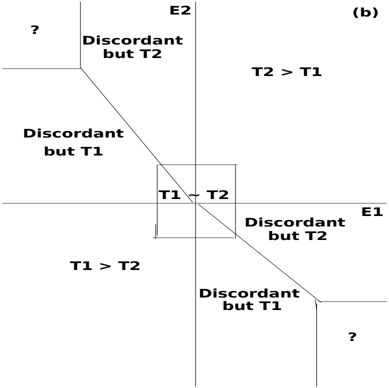
- Compare two treatments based on a bivariate outcome
  - Viral load and CD_4
  - Efficacy and SAE incidence
- Construct R^2 regions of equivalence and advantage
- Inherently R^2 regions can capture clinically important trade-offs
  - But, only generalized rectangles result from combining single-endpoint, univariate regions
- The Bayesian formalism is needed to compute,

$$\mathbf{pr}\ \mathbf{(region} \mid \mathbf{data)}$$

Combining endpoint-specific, univariate regions



Inherently R^2 Regions



Adaptive design based on relations among parameters

- Single parameter assessments
  1. if $pr(\theta > \theta_{safety} > 0 \mid \text{data}) > 0.20$, stop
  2. if $pr(\theta < \theta_{efficacy} < 0 \mid \text{data}) > 0.98$, stop
  3. if pr(either 1 or 2 by end of study $\mid$ data) > 0.90, continue as is, otherwise, either stop for futility or increase accrual/clinics
     - Requires simulating futures, conditional on current information
     - This requires assumptions on accrual, dropouts, cross-overs, . . .
- Parameter relations
  - if $\text{pr}(\text{Rel}(\theta_1, \theta_2) > 0 \mid \text{data}) > 0.98$, stop

Don't insist on strict frequentist goals

Continue or stop a dose

- Start with doses $(d_1, \ldots, d_m)$
- $P(d, \boldsymbol{\theta}) = pr(\text{favorable response} \mid d, \boldsymbol{\theta})$
  - If $P(d, \boldsymbol{\theta} \mid \text{data}) \geq 0.75$, continue accruing to the dose
  - If $P(d, \boldsymbol{\theta} \mid \text{data}) < 0.75$, stop accruing to the dose
- More generally, when allocating to doses, trade-off gaining information on $\boldsymbol{\theta}$ and doing the best for the next patient

Allocation on Outcome

- Controversial in clinical trials, but can be effective
- Less controversial: Adaptive randomization stratification
- Best approaches use Bayesian structuring for either Bayes or Frequentist goals

$\approx$ Louis 1975 Biometrika

- Gaussian Responses, treatments T_A and T_B
- SPRT Stopping based on the likelihood-ratio (L_{mn}) after m responses T_A and n on T_B
    - Continue if $0 < A < L_{mn} < B < \infty$
    - No maximum accrual
- For non-anticipating, adaptive allocation rules, frequentist type I and II errors are controlled

Approximately the Louis (1975) rule

- $\pi_{mn} = pr(T_B > T_A \mid \text{data}) = L_{mn}/(1 + L_{mn})$ for a 50/50 prior
  - Can use $\pi_{00} \neq 0.5$, but equipoise requires close to 0.5

- Select an imbalance parameter: $0.5 \leq \phi < 1.0$
- Allocate to keep

$$m/(m+n) \approx \phi\pi_{mn} + (1-\phi)(1-\pi_{mn})$$

Simulation Results, Treatment A is better

$100\phi \rightarrow$	50	55	70
M_ϕ	78.2	87.6	127.5
N_ϕ	77.7	71.7	57.2
$M_\phi + N_\phi$	155.9	159.3	184.7
Cost	0	**3.4**	**28.8**
Benefit	0	**6.0**	**20.5**

- M_ϕ and N_ϕ are expected sample sizes
- Cost $= (M_\phi + N_\phi) - (M_0.5 + N_0.5)$
- Benefit $= N_{0.5} - N_\phi$

Bayes & Multiplicity

- The prior to posterior mapping doesn't "know" about multiple comparisons
- With additive, component-specific losses each comparison is optimized separately with no accounting for the number of comparisons
- However, use of a hyper-prior (or EB) links the components since the posterior "borrows information"
  - Inducing shrinkage as a multiplicity control
- If collective penalties are needed, use a multiplicity-explicit loss function

The k-ratio, Z test

RE ANOVA

- $\theta_1, \ldots, \theta_K \quad iid \quad N(\mu, \tau^2)$
- $[Y_{ik} \mid \theta_k] \quad ind \quad N(\theta_k, \sigma^2)$
- $[\theta_k \mid Y_{.k}] \quad \sim \quad N\left(\mu + (1-B)(Y_{.k} - \mu), (1-B)\frac{\sigma^2}{n}\right)$

$$F \quad = \quad 1/\hat{B}$$

Compare columns 1 and 2:

$$Z_{12}^{Bayes} = Z_{12}^{freq}\left\{\frac{(F-1)^+}{F}\right\}^{\frac{1}{2}} = \left(\frac{\sqrt{n}(Y_{.1}-Y_{.2})}{\hat{\sigma}\sqrt{2}}\right)\left\{\frac{(F-1)^+}{F}\right\}^{\frac{1}{2}}$$

Comments

- The magnitude of F adjusts the test statistic
- For large K, under the global null hypothesis ($\tau^2 = 0$), pr[all $Z_{ij} = 0$] ≥ 0.5
- The FW rejection rate is much smaller than 0.5
- "Scoping" is important because the number of candidate comparisons influences the value of $\hat{\mu}$ and $\hat{B}$ and performance more generally
- Non-additive loss functions can be used
  - e.g., $1 + 1 = 2.5$
- These link inferences among components in addition to that induced by shrinkage

Bayes and Subgroups: HDFP

- Randomized between Referred Care (RC) and Stepped Care (SC)
- Outcome: 5-year death rate, overall and in 12 strata
- $Y = 1000 \log[\mathrm{OR(SC{:}RC)}]$
- Strata
  - Initial diastolic blood pressure

    $$\begin{array}{rcr} \mathrm{I} & = & 90\text{-}104 \\ \mathrm{II} & = & 105\text{-}114 \\ \mathrm{III} & = & \geq 115 \end{array}$$
  - Race (B/W)
  - Gender (F/M)

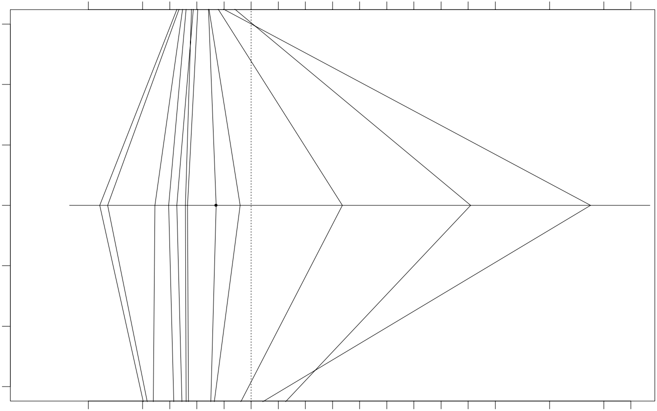
HDFP Results

Group		Y	$\hat{\theta}$	$1-B$	$\hat{\sigma}$	PSD
I	BM	–129	–157	54	170	125
	BF	–304	–240	44	206	137
	WM	–242	–220	59	153	117
	WF	–355	–253	39	231	144
II	BM	–274	–213	29	290	155
	BF	–529	–266	23	337	161
	WM	–41	–156	22	349	162
	WF	809	–61	13	479	171
III	BM	–558	–273	23	337	161
	BF	–235	–197	18	389	166
	WM	336	–122	13	483	171
	WF	1251	–103	6	730	178

All posterior means are negative

HDFP Subgroup Analysis: Ensemble Estimates

$(1-B)^{\frac{1}{2}}$ on data rather than $(1-B)$



Top:PMs Middle:MLEs Bottom:Ensemble

Bayesian Monitoring

CPCRA-TOXO: Prevention of Toxoplasmosis

- Eligibility
  - Either an AIDS defining illness
    or CD4 < 200
  - A positive titre for *toxoplasma gondii*
- Originally designed with four treatment groups
  - Active & placebo clindamycin, 2:1
  - Active & placebo pyrimethamine, 2:1
- The clindamycin arm was stopped after a few months
- We look at PYRI vs Placebo

Analysis of the Toxo Trial

WE

- Used the Cox model
  - Adjusted for baseline CD4
- Elicited priors from three HIV/AIDS clinicians, one PWA conducting AIDS research and one AIDS epidemiologist
- Monitored the trial after-the-fact
  - The DSMB monitored it during-the-fact
- "Stopped" when the posterior probability of benefit or the posterior probability of harm got sufficiently high
- Used a variety of prior distributions, including an equally-weighted mixture of the five elicited priors

The Cox Model

- Partial likelihood:

$$L(\theta_1,\ \theta_2) = \prod_{j=1}^{d} \left(\frac{e^{\theta_1 z_{1j} + \theta_2 z_{2j}}}{\sum_{\nu \in \mathcal{R}_j} e^{\theta_1 z_{1\nu} + \theta_2 z_{2\nu}}} \right)$$

- d is the number of individuals experiencing the endpoint (death or TE)
- $\mathcal{R}_j$ is the j^{th} risk set
  - The collection of individuals alive and in the study immediately preceding the j^{th} endpoint
- Covariates
  - Treatment group status: $z_{1j} = 1$ or 0 a.a. person j received pyrimethamine or placebo
  - CD4 cell count at study entry: (z_{2j})
- Negative values of θ_1 indicate a benefit for pyrimethamine

Prior Distributions

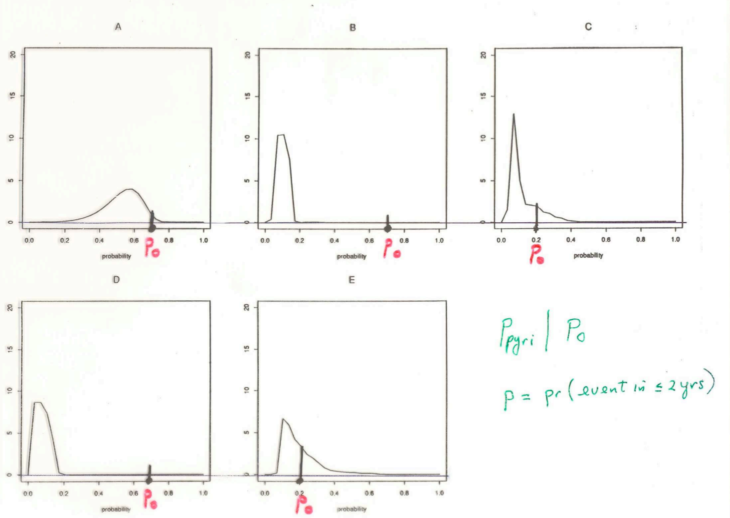
- We put a flat prior on the CD4 effect (θ_2)
- We elicited priors for the Pryimethamine effect (θ_1)

Elicitation

- Ask about potential observables
- P = pr[event in two years]
- P_0 = best guess for the placebo
    - mode, median, mean
- Then, distribution of $P_{pyri} \mid P_0$
    - percentiles
    - draw a picture
- Convert to Cox model parameter:

$$\theta_1 = \log(1 - P_0) - \log(1 - P_{pyri})$$

Elicited Priors



$P_{pyri} \mid P_0$

$p = pr(\text{event in} \le 2\text{yrs})$

Fig 2: the prior distributions on the probabilities

Actual TOXO Monitoring

- Monitored for file closing dates:
  01/15/91, 07/31/91, and 12/31/91
- At its final meeting the board recommended stopping
- The pyrimethamine group had not shown significantly fewer TE events and the low overall TE rate made a statistically significant difference unlikely to emerge.
- Also, an *increase* in the number of deaths in the pyrimethamine group was noted

Posteriors for a flat prior

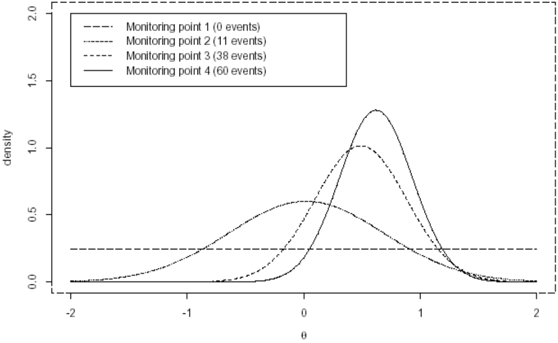
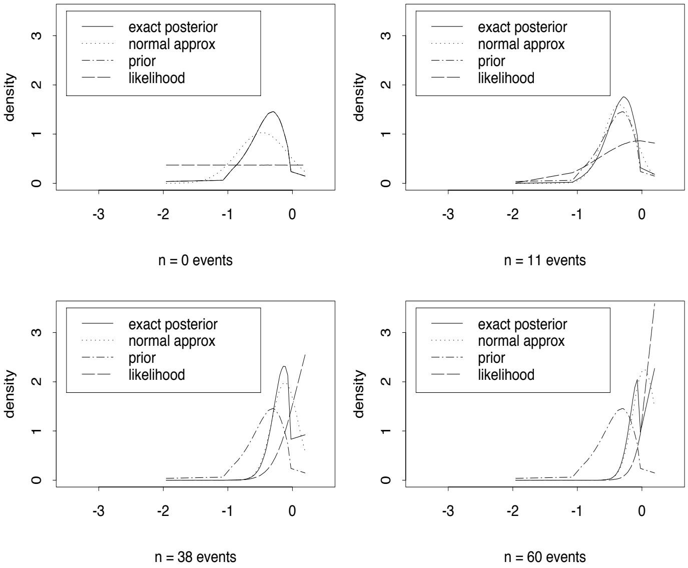


Figure 3: Posterior for the treatment effect under a flat prior, TE trial data. Endpoint is TE or death; Covariate is baseline CD4 count

Various Posterior Distributions



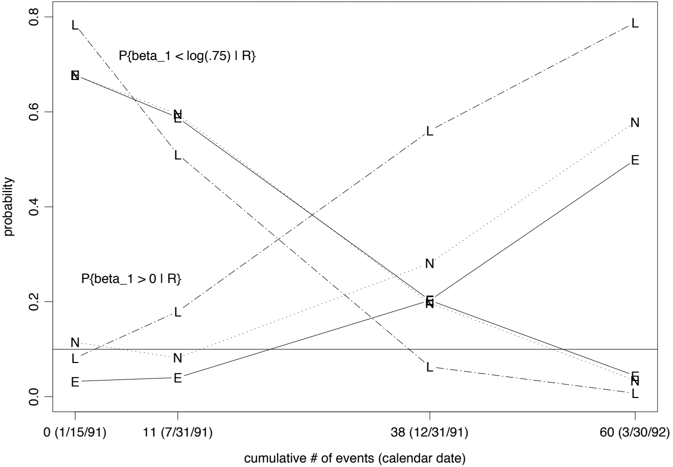
n = 0 events

n = 11 events

n = 38 events

n = 60 events

Posterior Probabilities of regions
(Bayes can take longer to stop!)



E = exact; N = normal approximation; L = likelihood

After the Fact Monitoring

- The elicited priors bear almost no resemblance to the eventual data
- Our experts believed
  - That TE is common in this patient population
  - That pyrimethamine has a substantial prophylactic effect
- Yet, eventually the data overwhelmed the elicited priors

Would it have been ethical to wait so that these experts were convinced?

Summary

- There have been many Bayesian successes, but much remains to be done
  - Methodologically
  - Sociologically
- CDRH, its encouragement and guidance have accelerated adoption and innovation
  - *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*
- The CDRH stem cell is seeding metastases to other FDA Centers

Recommendations

1. Encourage Bayesian design for frequentist analysis
    - To promote formal assembly of prior information
    - To produce realistic designs in the context of important uncertainties
2. Encourage use of the Bayesian formalism to develop all monitoring plans

    ◦ Sample size adjustment, accrual termination, follow-up termination (for efficacy or curtailment)
    - Priors and losses as tuning parameters for frequentist goals
    - Bayesian goals
3. Evaluate and introduce fully Bayesian designs and analyses

Closing

- Potential Bayesian benefits are substantial, but validity and effectiveness require expertise and care
- Bayes isn't always worth the bother, but acceptance and benefits burgeon
- The philosophy and formalism are by no means panaceas
- There are no free lunches in statistics

Happily, there are a broad array of reduced-price meals

Many based on Bayesian recipes!