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
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 $\Rightarrow$ Posterior Distribution
5. Do something with it, possibly structured by a loss function
   - $(\ldots)^2$: Posterior Mean
   - $| \ldots |$: Posterior median
   - $0/1 + c \times \text{volume}$: Tolerance Interval (CI)
   - $0/1$: Hypothesis Test/Model Choice
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 \(\Rightarrow\) Posterior Distribution
5. Do something with it, possibly structured by a loss function
   - \((\ldots)^2\): Posterior Mean
   - \(|\ldots|\): Posterior median
   - \(0/1 + c \times \text{volume}\): 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
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)
• 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: $\sigma^2$
- $L$ is the maximal total length of the CI length
- For two-sided coverage probability $(1 - \alpha)$:
  $$n(\sigma, L, \alpha) = 4Z^2 \left( \frac{\sigma}{L} \right)^2$$

- If we don't know $\sigma^2$, then CI length is a RV
Bayesian Design to Control Frequentist CI Length

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

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

- If we don’t know $\sigma^2$, then CI length is a RV
- Can do a series of “what ifs” or a “worst case”
- Can use a probability distribution (Bayes): $[\sigma^2 \mid \text{prior}]$
- Can also adapt: $[\sigma^2 \mid Y_{\text{available}}, \text{prior}]$
Frequentist CI Length: The Bayesian approach

- Background data or prior elicitation provide,
  
  \[
  [\sigma^2|\text{data/opinion}] \sim G \{\text{e.g., log-normal}\}
  \]
  
  \[
  E(\sigma^2|\text{data/opinion}) = \bar{\sigma}^2
  \]
  
  \[
  CoefVar(\sigma^2|\text{data/opinion}) = \eta
  \]

- Goals:
  \[
  E_G(\text{CI length} | \text{design}_n) < L
  \]
Frequentist CI Length: The Bayesian approach

- Background data or prior elicitation provide,

\[
\sigma^2 | \text{data/opinion} \sim G \{\text{e.g., log-normal}\}
\]

\[
E(\sigma^2 | \text{data/opinion}) = \bar{\sigma}^2
\]

\[
\text{CoefVar}(\sigma^2 | \text{data/opinion}) = \eta
\]

- Goals:

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

\[
pr_G(\text{CI length} > L | \text{design}_n) \leq \gamma
\]
Frequentist CI Length: The Bayesian approach

- Background data or prior elicitation provide,

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

- 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 $\sigma$

**SAMPLE SIZE FACTOR FOR A LOG NORMAL VARIANCE**

- Gamma = 0.50
- Gamma = 0.25
- Gamma = 0.10

**SAMPLE SIZE FACTOR FOR A LOG NORMAL DISTRIBUTED VARIANCE**

- Gamma = 0.50
- Gamma = 0.25
- Gamma = 0.10

---

Clinical Trials: Past, Present & Future  T. A. Louis: Bayesian Clinical Trials page 12
Monitor to adjust sample size in the context of accruing information on $\sigma^2$
The Basic, Hierarchical Model

\[
[\theta \mid \eta] \sim g(\cdot \mid \eta) \quad \text{Prior}
\]

\[
[Y \mid \theta] \sim f(y \mid \theta) \quad \text{Likelihood}
\]

\[
g(\theta \mid y, \eta) = \frac{f(y \mid \theta)g(\theta \mid \eta)}{f_G(y \mid \eta)} \quad \text{Posterior}
\]

\[
f_G(y \mid \eta) = \int f(y \mid \theta)g(\theta \mid \eta)d\theta \quad \text{Marginal}
\]

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

\[
g(\theta \mid y) = \int g(\theta \mid y, \eta)h(\eta \mid y)d\eta
\]
Compound Sampling, the Objectivity Enabler
Shrinkage, Variance Reduction, Borrowing Information

Multiple draws from the prior: Gaussian Case

\[ \theta_1, \ldots, \theta_K \sim iid \quad N(\mu, \tau^2) \]

\[ [Y_k \mid \theta_k] \sim ind \quad N(\theta_k, \sigma_k^2) \]

\[ [\theta_k \mid Y_k] \sim N(\mu + (1 - B_k)(Y_k - \mu), (1 - B_k)\sigma_k^2) \]

\[ B_k = \frac{\sigma_k^2}{\sigma_k^2 + \tau^2} \]

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

\[ \hat{\mu} = Y \]

\[ \hat{\tau}^2 = (S^2 - \sigma^2)^+ = \sigma^2(F - 1)^+ \]
The relatively high-SE estimates are pulled in more, reducing MSE by striking an effective variance/bias trade-off.
### Historical Controls

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>E</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>No Tumor</td>
<td>50</td>
<td>47</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>E</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>No Tumor</td>
<td>500</td>
<td>47</td>
<td>547</td>
</tr>
</tbody>
</table>

Fisher’s exact one-sided $P \approx .0075$

**Biological and Statistical significance!**
Bringing In History

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

<table>
<thead>
<tr>
<th>Tumor</th>
<th>N</th>
<th>$\hat{M}$</th>
<th>$\hat{\mu}$</th>
<th>$\frac{\hat{M}}{N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>1805</td>
<td>513</td>
<td>.022</td>
<td>28.4%</td>
</tr>
<tr>
<td>Stromal Polyp</td>
<td>1725</td>
<td>16</td>
<td>.147</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

- 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 \( V(\text{treatment comparison}) \), when computed under the assumption of independence is \( V_{ind} \)
- Adjust this by the among-clinic variance component

\[
V_{icc} = V_{ind} \times [1 + \rho (n - 1)] = V_{ind} \times [\text{design effect}]
\]

\[
\rho = \frac{\tau^2}{\sigma^2} + \tau^2 \quad (\text{the ICC})
\]

\[
\tau^2 = \left( \frac{\rho}{1 - \rho} \right) \sigma^2 \quad (\text{the among-clinic variance})
\]

\[\sigma^2 = \text{single-observation variance}\]
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 \((\sigma^2)\)
  - The trial will provide sufficient information
- Among-clusters: within \((\tau^2_w)\) and between \((\tau^2_b)\) cluster pairs with \((\tau^2 = \tau^2_w + \tau^2_b)\)
The need for an informative prior

- With a small number of clusters, the trial will provide little information on $\tau^2$ and even less information on $\gamma = \tau_w^2 / (\tau_w^2 + \tau_b^2)$.
- Without informative priors, an “honest” computation of posterior uncertainty (one that integrates over uncertainty in $\tau^2$ and $\gamma$) 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 $\tau^2$, but not so for $\gamma$.
  - For $\gamma$, we need to rely primarily on expert opinion and sensitivity analysis.
A Bayesian Model

- Use an informative, data-based prior for $\tau^2$ and a small-mean, small-variance prior for $\gamma$

  $$\tau^2 \sim IG = \tau^2_{50} \text{ with } \tau^2_{95} = 2 \times \tau^2_{50}$$

  $$[\gamma \mid \epsilon, M] \sim Beta(\epsilon, M)$$

  $$E(\gamma) = \epsilon, \quad V(\gamma) = \epsilon(1 - \epsilon)/M$$

- Take the “best estimates” of $(\sigma^2, \rho)$ 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)

  $$\Rightarrow 10^4 \times \tau^2 = (7.0, 11.7, 60.8),$$

  $$10^4 \tau^2_{50} = 11.7, \quad 10^4 \tau^2_{95} = 23.4$$

- Use $\epsilon \approx 0.10$ and a relatively large $M = 15$
  - The 90th 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

- $\Delta$ 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 $\Delta$ 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,
  $$\text{pr} \ (\text{region} \ | \ \text{data})$$
Combining endpoint-specific, univariate regions
Inherently $R^2$ Regions
Adaptive design based on relations among parameters

- **Single parameter assessments**
  1. if $pr(\theta > \theta_{\text{safety}} > 0 \mid \text{data}) > 0.20$, stop
  2. if $pr(\theta < \theta_{\text{efficacy}} < 0 \mid \text{data}) > 0.98$, stop
  3. if $pr(\text{either 1 or 2 by end of study} \mid \text{data}) > 0.90$, continue as is, otherwise, either stop for futility or increase accrual/clinics

- **Parameter relations**
  - if $pr(\text{Rel}(\theta_1, \theta_2) > 0 \mid \text{data}) > 0.98$, stop

Requires simulating futures, conditional on current information

This requires assumptions on accrual, dropouts, cross-overs, . . .
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 \( pr(\text{either 1 or 2 by end of study} \mid \text{data}) > 0.90 \), 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 \( pr(\text{Rel}(\theta_1, \theta_2) > 0 \mid \text{data}) > 0.98 \), stop

Don’t insist on strict frequentist goals
Continue or stop a dose

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

<table>
<thead>
<tr>
<th>$100\phi \rightarrow$</th>
<th>50</th>
<th>55</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_\phi$</td>
<td>78.2</td>
<td>87.6</td>
<td>127.5</td>
</tr>
<tr>
<td>$N_\phi$</td>
<td>77.7</td>
<td>71.7</td>
<td>57.2</td>
</tr>
<tr>
<td>$M_\phi + N_\phi$</td>
<td>155.9</td>
<td>159.3</td>
<td>184.7</td>
</tr>
<tr>
<td>Cost</td>
<td>0</td>
<td>3.4</td>
<td>28.8</td>
</tr>
<tr>
<td>Benefit</td>
<td>0</td>
<td>6.0</td>
<td>20.5</td>
</tr>
</tbody>
</table>

- $M_\phi$ and $N_\phi$ 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

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

$$F = \frac{1}{\hat{B}}$$

Compare columns 1 and 2:

$$Z_{12}^{Bayes} = Z_{12}^{freq} \left\{ \frac{(F-1)^+}{F} \right\}^{\frac{1}{2}} = \left( \frac{\sqrt{n}(Y_{.1} - Y_{.2})}{\hat{\sigma} \sqrt{2}} \right) \left\{ \frac{(F-1)^+}{F} \right\}^{\frac{1}{2}}$$
The magnitude of F adjusts the test statistic

For large K, under the global null hypothesis ($\tau^2 = 0$),

$$\text{pr}[\text{all } Z_{ij} = 0] \geq 0.5$$

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
- \( Y = 1000 \log[\text{OR(SC:RC)}] \)
- Strata
  - Initial diastolic blood pressure
    - I = 90-104
    - II = 105-114
    - III = \( \geq 115 \)
  - Race (B/W)
  - Gender (F/M)
## HDFP Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Y</th>
<th>$\hat{\theta}$</th>
<th>$1 - B$</th>
<th>$\hat{\sigma}$</th>
<th>PSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>BM</td>
<td>-129</td>
<td>-157</td>
<td>54</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>-304</td>
<td>-240</td>
<td>44</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>-242</td>
<td>-220</td>
<td>59</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>WF</td>
<td>-355</td>
<td>-253</td>
<td>39</td>
<td>231</td>
</tr>
<tr>
<td>II</td>
<td>BM</td>
<td>-274</td>
<td>-213</td>
<td>29</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>-529</td>
<td>-266</td>
<td>23</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>-41</td>
<td>-156</td>
<td>22</td>
<td>349</td>
</tr>
<tr>
<td></td>
<td>WF</td>
<td>809</td>
<td>-61</td>
<td>13</td>
<td>479</td>
</tr>
<tr>
<td>III</td>
<td>BM</td>
<td>-558</td>
<td>-273</td>
<td>23</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>-235</td>
<td>-197</td>
<td>18</td>
<td>389</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>336</td>
<td>-122</td>
<td>13</td>
<td>483</td>
</tr>
<tr>
<td></td>
<td>WF</td>
<td>1251</td>
<td>-103</td>
<td>6</td>
<td>730</td>
</tr>
</tbody>
</table>

All posterior means are negative
HDFP Subgroup Analysis: Ensemble Estimates

\[(1 - B)^{\frac{1}{2}} \text{ on data rather than } (1 - B)\]
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:

\[
L(\theta_1, \theta_2) = \prod_{j=1}^{d} \left( \frac{e^{\theta_1 z_{1j} + \theta_2 z_{2j}}}{\sum_{\nu \in R_j} e^{\theta_1 z_{1\nu} + \theta_2 z_{2\nu}}} \right)
\]

- \(d\) is the number of individuals experiencing the endpoint (death or TE)
- \(R_j\) 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 \(\theta_1\) indicate a benefit for pyrimethamine
Prior Distributions

- We put a flat prior on the CD4 effect ($\theta_2$)
- We elicited priors for the Pryimethamine effect ($\theta_1$)
Ask about potential observables

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

\[
\theta_1 = \log(1 - P_0) - \log(1 - P_{\text{pyri}})
\]
Elicited Priors

Fig 2: the prior distributions on the probabilities

$p_{prior} \mid p_0$

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

Clinical Trials: Past, Present & Future
Posterior Probabilities of regions
(Bayes can take longer to stop!)

\[ P\{\beta_1 > 0 \mid R\} \]

\[ P\{\beta_1 < \log(0.75) \mid R\} \]

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
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
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.**
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!