Bayesian Hierarchical Models for Subgroup Analysis in Clinical Studies

Gene Pennello, PhD
Division of Biostatistics
FDA / CDRH
Silver Spring, MD
Abstract

Bayesian hierarchical models for subgroup analysis will be introduced and considered for early phase biomarker discovery studies, pre-market validation studies, and post-market signal detection. In Bayesian hierarchical modeling, the subgroups are considered jointly. That is, the Bayesian estimate for a subgroup considers not just the data from that subgroup, but the data from all the other subgroups as well. When appropriate, joint Bayesian estimates of subgroup-specific treatment effects are attractive: they tend to have fewer “random highs”, greater precision, and an improved rank ordering relative to the point estimates for the subgroups when they are each considered alone.
The Subgroup Problem

- Patients enrolled into a clinical study can be heterogeneous.
- Treatment effects may therefore be heterogeneous among subgroups.
- If treatment effects are estimated separately within each subgroup, their variation will tend to be overstated.
- The chance of a falsely positive subgroup finding can be out of control.
• Subgroup specific treatment effects can be falsely significant (statistically or clinically).
Frequentist Subgroup Procedures

- Analyze subgroups separately, but
  - Control the *familywise* type 1 error rate, the probability of falsely rejecting at least one true null hypothesis among the family of subgroups being tested.
  - Lower significance level for subgroup hypothesis test.
  - Assess significance with adjusted p value (smallest familywise level at which subgroup is significant).
  - Widen confidence interval about subgroup estimate.

- Gatekeepers (possibly underpowered):
  - Significant treatment by subgroup interaction
  - Significant overall treatment effect
Bayesian Hierarchical Model

• Analyze subgroups not separately, but *jointly*.
• Assume subgroup treatment effects are *exchangeable*, which relates their estimation.
• Bayesian estimate (posterior mean) of a subgroup-specific treatment effect is a *weighted average* of
  – the sample estimate in the subgroup, and
  – the overall treatment effect estimate.
• Bayesian inference for a subgroup depends on
  – the data on the other subgroups, but
  – not on whether they are in the family of subgroups being formally tested (inference is *comparisonwise*)
Bayesian Hierarchical Model

• Relative to separate analyses of subgroups,
  – Bayesian subgroup estimates are *more* precise (borrows strength from overall estimate).
  – Bayesian credible intervals are *shorter* (due to borrowing strength).
  – Bayesian subgroup analysis exhibits *fewer* “random highs” (subgroup sample estimate is “shrunk” toward overall estimate).
  – Significant differences between subgroups are harder to declare (shrinkage toward overall estimate is faster than increase in precision).
Extended D&S is a 3-way Bayesian hierarchical model.
Early Clinical Studies

• In a feasibility, exploratory, or Phase II study, suppose
  – Sample Size is Small to Moderate.
  – Heterogeneous Treatment Effects are Plausible
  – Truly Zero Effects not likely.

• Bayesian HM Subgroup Analysis
  – increases precision of subgroup estimates,
  – obtains an ‘honest’ subgroup estimate (via shrinkage), useful for
    • deciding whether to validate the treatment in a pivotal or Phase III study of just one of the subgroups.
    • sizing that validation study.
Pivotal Clinical Validation

• A subgroup finding can be assessed for statistical and clinical significance using Bayesian hierarchical model subgroup analysis.

• Clinical significance of the subgroup finding can be assessed by
  – using the posterior mean shrinkage estimate and associated posterior standard deviation,
  – not by taking the subgroup sample estimate at face value.
Bayesian Analysis

• Prior Distribution
  – A priori distribution of parameter

• Likelihood
  – Distribution of data given parameter

• Posterior Distribution
  – Update of prior distribution of parameter given likelihood
Hierarchical Model: **Subgroups Have Different But Related Effects**

**Level 1:** Patients *exchangeable* within subgroups

**Level 2:** Subgroups *exchangeable*.

*All patients are related:* more so within **subgroups** than between subgroups.
Are Subgroups Exchangeable?

• Consider estimating a treatment effect.
• **Subgroups** are *exchangeable* in their treatment effects if any ordering of them is considered equally likely a priori (i.e., before seeing the data).
• Put another way, if the subgroup-specific treatment effects were revealed, but their subgroup labels were not, then the effects would not be helpful in predicting their labels.
Beta-blocker for Hypertension

Bayesian subgroup analysis
Beta-blocker for Hypertension

Losartan versus atenolol randomized trial

**Endpoint:** composite of Stroke/ MI/ CV Death

N=9193   losartan (4605),
          atenolol (4588)

# Events   losartan (508),
           atenolol (588)

80% European Caucasians 55-80 years old.

http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020386s058lbl.pdf
# Beta-blocker for Hypertension

## Cox Analysis

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>logHR</th>
<th>SE</th>
<th>HR</th>
<th>.95%CI</th>
<th>p val</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>9193</td>
<td>0.87</td>
<td>0.77,0.98</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Race Subgroups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Black</td>
<td>8660</td>
<td>–.19</td>
<td>.06</td>
<td>0.83</td>
<td>0.73,0.94</td>
<td>0.003</td>
</tr>
<tr>
<td>Black</td>
<td>533</td>
<td>.51</td>
<td>.24</td>
<td>1.67</td>
<td>1.04,2.66</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Is the Finding Among Blacks Real or a Directional Error due to Multiplicity?
Bayesian Hierarchical Model

For subgroups $i = 1, 2$ with $n_i$ patients, assume observed log hazard ratio

$$y_i \sim N(\mu_i, \sigma_i^2 / n_i),$$

$$(\text{se of } y_i)^2 = s_i^2 \sim \Gamma(f_i / 2, n_i f_i / 2\sigma_i^2)$$

Null Hypothesis: $H : \mu_i > 0$
Bayesian Hierarchical Model

Data:
\[ y_i \sim N(\mu_i, \sigma_i^2 / n_i) \]
\[ s_i^2 \sim \Gamma(f_i / 2, n_i f_i / 2\sigma_i^2) \]

Prior:
\[ \mu_i \sim N(\mu_0, \sigma_\mu^2), \quad i = 1, 2 \]
\[ \mu_0 \sim N(0, 1000) \]
\[ \sigma_i^{-2}, \sigma_\mu^{-2} \sim \Gamma(.001, .001) \]
Bayesian HM Analysis

<table>
<thead>
<tr>
<th></th>
<th>logHR</th>
<th>se/sd</th>
<th>HR</th>
<th>.95%CI</th>
<th>p val</th>
<th>Pr HR&gt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>non-black</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequentist</td>
<td>-.19</td>
<td>.06</td>
<td>0.83</td>
<td>0.73, 0.94</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Bayesian</td>
<td>-.18</td>
<td>.06</td>
<td>0.84</td>
<td>0.74, 0.95</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td><strong>black</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequentist</td>
<td>.51</td>
<td>.24</td>
<td>1.67</td>
<td>1.04, 2.67</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>Bayesian</td>
<td>.38</td>
<td>.27</td>
<td>1.47</td>
<td>0.87, 2.44</td>
<td>0.914</td>
<td></td>
</tr>
</tbody>
</table>

Bayesian analysis casts some doubt on the surprising finding among blacks. Shrinkage of effect size in blacks is predicated on exchangeability, that is, not expecting a priori a worse effect in blacks than in non-blacks.
Other Clinical Situations

- **Biomarker Discovery**
  - A huge number of biomarker defined subgroups may be examined (e.g., GWAS SNP based subgroups).

- **Post-Market Signal Detection**
  - Assess disproportionality in the reporting of an adverse event with the use of a medical product (a drug, biologic, or device).
  - Number of combinations of type of adverse event and medical product is huge.
  - For most combinations, disproportionality is not expected.

- **Variants on hierarchical model for these cases:**
  - Model subgroup effects as being drawn from a mixture of two distributions, e.g., a distribution of non-zero effects and a degenerate distribution of zero effects.
Bayesian Hierarchical Model, Normal Data, Normal Prior

Subgroups \( j = 1, 2, \ldots, J \)

Treatment effect data

\[
y_j \sim N(\mu_j, \tau_y^{-1}), \tau_y^{-1} = \sigma_y^2
\]

Prior

\[
\mu_j \sim N(\mu_0, \tau_\mu^{-1}), \tau_\mu^{-1} = \sigma_\mu^2
\]

(true treatment effects \( \{\mu_j\} \) are exchangeable)

For simplicity, assume hypermean \( \mu_0 \), variances \( \sigma_\mu^2 \), \( \sigma_y^2 \) are known.
Bayesian Hierarchical Model, Normal Data, Normal Prior

- Posterior distribution of treatment effect in subgroup $j$:

$$
\mu_j \mid \{y_j\} \sim N\left( (1-S)\mu_0 + Sy_j, S\sigma^2_y \right)
$$

$$
S = \frac{1}{1 + \Phi} \text{ is the shrinkage factor}
$$

$$
\Phi = \frac{(\sigma^2_y + \sigma^2_\mu)}{\sigma^2_y} \text{ is the 'true' F ratio}
$$
Bayesian Hierarchical Model, Normal Data, Normal Prior

Posterior distribution of treatment effect difference between two subgroups:

\[
\delta = \mu_1 - \mu_2 \mid \{y_j\} \sim N\left(Sd, S\sigma_d^2\right)
\]

\[
d = y_1 - y_2
\]

\[
\sigma_d^2 = 2\sigma_y^2
\]
Bayesian Hierarchical Model, Normal Data, Normal Prior

Critical Region:

\[ \Pr(\delta > 0 \mid \{y_j\}) > 1 - \alpha \]

iff

\[ z_d = \frac{d}{\sigma_d} > \frac{z_{1-\alpha}}{\sqrt{S}} \]

Critical \( z_d \) value \( \frac{z_{1-\alpha}}{\sqrt{S}} \uparrow \) as \( \Phi, \frac{\sigma^2_\mu}{\sigma^2_y} \downarrow \)
Fully Bayesian Critical $t$ Values
Bayesian Two-Way Hierarchical Model

\[
\bar{y}_{ij} \sim N(\mu_{ij}, \sigma^2 / n), \quad i = 1, \ldots a, \quad j = 1, \ldots b
\]

\[
fs^2 / \sigma^2 \sim \chi^2(f), \quad f = ab(r - 1)
\]

\[
\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}
\]

\[
\alpha_i \sim N(0, \sigma^2_{\alpha}), \quad \beta_j \sim N(0, \sigma^2_{\beta}), \quad \gamma_{ij} \sim N(0, \sigma^2_{\gamma})
\]

Jeffreys prior on \((\mu, \sigma^2, \sigma^2_{\alpha}, \sigma^2_{\beta}, \sigma^2_{\gamma})\)
Known Variances Inference

Subgroup Problem: \[ \delta_{12,j} = \mu_{1_j} - \mu_{2,j} \]

Posterior
\[ \delta_{12,j} \mid y, \sigma^2 \sim N(S_A d_{12.} + S_C d_C, (S_A + (b-1)S_C)\sigma_d^2 / b)) \]

\[ d_{12.} = \bar{y}_1 - \bar{y}_2, \quad d_C = d_{12,j} - d_{12.} \]

\[ S_A = 1 - 1/\Phi_A, \quad \Phi_A = \sigma_A^2 / \sigma^2, \quad \sigma_A^2 = \sigma_C^2 + b r \sigma_\alpha^2 \]

\[ S_C = 1 - 1/\Phi_C, \quad \Phi_C = \sigma_C^2 / \sigma^2, \quad \sigma_C^2 = r \sigma_\gamma^2 + \sigma^2 \]
Bayes Decision Rule

Let \[ z_{12,j} = d_{12,j} / \sigma \sqrt{2 / r} \], \[ d_{12,j} = \bar{y}_{1j} - \bar{y}_{2j} \]

\[ z_{12,.} = d_{12,.} / \sigma \sqrt{2 / br} \], \[ d_{12,.} = \bar{y}_{1.} - \bar{y}_{2.} \]
Bayes Critical z Value

Decide $\delta_{12,j} > 0$ if

$$z_{12,j} > \frac{z_k}{\sqrt{S_C}} \left\{ \frac{S_A}{bS_C} + \frac{b - 1}{b} \right\}^{1/2} - \frac{z_{12,\cdot}}{\sqrt{b}} \left\{ \frac{S_A}{S_C} - 1 \right\}$$

Linear dependence on standardized marginal treatment effect $z_{12,\cdot}$.

↑ with ↓ interaction (↑ $S_A / S_C$)

↓ with ↑ # subgroups $b$.  

29
Fully Bayes Critical $t$ Boundaries

$F_c = 3.5$

$F_c = 1.5$

$R_-$

$R_0$

$R_+$

$t_{12,\cdot}$

$t_{12,j}$
Concluding Remarks

• Bayesian subgroup analysis offers efficiency in estimation.

• Using a Bayesian hierarchical model,
  – a sample estimate of a subgroup treatment effect that is large and clinically impressive in magnitude could be shrunk to a much smaller, less compelling value.
  – Such dramatic shrinkage could suggest that the large point estimate may have been a random high due to multiple testing and large standard errors within subgroups.
Concluding Remarks

- Bayesian subgroup analysis may be most useful for clinical studies with small sample sizes (feasibility, exploratory, Phase II) or for rare event endpoints (safety).
- In clinical validation studies (pivotal, Phase III), regulatory authorities may still prefer the frequentist approach to formally control operating characteristics (studywise type 1 error).
- When the direction of subgroup effects can be anticipated (e.g., biomarker measurement of the target of a drug), a priori exchangeability of subgroups may not be warranted.
Concluding Remarks

• Tools are available for assessing Bayesian model fit
  – Exchangeability of subgroups can be assessed a posteriori by computing the Bayesian p value of a sensitive statistic based on its predictive distribution (Pennello, Thompson, 2008).

• Bayesian hierarchical models and false discovery rate:
  – For some loss functions, the directional FDR is controlled
  – The posterior probability of a null hypothesis is like a *local* false discovery rate
References

- DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Statist*. 1999; 53, 170-190.
- FDA. *Draft Guidance: Design Considerations for Pivotal Clinical Investigations for Medical Devices.*
- FDA. *Guidance for Use of Bayesian Statistics in Medical Device Clinical Trials.*
References

- Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons, *J Research on Educational Effectiveness* 2012; 5: 189-211.
References

• Pennello, GA. The k-ratio multiple comparison Bayes rule for the balanced two-way design, *J Amer Statist Assoc* 1997; 93: 675-684.