AMGEN®

Pioneering science delivers vital medicines™

Early Safety Signal Detection

Brian Smith[1], Haijun Ma[1]

Jeffrey Zhang[1], Amy Xia[1], Wenhua Hu[2]

1Amgen, Inc. 2Bristol-Myers Squibb

BASS 2013

Philosophy of Safety Signal Detection in Early Drug Development

TIME TO CLARIFY

Typical FIH Study

- 7 doses of drug and placebo
- Vital signs (SBP, DBP, HR, RR, Temp) measured at 12 time points
- ECG measured at 15 time points
- Laboratory measurments 2 time points

Multiplicity

- **If you did all pairwise comparisons to placebo for all endpoints at all times, there could be ~1500 p-values generated**
- **Using a p-value of 0.05 we would expect around 75 false positives if the compound completely safe**
- **BTW this is not counting all the potential analyses that could come from AEs**

Why so concerned about false positives?

- **For a safety endpoint, a signal could lead to the end of development for the molecule**
- **Now, if that signal is false…**
- **The traditional p-value < 0.05 along with 1500 tests will not work**
  - **Contingent on the notion that p-value < 0.05 → proves effect**

A Common Refrain

- **P-values should not be generated for safety data**
- **Use confidence intervals**
- **Judge the signal through clinical relevance**

The Fallacy of The Refrain

- **No matter whether we use p-values or clinical judgment to interpret a signal**
- **False Positives Will Occur**
- **The problem is that with clinical judgment we are not in a position to control the rate of false positives**

New Refrain

- **We will use a statistical system that does not use p-values**
- **Let's be Bayesian**
- **Problem – As long as there is some sort of decision rule, false positives will occur**
- **Seems like we are stuck in quick sand**

A Hint from RA Fisher

- **If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance. (RA Fisher)**

The Four Pillars of Early Safety Evaluation

1. **Repetition**

2. **Bayesian Thought**
   - **Not necessarily Bayesian statistics**
   - **I know Fisher would have hated this bullet!!**

3. **Clinical Judgment**

4. **Potential Subject Risk**

Summary of Philosophy

- **For an early safety signal detector to be used we must broaden our and our colleagues understanding of**
  - **Probability**
  - **Multiplicity**
  - **P-values**
  - **False Positives**
  - **Bayesian Thinking**

Further Motivation of an Early Development Signal Detector

To Kill or not to Kill

- **…the kill decision, especially an early kill decision, allows the sponsor to reallocate people and money to other development programs promising more benefit. – Dan Weiner, Pharsight**

- **Early Safety Signal Detect can help facilitate a kill decision**

Awareness makes us wiser

- **Imagine a less serious AE (Headaches)**
- **Reduction in Dose**
- **Early Signal Detector would help facilitate**

Ideal Properties of an Early Detector

- **Accommodates Past Information**
- **Could update**
- **Relatively automated**
- **Has reasonable power**
- **Signal spotter not signal prover**

Component 1 - Use Past Data

- **We collect placebo data in a lot of trials (healthy subject)**
- **We should be able to know what the average ALT is for healthy subjects**
- **Let’s use this information**

Component 2 - Take advantage of continuous data

- **Categorical – Is ALT > 3xULN**

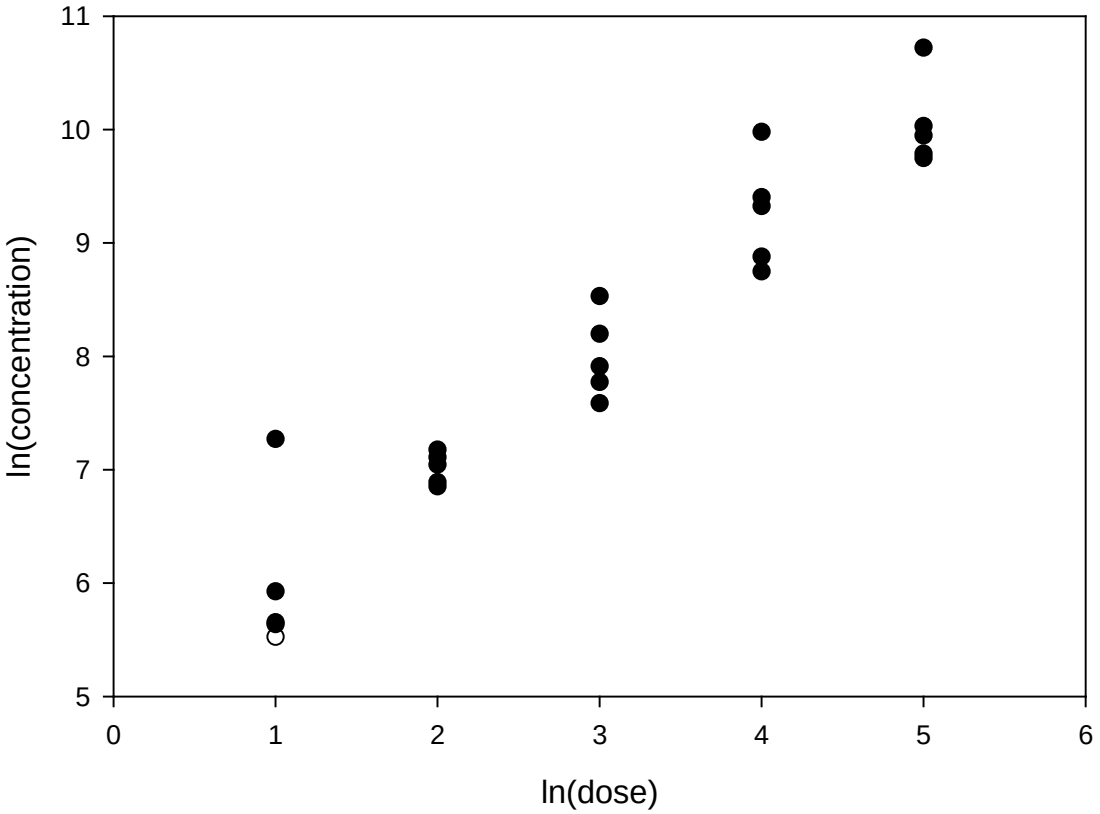
- **Understand Distribution**
  - **Mean ALT for Treatment**
  - **Mean ALT for Placebo**
  - **Variance**

- **Predict**
  - **P(ALT > 3xULN) for Treatment**
  - **P(ALT > 3xULN) for Placebo**
  - **Relative Risk**

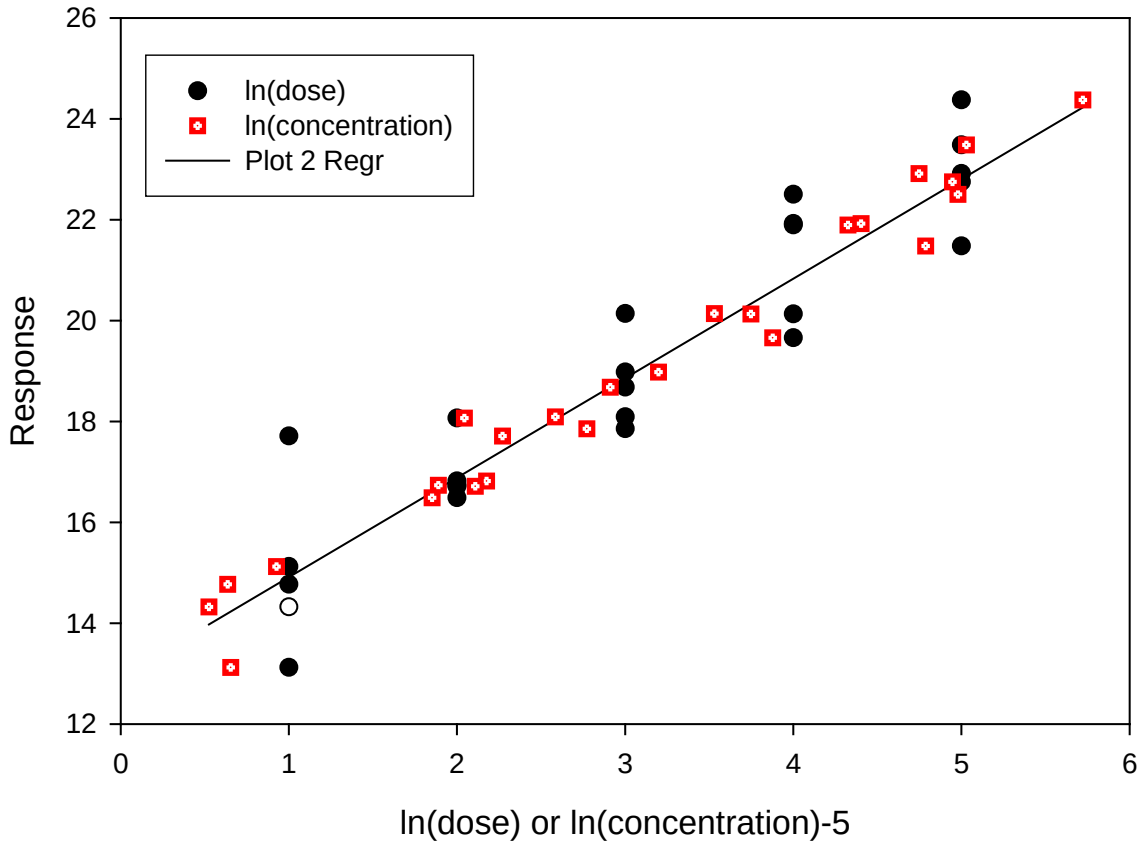
Component 3 - Take advantage of covariates

- **ANCOVA more powerful than ANOVA**

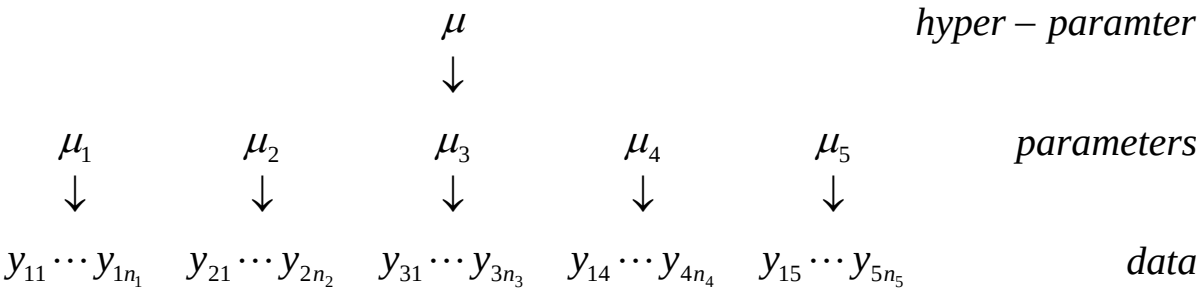
Component 4 - Use concentration/response instead of dose/response



Use concentration/response instead of dose/response



Component 5 - Hierarchical Bayesian Models



- Each study has estimated means $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, and\ \bar{x}_5$

- Suppose means all very similar
  - Mean of the 5th study can be based on all of the data
  - Precision of the estimate will be based on all of the data

- Suppose mean of 5th study very different
  - Mean of 5th study should be based on data from the 5th study
  - Precision of the estimate will be based on data from the 5th study

AMGEN®

Pioneering science delivers vital medicines™

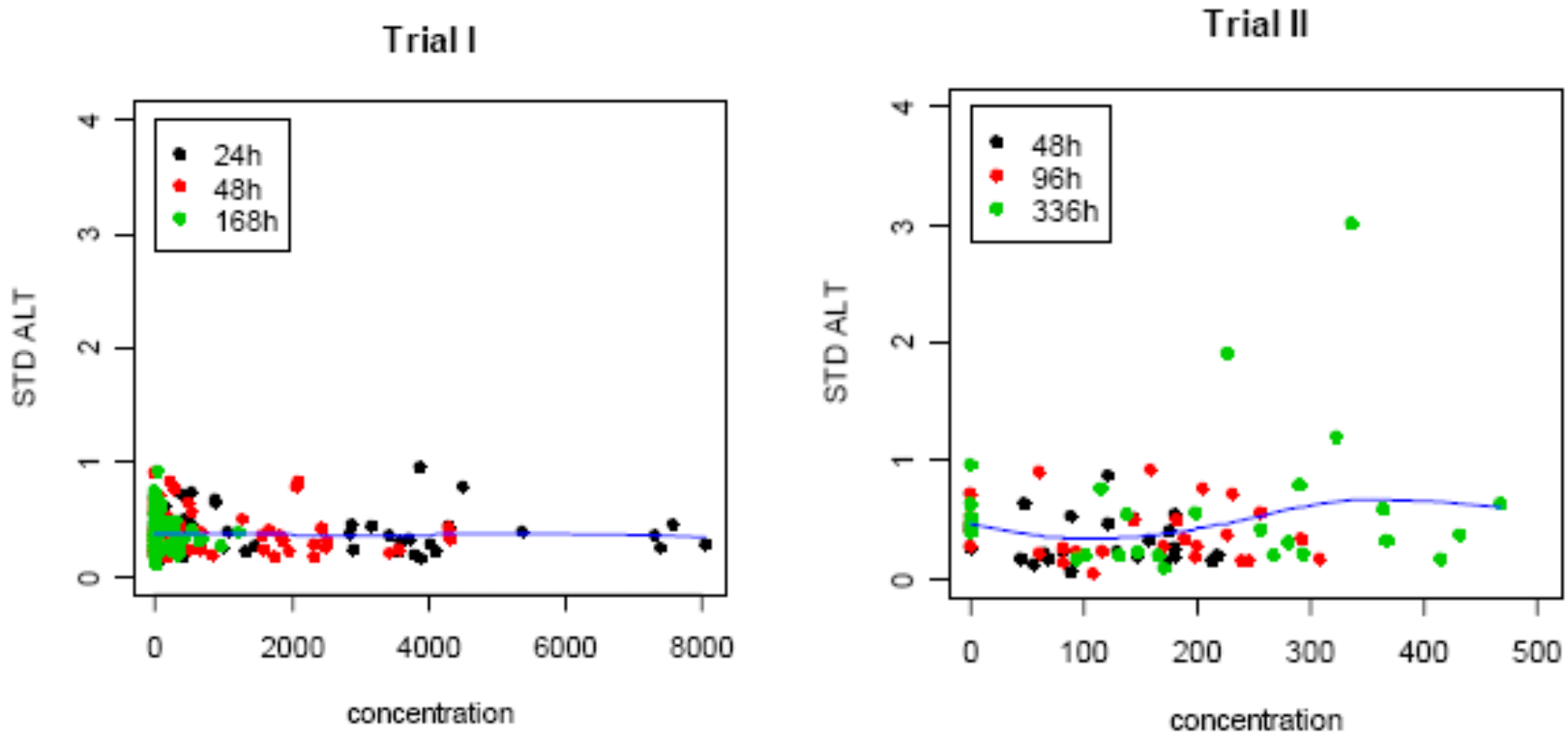
First Attempt

An Example

Example of Signal Detection

- Single dose first in human study (trial I)
  - 58 dosed, 26 placebo
- Followed by multiple dose study (trial II)
  - 22 dosed, 6 placebo
- Both trials in healthy subjects.
- Some high ALTs seen in second trial.

Data in two trials



$$STD\ ALT = \frac{ALT}{ULN}$$

Challenges

- Sample size on current trials may be too small for inference
  - Not enough precision to quantify the likelihood of observing the abnormal observations
- Interested in finding a concentration-response relationship for signal screening
- Individual lab observations often affected by baseline conditions and often have strong within-subject correlation

Use of Historical Placebo Data

Historical data

- 25 phase I trials on healthy subjects in placebo group between 1997-2007 in Amgen
- N = 309
- Common covariates: demographical variables

Use of Historical Data

- Determine the underlying distribution for the response
- Find important covariates in order to reduce variance
- Help to interpret results of the current study

What factors affect the response ALT?

Effect	Estimate	Standard Error	Pr > |t|
Intercept	1.99	0.14	<.0001
Female (ref: Male)	-0.36	0.06	<.0001
AGE	0.0052	0.0018	0.0055
B_BMI	0.036	0.006	<.0001

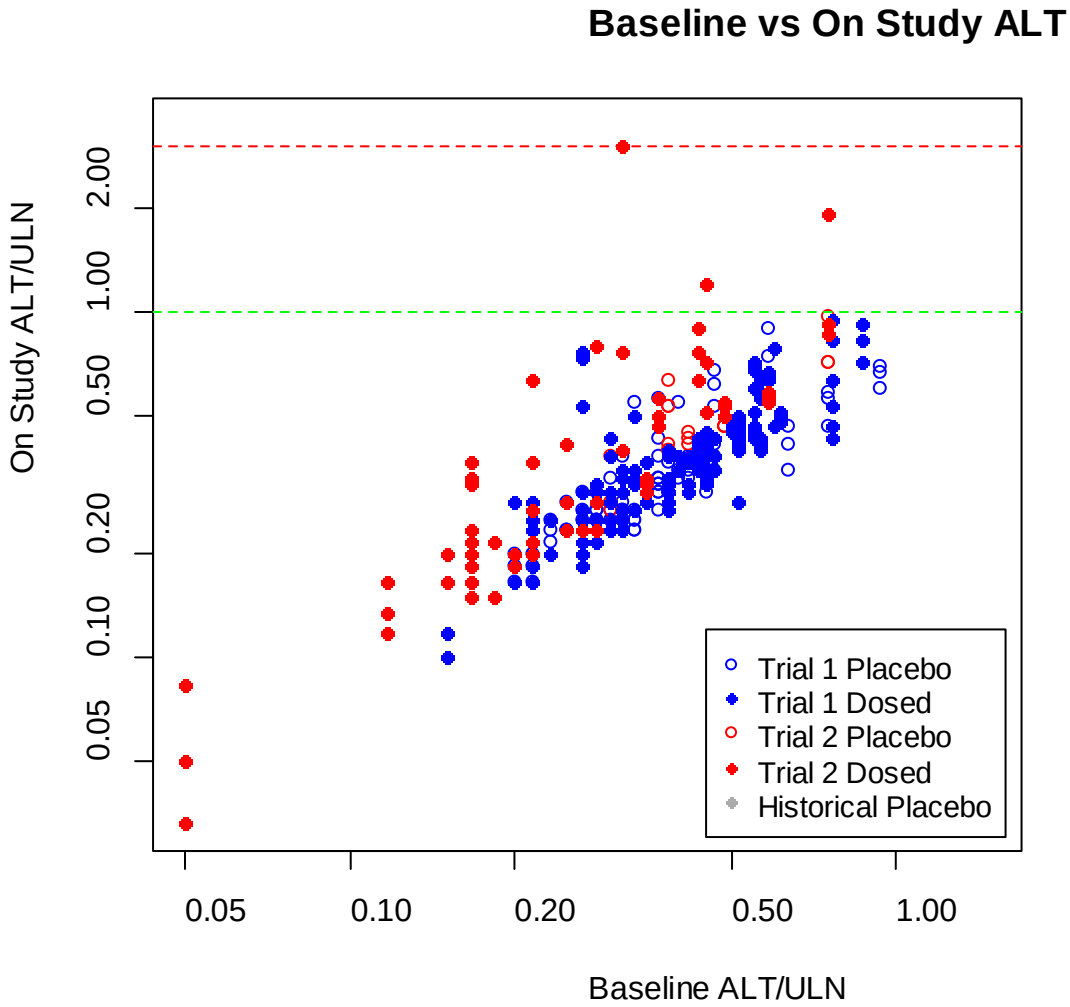
- Investigated: gender, race, BMI, age, visit time, weight and height
- Sex, Age, Baseline BMI are significantly related to ALT
- Consistent to the previous work (Eran Elinav et al, 2005)

What factors affect the response ALT?

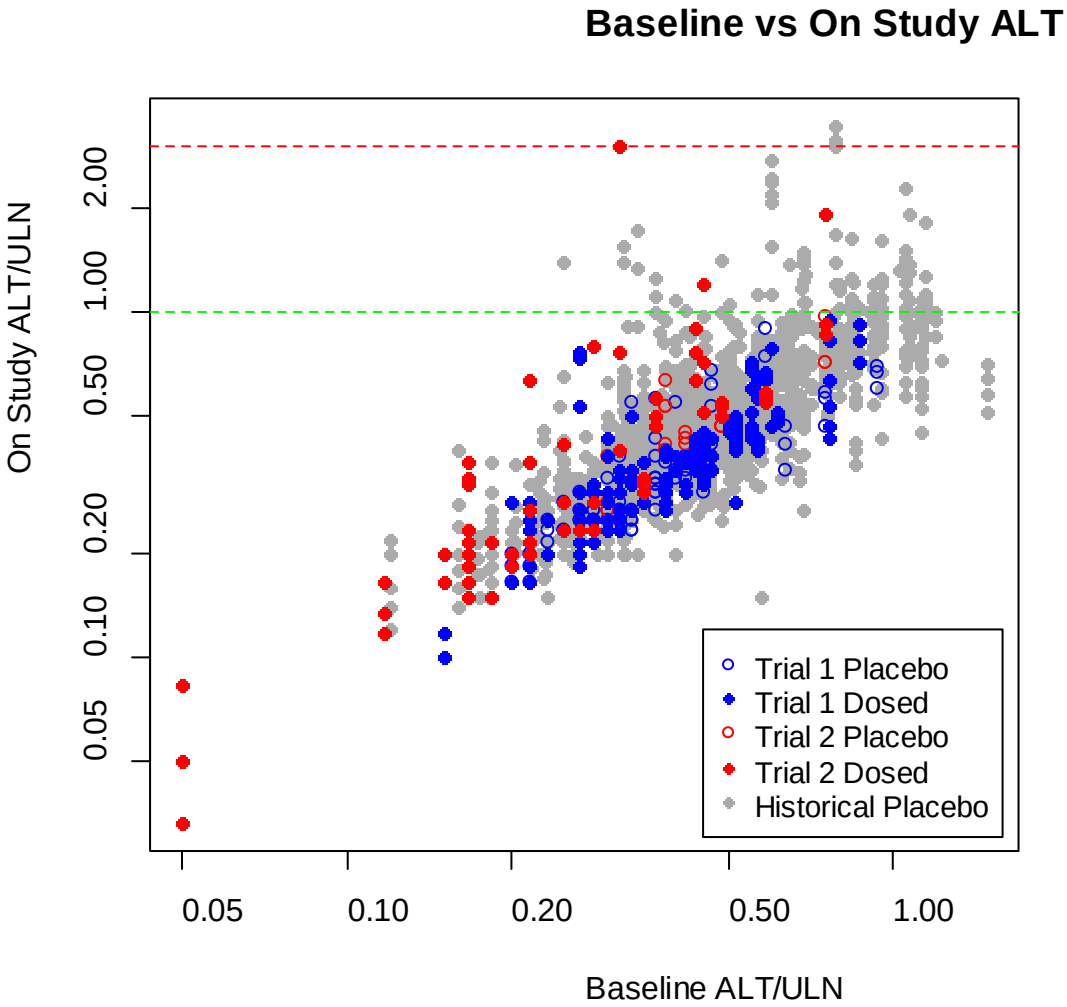
Effect	Estimate	Standard Error	Pr > |t|
Intercept	0.35	0.09	0.0002
Baseline log(ALT)	0.87	0.03	<.0001
Female (ref: Male)	-0.019	0.035	0.60
AGE	-0.0009	0.0010	0.36
Baseline BMI	0.004	0.003	0.20

- Baseline ALT breaks down significance of Sex, Age, Baseline BMI

Baseline is the most Important Covariate for ALT Response



Baseline is the most Important Covariate for ALT Response

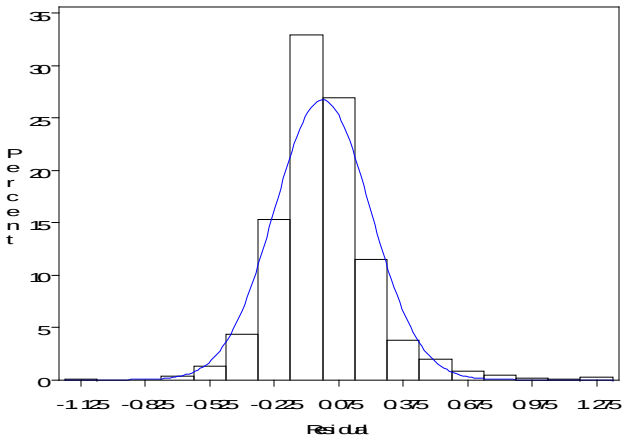


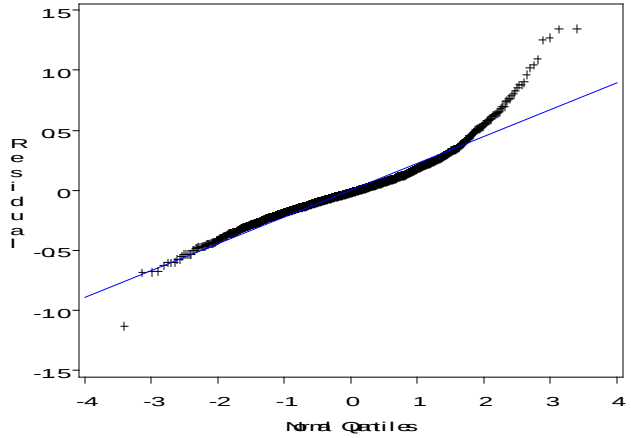
What assumption should we make for residuals?

$$Y_{kij} = \beta_{0k} + \beta_1 * Y_{ki0} + b_{ki} + \varepsilon_{kij}$$

Y---log(ALT)

k: study, i: subject, j: observation





Skewness: 0.97

Kolmogorov-Smirnov Test for normality: p<0.01

Model We Want:

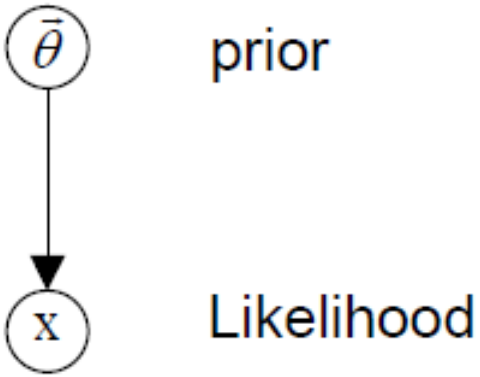
- **Borrow information from historical placebo data**
- **Borrow information between the two trials when appropriate**
- **Interested in identifying a dose–response relationship**
- **Model can handle tail behavior appropriately**

Incorporation of Historical Data in Bayesian Hierarchical Modeling of Extreme Lab Values

Bayesian Hierarchical Modeling

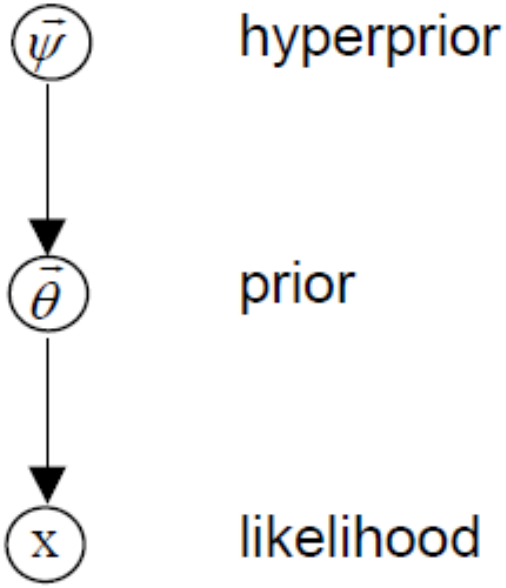
- **Bayes Theorem**

$$p(\vec{\theta} \mid x) \propto p(x \mid \vec{\theta})\pi(\vec{\theta})$$



Bayesian Hierarchical Modeling

$$p(\vec{\theta},\vec{\psi} \mid x) \propto p(x \mid \vec{\theta})\pi(\vec{\theta} \mid \vec{\psi})\pi(\vec{\psi})$$



30

Objectives

- Bayesian hierarchical models that
  - Model the mean trend
    - Removing baseline effect
    - Using repeated laboratory measurements
    - Incorporating information from historical data
  - Model residuals with more robust distributions that allow for heavy tails
- The model could be used for signal screening and event prediction for future studies
- The model can be updated when new data become available

Our proposal: Mean Trend for log(ALT)

- Y_{i0}: baseline ALT, Y_{kij}: ALT from study k, subject i observation j
- $C_{26,ij}$, $C_{27,ij}$: concentration in two trials (26 is the trial with abnormal ALT elevation)

$$\log(y_{kij}) = b_1 \log(y_{i0}) + b_2 C_{26,ij} + b_3 C_{27,ij} + \lambda_k + U_{ki} + \varepsilon_{kij}$$

Parameter	Interpretation	Prior Distribution
b_1	Coefficient of baseline ALT	$N(0, 10^{10})$
b_2	effect of concentration in trial 1	$N(0, 10^{10})$
b_3	effect of concentration in trial 2	$N(0, 10^{10})$
λ_k	Study level random effects	$N(\alpha_0, \sigma_0^2)$
U_{ki}	Subject level random effects	$N(0, \sigma_1^2)$

Non-informative hyper priors for α_0, σ_0^2 and σ_1^2

Robust Inference with Student-t Assumption

- Replacing the normality assumption of measurement error with the t-distribution provides a robust method for outliers (Sutradhar and Ali 1986; Lange, Little, and Taylor 1989).
- For Bayesian hierarchical model, one more step could be added in order to use t-distribution with d.f. v.

$$y_i | V_i \sim N(\mu, V_i)$$

$$V_i \sim Inv - \chi^2(\nu, \sigma^2)$$

Extreme Value Modeling

- Extreme value modeling seeks to analyze observed extremes and forecast the occurrence and magnitude of further extremes
- The generalized Pareto distribution (GPD) is often used to model the tails of another distribution
  - Commonly used in environmental, financial and engineering data analysis
  - Southworth and Heffernana (2012) applied GPD for safety laboratory data analysis

Our proposal: Mixture Model for Residuals

- Model residuals as a mixture of truncated t-distribution and Generalized Pareto Distribution

$$F(\varepsilon_{kij}) = 0.5 * \tilde{F}_{<0}(\varepsilon_{kij}) + 0.5 F_{>=0}(\varepsilon_{kij}, \phi(c)_{kij}, \xi(c)_{kij})$$

$$F_{>u}(x) = 1 - \left\{1 + \xi\left(\frac{x-u}{\sigma}\right)\right\}^{-1/\xi} \quad \text{for } x > u,$$

- σ is scale parameter
- ξ is shape parameter
- μ is fitting threshold
- $\mu < x \leq \mu - \sigma/\xi \quad if \quad \xi < 0$
- $\mu < x \leq \infty \quad if \quad \xi \geq 0$

Priors: High uncertainty in estimation of parameters of ϕ, ζ due to small sample size.

Priors help to share occurrence of extreme value of clinical variables among studies

- $\phi_k \sim normal(\phi_0, \sigma_\phi^2)$
- $\xi_k \sim normal(\xi_0, \sigma_\xi^2)$

Results: DIC

	Model of Residuals	Dbar + pD=DIC
M1	Normal distribution	-202 + 279 = 77
M2	T distribution	-660 + 293 = -367
M3	T+GPD Mixture distribution	-840 + 316 = -524

- Deviance information criterion (DIC) is used for model selection
- Smaller DIC indicates better model fitting

Results: Parameter Estimates

parameter	Interpretation	Normal	T	Mixture
b_1	Coefficient of baseline ALT	0.93 (0.02)	0.92 (0.02)	0.90 (0.01)
b_2	effect of concentration in first trial	3.623E-6 (1.27E-5)	2.751E-6 (9.035E-6)	2.357E-6 (8.681E-6)
b_3	effect of concentration in second trial	0.002 (0.0003)	8.978E-4 (2.801E-4)	6.37E-4 (2.554E-4)
σ_0^2	Variability of study random effects	0.003 (0.002)	0.002 (0.001)	0.002 (0.001)
σ_1^2	Variability of subject random effects	0.024 (0.003)	0.018 (0.002)	0.017 (0.002)
σ_z^2	Variability of within subjects obs.	0.053 (0.002)	0.019 (0.001)	0.020 (0.001)
kappa	Df for t-distribution		2.66 (0.21)	4.9(0.26)

Posterior Predictive Probability (%) of ALT>3ULN (Trial II)

Percentiles (BL STD ALT)	50% (0.40)			95% (0.84)			100% (1.49)		
Percentiles (Concentration)	Nml	T	T+GPD	Nml	T	T+GPD	Nml	T	T+GPD
0 (0)	0	0.08	0	0	0.2	0.3	0.3	0.8	1.8
25 (59)	0	0.04	0.5	0	0.2	2.0	0.5	0.9	6.9
50 (140)	0	0.04	0.8	0	0.2	2.8	1.6	1.2	8.6
75 (214)	0	0.08	1.0	0	0.3	3.2	4.1	1.8	9.4
90 (292)	0	0.08	1.1	0.1	0.4	3.6	9.7	2.6	10.2
95 (359)	0	0.1	1.2	0.3	0.4	3.8	17.4	3.5	10.5
100 (466)	0	0.07	1.4	1.5	0.7	4.1	35.8	7.2	11.2

Relative predictive probability at the highest concentration vs placebo for subjects with baseline ALT of 1.49 over upper normal limit was 119, 9 and 6 times respectively using normal, t and mixture distributions.

Nml: normal distribution; T: t-distribution ; T+GPD: mixture distribution

AMGEN

Results: Benefit of Using Historical Data

Parameter	Interpretation	T	T (no historical data)
b_1	Coefficient of baseline ALT	0.92 (0.02)	1.00 (0.04)
b_2	effect of concentration in trial I	2.751E-6 (9.0E-6)	3.052E-6 (7.4E-6)
b_3	effect of concentration in trial II	0.0001 (2.8E-4)	0.001 (4.2E-4)

- Precision increase for baseline ALT coefficient estimate
- Point estimate shift and precision increase for trial II concentration effect estimate

Summary of First Example

- Example used to show a feasible method for early safety signal detection
- We explored a concentration-response relationship
- Using historical data improved precision of population level parameter estimates and provided better prediction
- We made use of existing Amgen healthy subject placebo data
  - determine the underlying distribution for the response
  - find important covariates in order to reduce variance
  - help in interpreting results of a current study with more precision

Discussion

- **Bayesian hierarchical modeling is convenient to incorporate historical data and provide prediction**
- **Extreme value modeling focuses on tail behavior and could be used for abnormal laboratory modeling and prediction**
- **Plans include**
  - **Extending the model to other endpoints**
  - **Simulation study to compare the performance of different models**

Futuristic Thought

- **Could we establish industry wide placebo database for such efforts?**

BACKUP


Q&A