# Overview of Model Evaluation Methodologies

Marc R. Gastonguay, Ph.D.

metrum Research Group LLC

15 Ensign Drive, Avon, CT 06001
marc.gastonguay@snet.net

# Overview

- Validation?
- A risk-based approach
- Assumption checking
- Evaluation methods
  - Parameters
  - Predictive performance
- Sensitivity analysis

# Can we agree on a name?

- Model Appropriateness
- Model Checking
- Model Evaluation
- Model Qualification
- Model Validation
- Model Verification

# "Validation" Is Misleading

"All models are wrong but some are useful."

[Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer, and G. N. Wilkinson, (eds.) Robustness in Statistics. New York: Academic Press]

- Models are never completely valid, but the application of a model to a specific purpose(s) can be evaluated.

# A Risk-Based Approach to Model Evaluation

- Identify:
  - What are the deficiencies of the model?
  - What are the resulting risks for the modeling and decision-making process?
- Assess:
  - Will these deficiencies/risks impact the intended use of the model?
- Manage:
  - What are the strategies for managing these risks?

# What to Evaluate?

- The model itself?
  - Structural PK-PD model
  - Models for covariate-parameter relationships
  - Random effect models
- The performance of model-based applications and inferences?
  - Parameter estimates and confidence intervals
  - Hypothesis tests
  - Predictions/simulations with model

# How?

- Before implementing a model evaluation method:
  - Look at the data carefully
  - Review modeling objective
  - Develop model according to GOF criteria
  - Check model assumptions
  - Define model evaluation method and criteria for decision making

# Methods

- Assumption Checking
  - Randomization test
- Stability/precision of parameter estimates
  - Validation through parameter prediction errors (Bruno et al. *JPB* 24:153-172, 1996)
  - Log-Likelihood Profile
  - Bootstrap (parametric & non-parametric)
  - Cross-Validation/Leverage analysis
- Assessment of model performance
  - Prediction errors (from internal or external test set)
  - (Posterior) Predictive Check
- Sensitivity Analysis
  - Part of the simulation effort

# Assumption Checking

# Assumption Checking

- Test assumptions of structural model
- Test assumptions in statistical model
- Test estimation method assumptions
- Primarily a model-building and data analysis concern (especially when analysis involves hypothesis testing)

# Assumption Checking: Model

- Does the model fit the observed data?
- Has the search reached a global minimum?
- Are parameter estimates consistent with prior knowledge?
- Are random effect distributions consistent with modeling assumptions (e.g. Normal $\eta$ & $\varepsilon$ distributions, centered on zero)?
- What is the impact of assumptions in data recording/assembly? (try plausible scenarios)

# Assumption Checking: Estimation Method

- Are assumptions about likelihood approximation valid? (Try more rigorous methods and compare diagnostics).
- Are assumptions about test statistics accurate? (e.g. $\chi^2$ distribution of $\Delta$-2LogL)

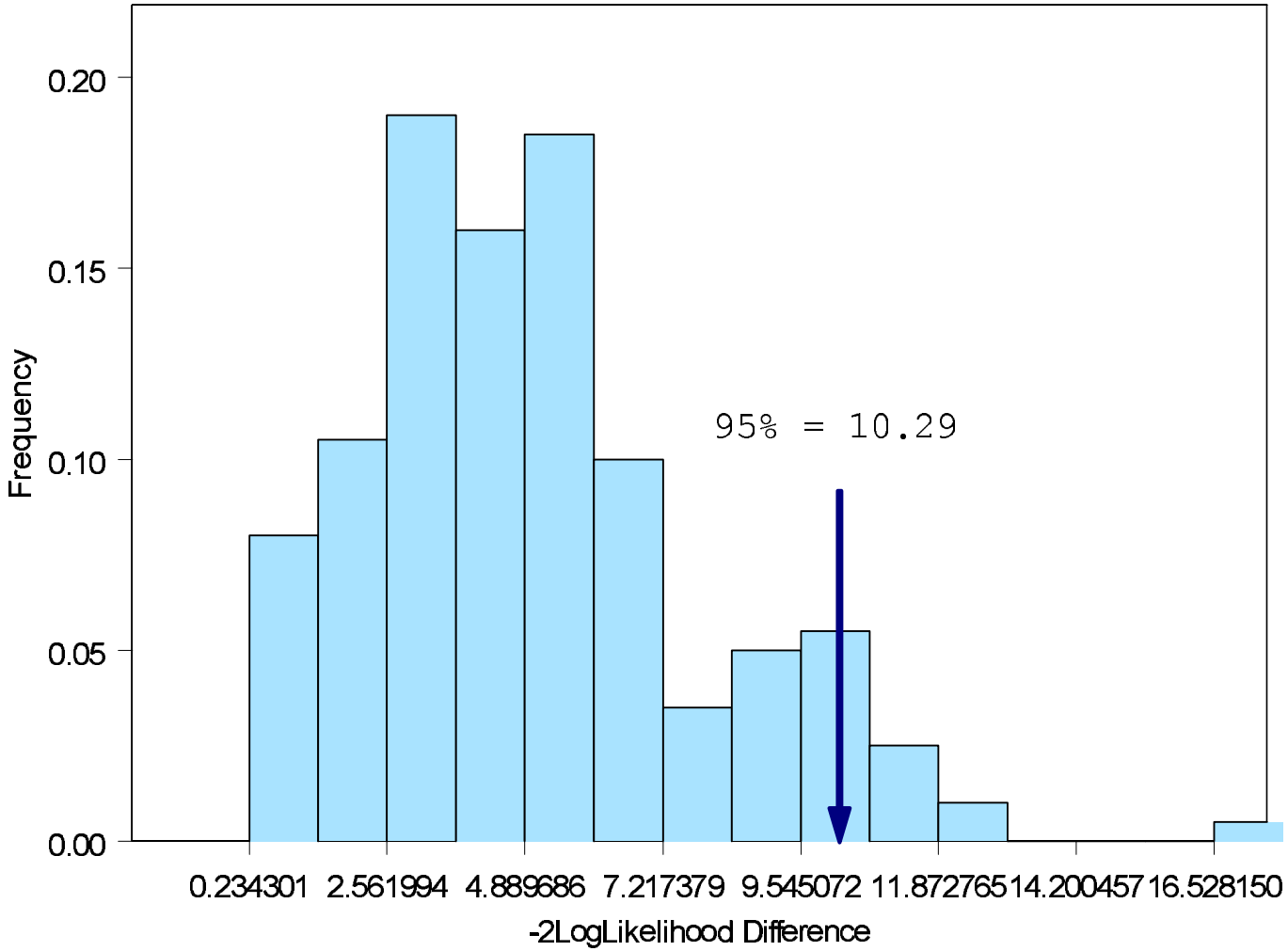
# Randomization Test

- Checks assumptions about $\chi^2$ distribution of $\Delta$-2LogL
- Determines actual significance level for a given model comparison (Full vs. Reduced)
- Distribution of $\Delta$-2LogL for null hypothesis is generated by fitting repeated random permutations of data set

# Randomization Test: Method

1. Fit Reduced ($H_0$) and Full ($H_1$) models to original data set and obtain Δ-2LogL
2. Randomly permute (scramble) variable of interest (e.g. covariate) across entire data set
3. Fit Full ($H_1$) model to permuted data set and compare to Reduced model to obtain Δ-2LogL
4. Repeat 2 – 3 several times (hundreds)
5. Generate distribution of Δ-2LogL and pick quantile of interest (95% for $p<0.05$)

# Distribution of Δ–2LogL for Null



# Model Evaluation: Test Data

- Internal
  - Data splitting (split data into model building & testing data sets)
  - Cross Validation (multiple splits)
  - Bootstrap (resampling or simulation)
- External (separate data set)

# Evaluation of Parameter Estimates

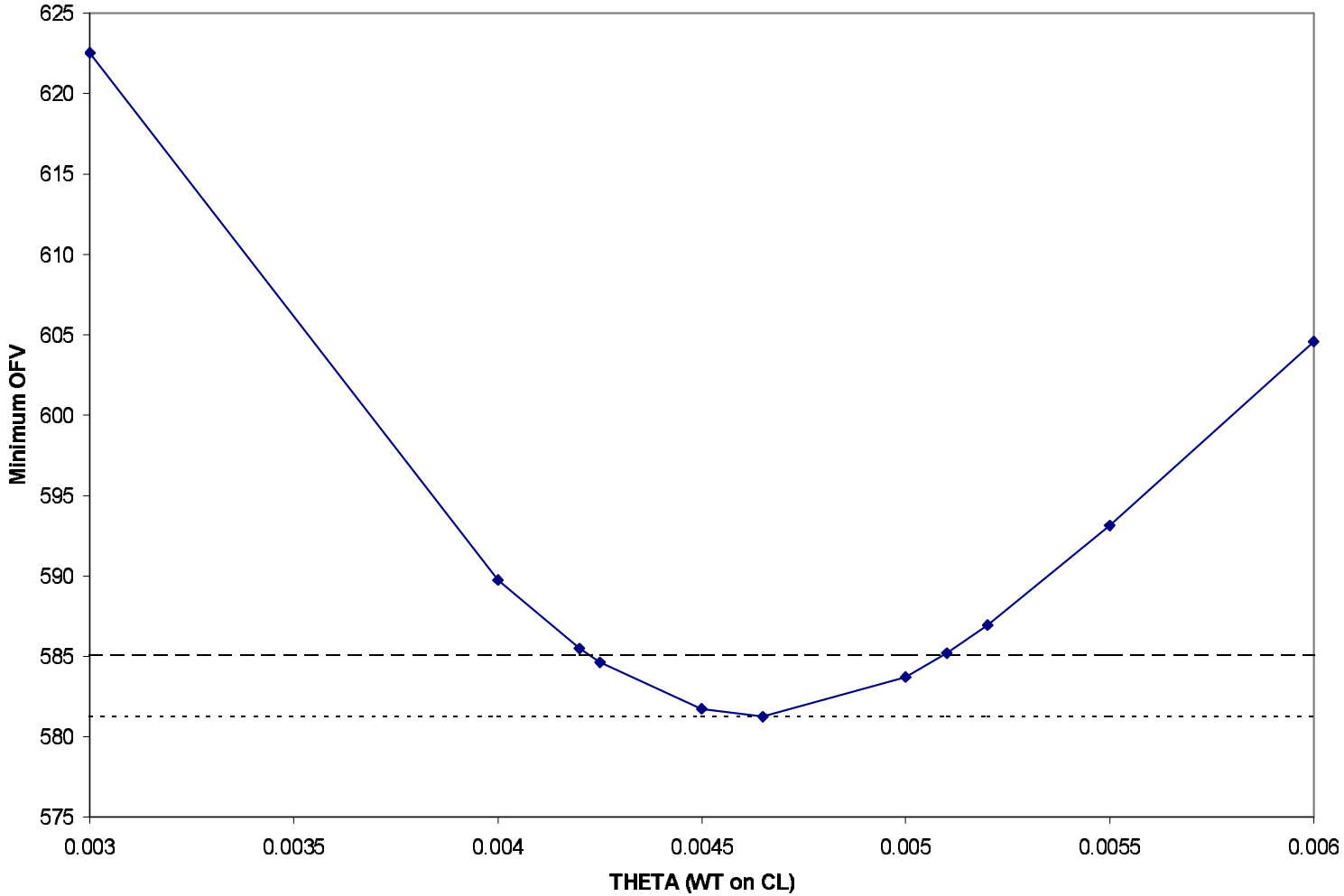
# Log-Likelihood Profile

1. Obtain final model and parameter estimates
2. Fix parameter of interest (e.g. THETA(1)) at range of values above and below the maximum likelihood estimate
3. Perform estimation runs for each of the fixed values of THETA(1)
4. Record minimum objective function value (MOFV) at each fixed value of THETA(1
5. Plot MOFV vs. THETA(1)

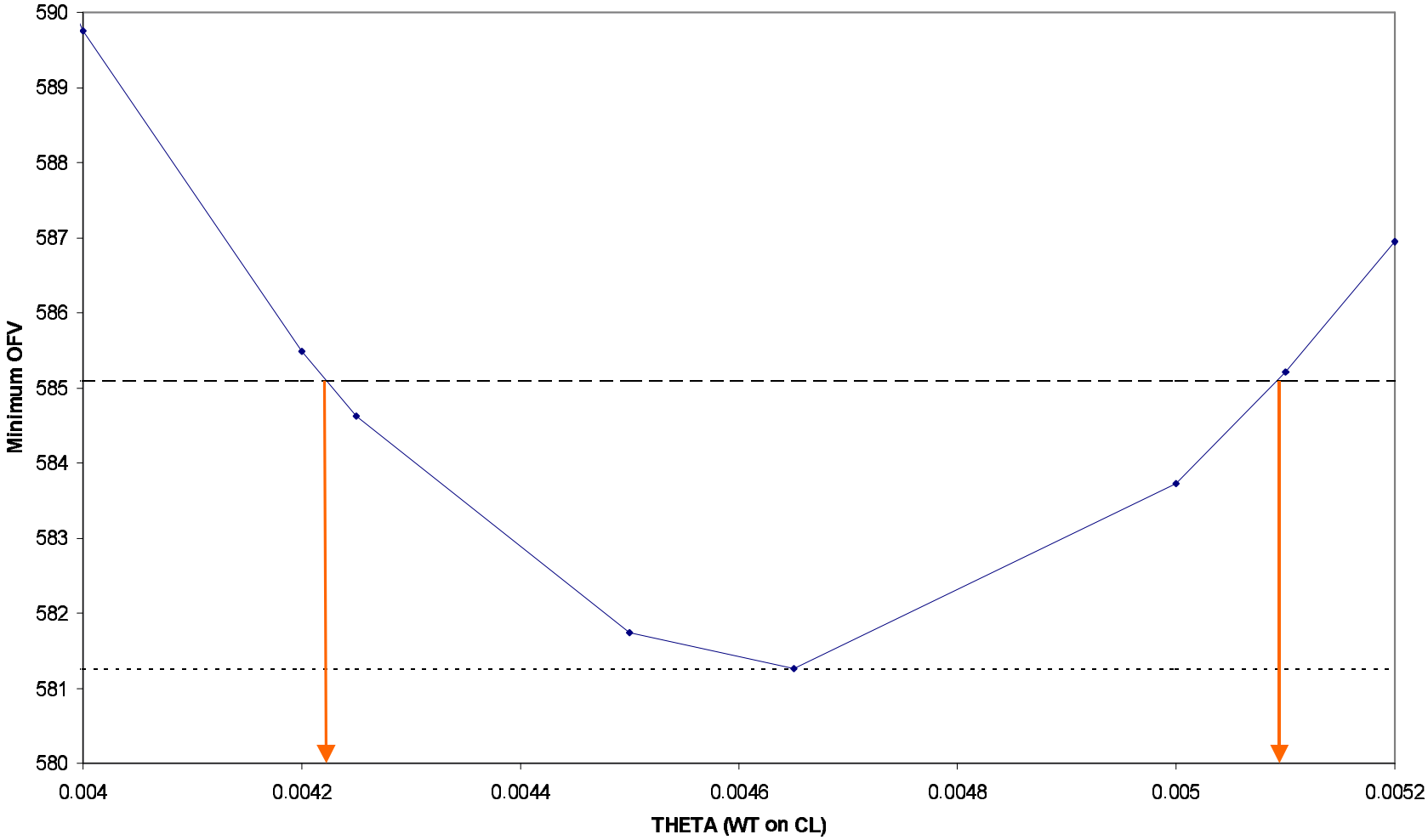
# Log-Likelihood Profile

- Values of THETA(1) that increase the objective function by 3.84 units are defined as 95% confidence interval
- Accuracy of this method is highly dependent upon accuracy of likelihood approximation in estimation method

# Log-Likelihood Profile



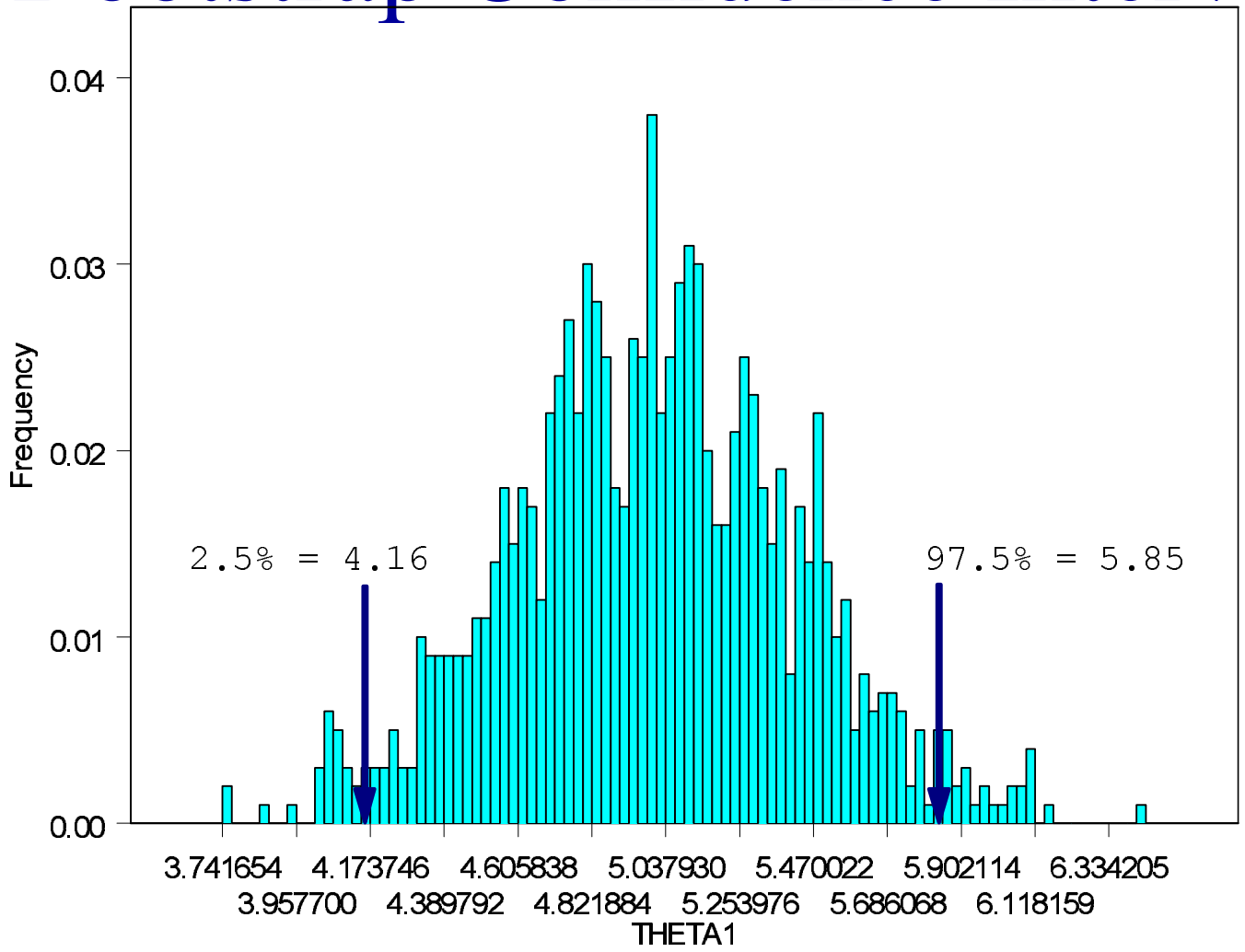
# Log-Likelihood Profile



# Bootstrap

1. Obtain final model and parameter estimates
2. Create several (hundreds) of replicate data sets by re-sampling (unit=individual) with replacement from original data set
3. Perform estimation run with final model on each of the replicate data sets
4. From distribution of population parameter estimates, obtain quantiles of interest
5. 95% CI is defined by $2.5^{th}$ and $97.5^{th}$ quantiles of parameter estimates

# Bootstrap Confidence Intervals



# Bootstrap

- Re-sampled data technique is know as non-parametric bootstrap. Results are conditional on the data.
- Parametric bootstrap can be conducted by simulating hundreds of data sets from final model (instead of re-sampling data). Results are conditional on model.
- Accuracy of both methods depends upon accuracy of likelihood approximation in estimation method

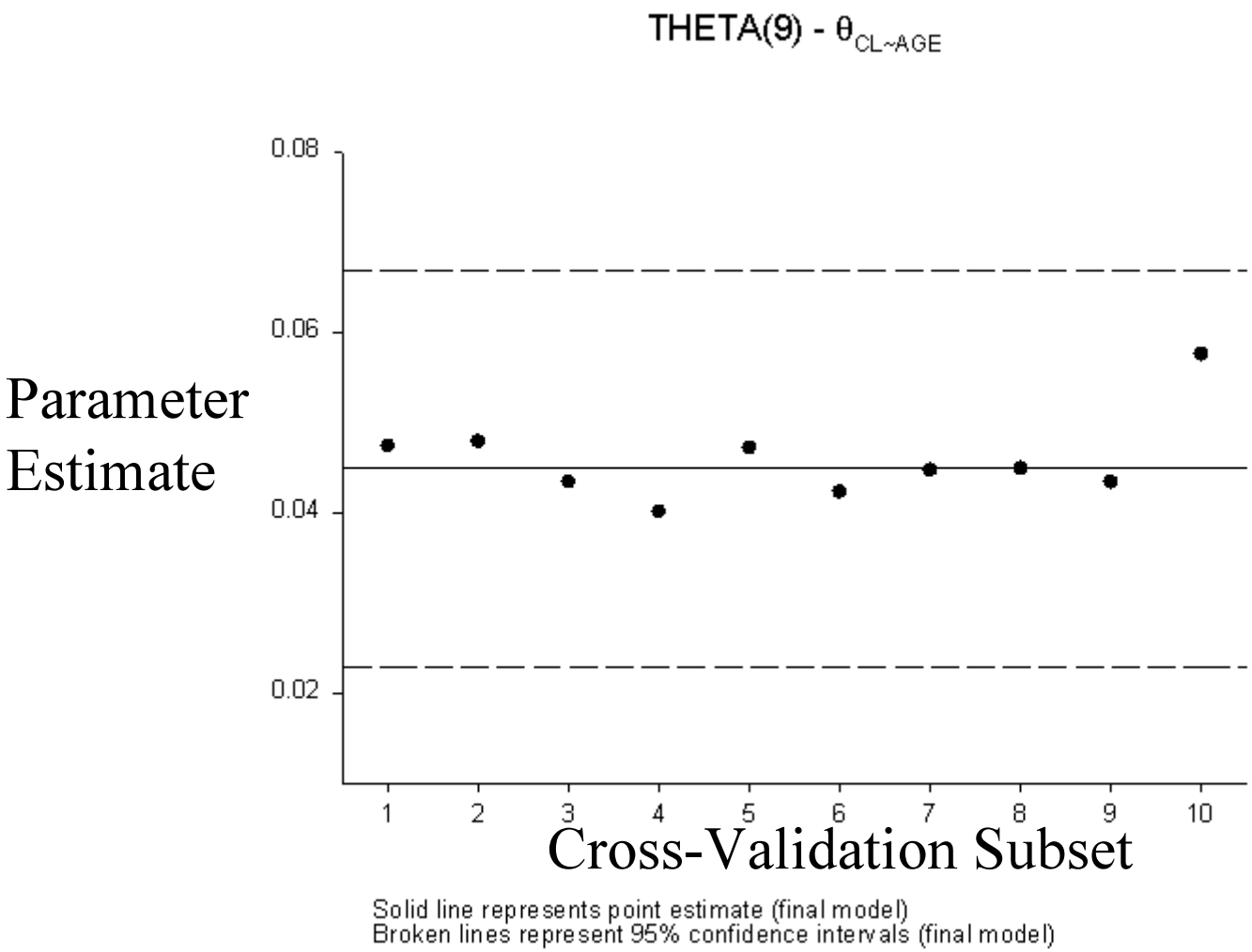
# Leverage Analysis

1. Obtain final model and parameter estimates with entire data set
2. Split data set into *m* approximately equal size (by individual) subsets
3. Fit the model to data from *m*-1 subsets
4. Record population parameter estimates
5. Repeat for each of the unique subsets

# Leverage Analysis

- Useful to explore stability of parameter estimates
- Assists in identifying highly influential individual(s) or outliers

# Leverage Analysis



# Evaluation Based on Predictive Performance

# A Simple Qualitative Evaluation

1. Predict into validation data set with model & parameters estimated from index data
2. Create diagnostic plots
   - Typical diagnostic scatter plots (PRED vs. DV, RES & WRES vs. TIME & PRED)
   - RES vs. covariates
   - PRED, DV vs. covariates

***SL Beal.*** *"Validation of a Population Model." E-mail to NONMEM Users Network, February 1, 1994.*

# Validation Through Predictions

- Prediction error: $PE_i = Cpred_i - Cobs_i$
- Summary metrics
  - Bias: MPE, MSPE
  - Precision: MSE, MAE, RMSE
- Statistical Issues
  - heteroscedastic variance

    $SPE_i = PE_i/SD_{i,}$ or use log(PE)
  - more than 1 observation per individual leads to correlated prediction errors and invalid statistical tests

# Single-Split or External Prediction

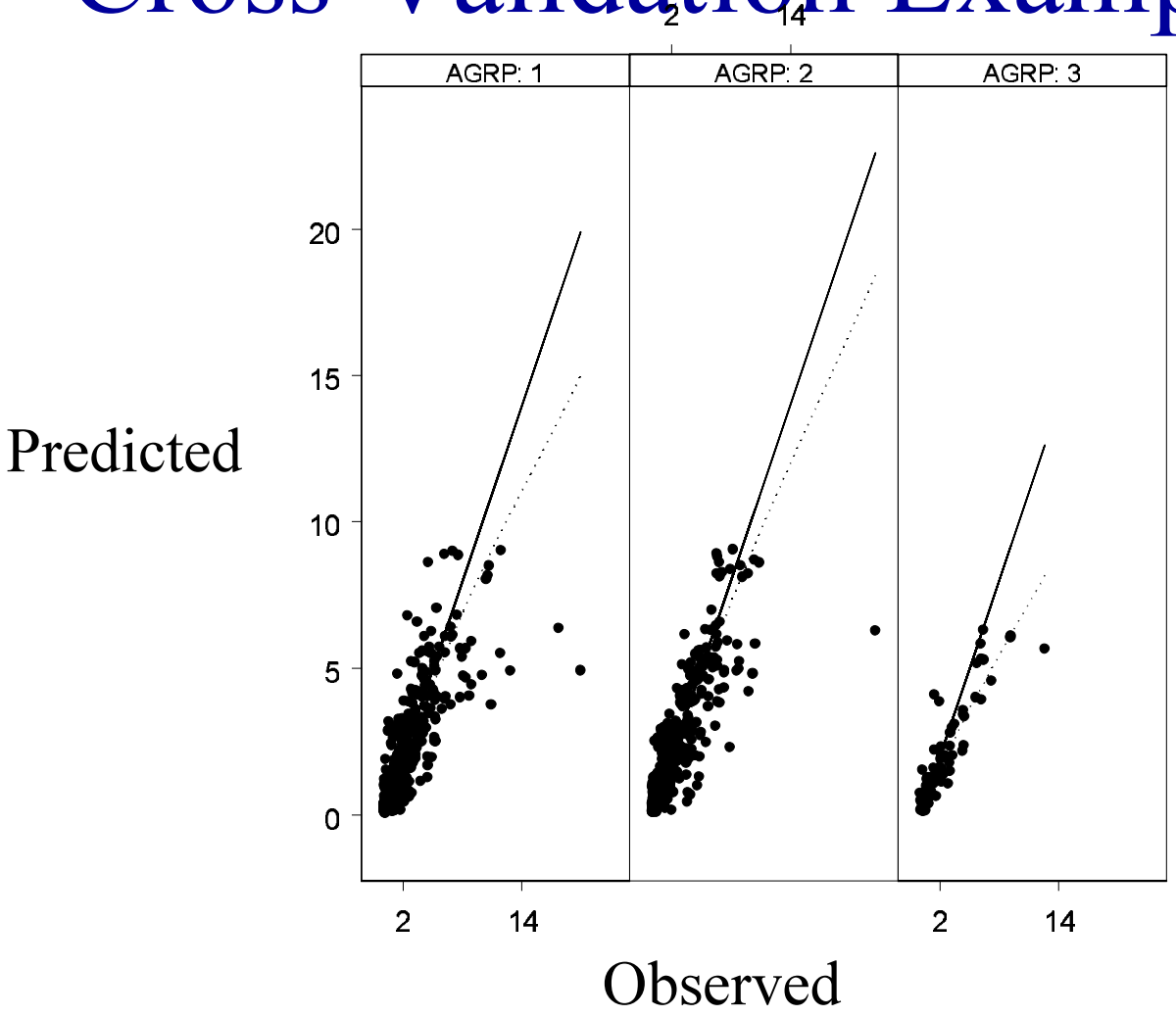
1. Split data once by individuals (e.g. 70/30)
2. Estimate with index set
3. Predict into validation set
4. Calculate prediction errors
5. Create diagnostic plots

OR- estimate with one data set and predict into an entirely new data set (external)

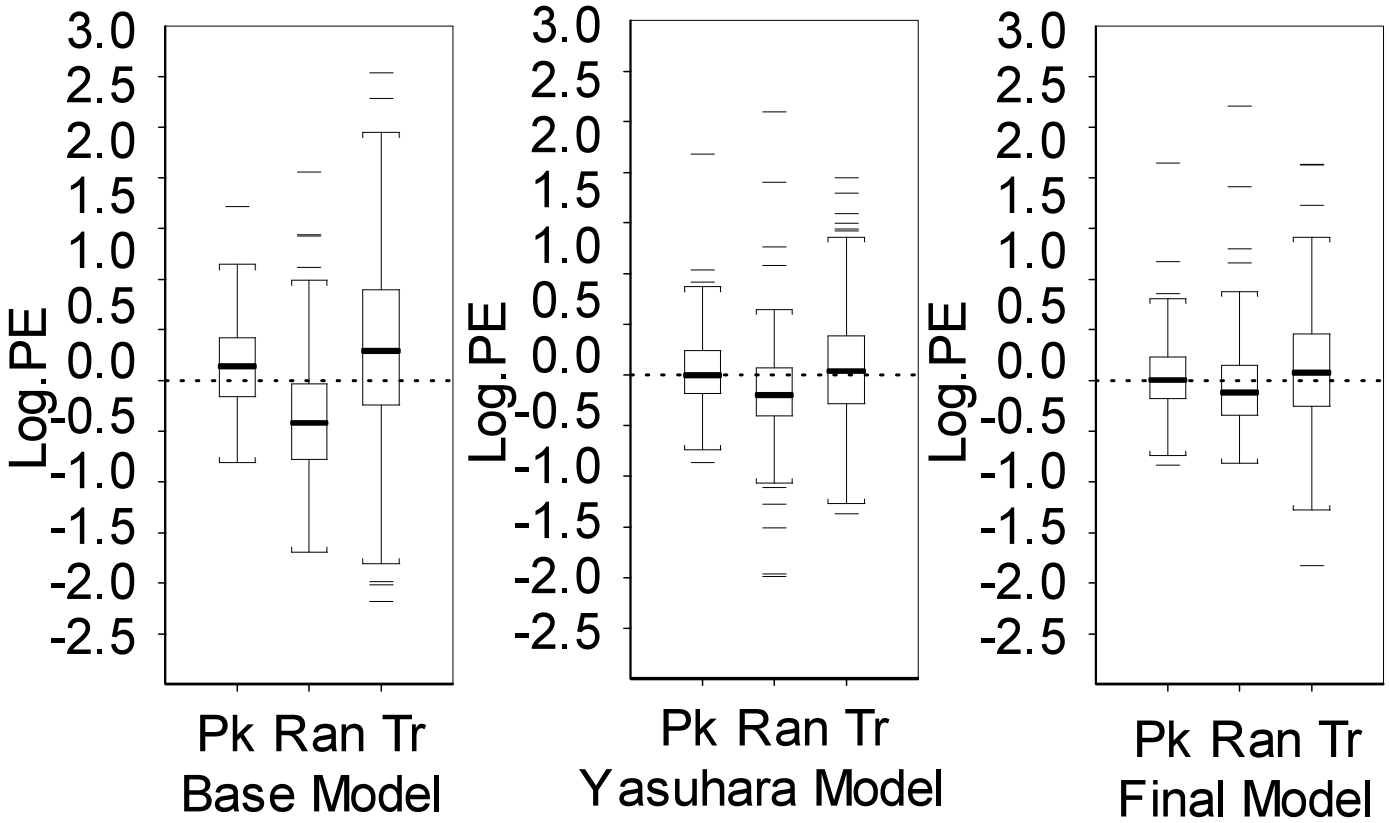
# Cross-Validation

1. Procedure similar to Leverage Analysis
2. Split data into multiple subsets
3. Each of the *m*-1 estimation subsets is used to predict into the remaining (unused) subset
4. Create diagnostic plots
5. Calculate prediction errors

# Cross-Validation Example 1



# Cross-Validation Example 2




*Data from: M. Riggs. Doctoral Thesis Dissertation, UConn, 2000.*

# Posterior Predictive Check

- Proposed to check performance of hierarchical Bayesian models
- Do simulations based on the model & parameters (mean and variance) result in parameter (or response) distributions that are similar to the observed distribution?

***Gelman et al.*** *Model Checking and Sensitivity Analysis. In Bayesian Data Analysis. Chapman and Hall: New York (1995).*

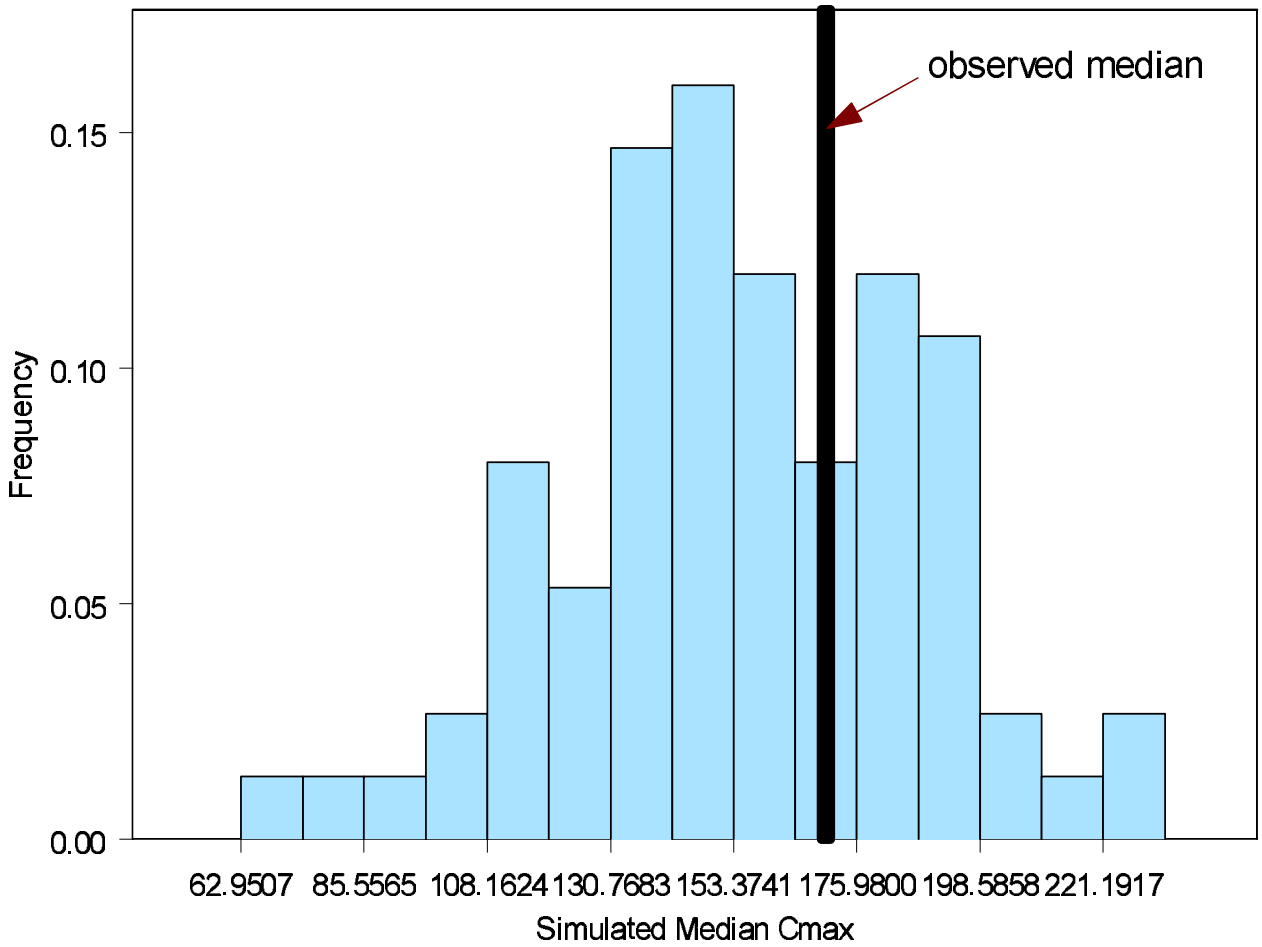
# Posterior Predictive Check

1. Obtain final model and parameter estimates
2. Simulate several (100+) replicates of the original data set using final model fixed and random effect parameters
3. Estimate population fixed and random effect parameters for each one of these replicates
4. Simulate several (100+) new replicates with each replicate using simulation parameters equal to one of the sets of estimated parameters from step 3.

# Posterior Predictive Check

5. From each of the simulations in step 4, calculate a characteristic of the data that is of interest (e.g. $C_{max}$)
6. Summarize $C_{max}$ across all simulations (e.g. median, 1$^{st}$ quartile, etc.)
7. Calculate the same summary statistic from the original data set
8. Plot distribution of simulated statistic with observed value

# Posterior Predictive Check



# Posterior Predictive Check

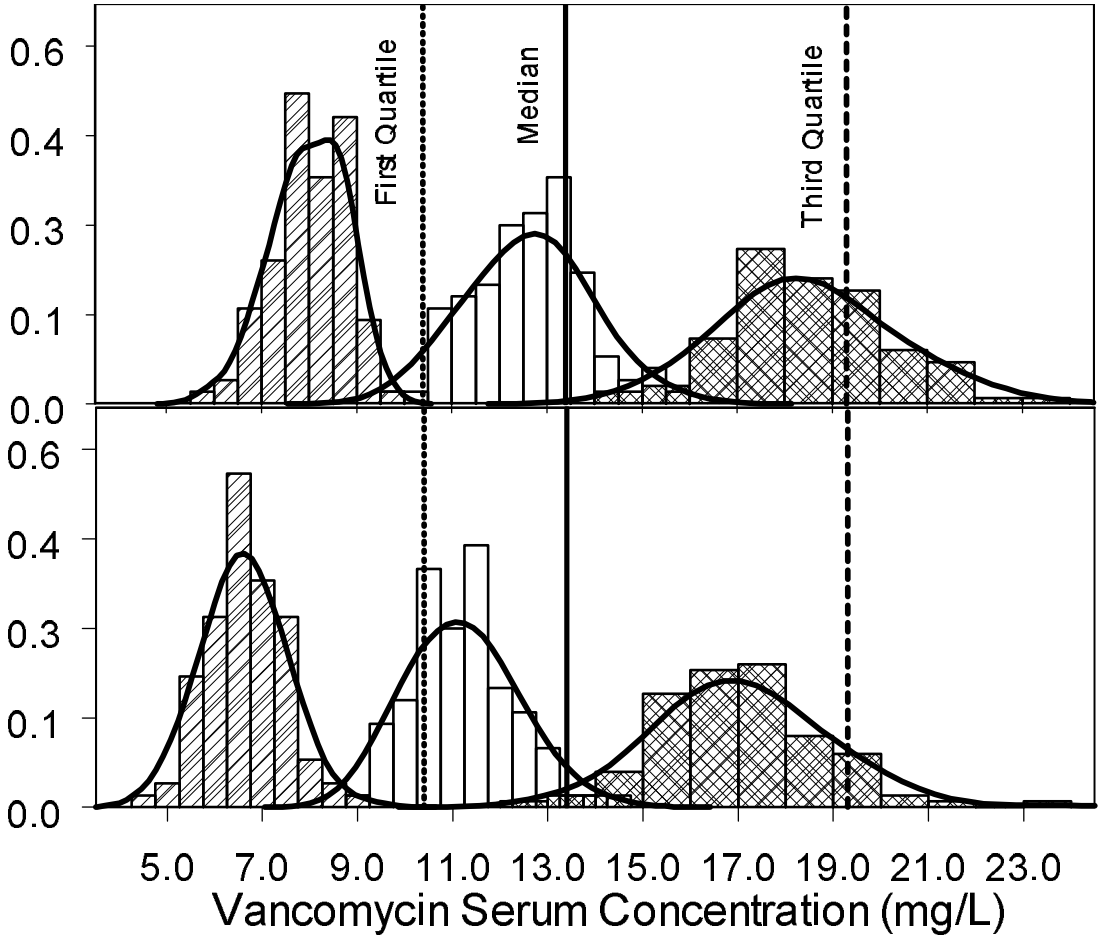
- The resulting distribution reflects both uncertainty in parameter estimates and random inter-individual & residual variability
- This process is very computationally intensive

# Predictive Check (A Shortcut)

1. Assume uncertainty in parameters is small relative to other sources of variability
2. Perform one set of simulations using final model parameters
3. Calculate statistic of interest as in Posterior Predictive Check (steps 5 – 8)

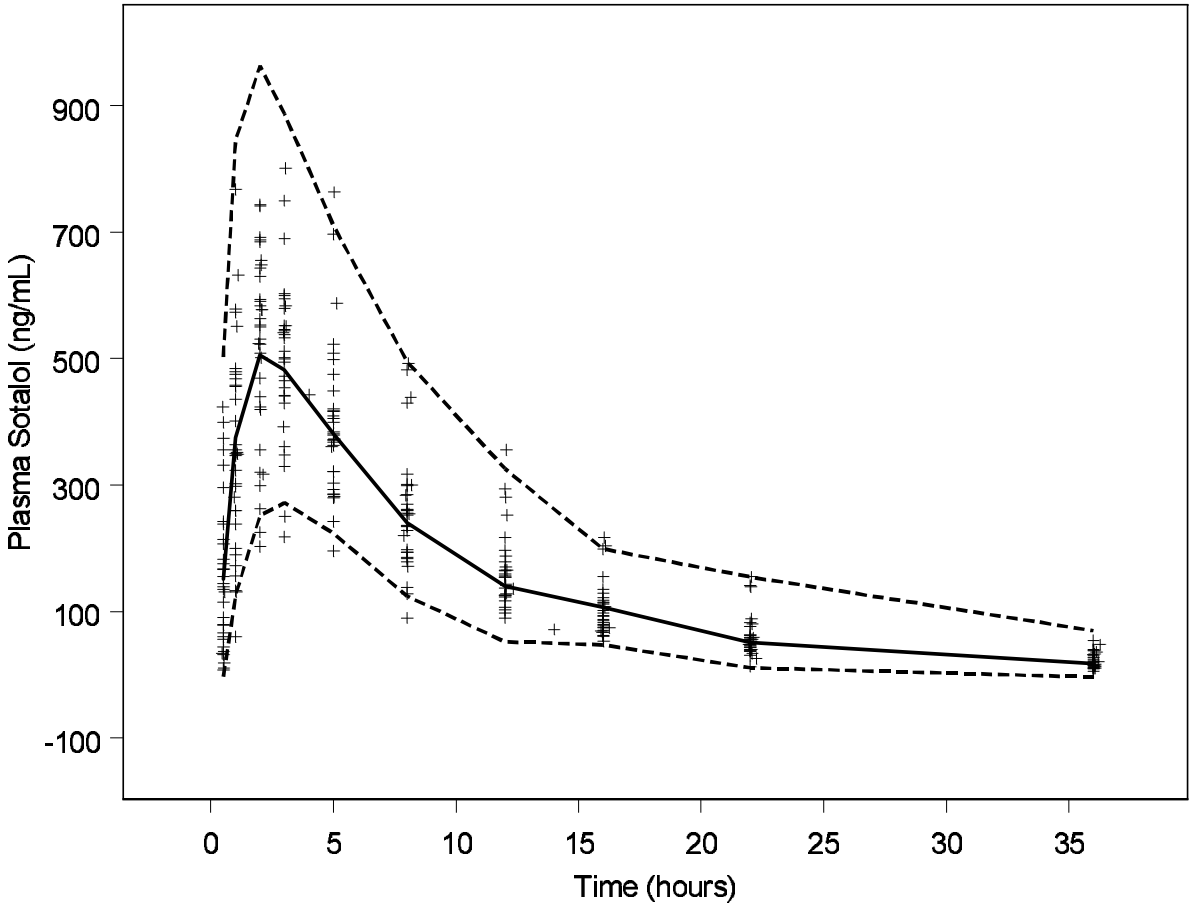
OR - Compare observed data with simulated data

# Predictive Check Example 1




*Data from: M. Riggs. Doctoral Thesis Dissertation, UConn, 2000.*

# Predictive Check Example 2




*adapted from:* Shi, J. et al. *JPKPD*, (28): 555-575, 2001.

# When is Evaluation Less Important?

- The model is purely descriptive
- The impact of model-derived inferences is insignificant
- The model is applied to a purpose that is not easily validated (extrapolation)
- The goal of model validation is simply to satisfy a requirement, or check a box

☑ valid model ☐ you've wasted the last 3 months

# Sensitivity Analysis

# Sensitivity Analysis

- Sometimes, the model is applied to a purpose that is not easily evaluated (extrapolation)
- It may be more important to determine how the model inadequacies affect the conclusions drawn from the modeling application
- Example: clinical trial simulations

# Types of Sensitivity Analysis

- Local sensitivity analysis
  (fixed-point perturbations)

- Global sensitivity analysis
  (based on uncertainty distributions across all parameters)

# Local Sensitivity Analysis

- Can be performed by evaluation of gradients w.r.t. each parameter:

  sensitivity=d(response)/d(parameter)

- Or by simulating with fixed-point perturbations in the parameter and subsequent comparison of simulation response endpoints.

| Fixed Value of ZDVSL | % Trials Successful[a] |
|---|---|
| 0.25 | 30.6% |
| 0.5 | 70.4% |
| 0.735 | 93.0% |
| 1.0 | 99.0% |

[a]Results reflect 500 simulated trials of 2000 patients

# Limitations of Local Sensitivity Analysis

- Only reflects sensitivity to uncertainty in 1 parameter (assumption) at a time
- Inefficient: Must repeat the simulation exercise for each parameter of interest
- Conclusions about sensitivity to assumptions are only accurate if fixed values of all other parameters are correct

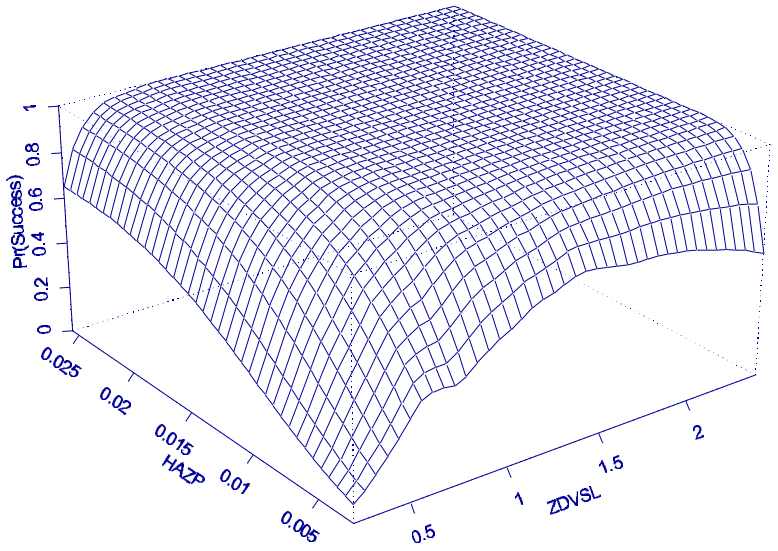
# Global Sensitivity Analysis

- Uncertainty is quantitatively defined for all parameters (models)
- Monte Carlo methods are required to simulate from uncertainty distributions (usually requires one set of simulations with large number of replicates)
- Uncertainty distributions are implemented as inter-trial variability

# Global Sensitivity Analysis

- Sensitivity of simulation outcome(s) to assumptions can be viewed over a continuous range of parameter uncertainty.

- Characterizes sensitivity to uncertainty in all model parameters (assumptions) simultaneously



# Global Sensitivity Analysis in PK-PD

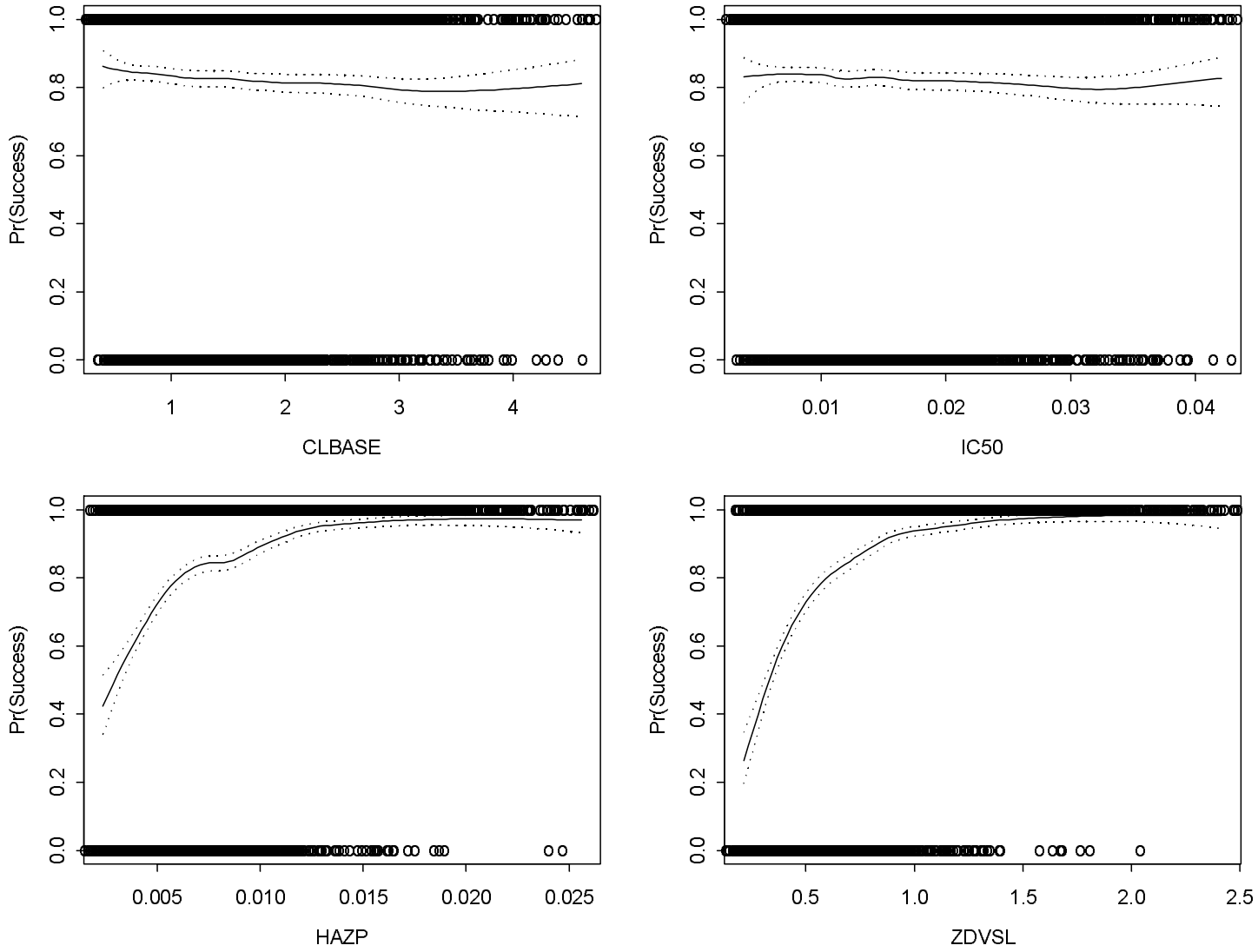
- Examples in PBPK literature: simulation from a range of parameter values
  - Bois, F. et al. *Toxicol. Appl. Pharmacol.* 110: 79-99, 1991.
- New methods proposed include use of fuzzy numbers
  - Nestorov, I. et al. *Drug Metab. Dispos.* 30:276-282, 2002.

- PBPK examples typically lump parameter uncertainty and inter/intra-individual variability together

# Global Sensitivity Analysis in CTS

- Incorporate parameter uncertainty in CTS using Bayesian prior probability distributions for mean and variance parameters in the simulation model
  - Gillespie, W.R., et al. *Clin. Pharmacol. Ther.* 65: PIII-21, 1999.
- A hierarchical model is employed where uncertainty is implemented as inter-trial variability in the model and parameters

# Global Sensitivity Analysis



# To Think About:

- What are the underlying assumptions and inadequacies of a particular model?

- Will the assumptions and inadequacies of a model have a significant impact on the inferences drawn from applications of the model?

# References


- LB Sheiner, SL Beal. Some Suggestions for Measuring Predictive Performance. *JPB* 9: 503-512 (1981).
- S Vozeh, PO Maitre, DR Stanski. Evaluation of Population (NONMEM) Pharmacokinetic Parameter Estimates. *JPB* 18:161-173 (1990).
- SL Beal. "Validation of a Population Model." E-mail to NONMEM Users Network, February 1, (1994). (http://www.cpb.uokhsc.edu/common/anonymous/nm/topic006.html)
- F Mentre, ME Ebelin. Validation of Population Pharmacokinetic/ Pharmacodynamic Analyses: Review of Proposed Approaches. In L Aarons, LP Balant, M Danhof, M Gex-Fabry, UA Gundert-Remy, MO Karlsson, F Mentre, PL Morselli, F Rombout, M Rowland, J-L Steimer & S Vozeh. eds). *The Population Approach: Measuring and Managing Variability in Response, Concentration and Dose*, COST, Brussels, (1997).
- Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic /pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn 2001; 28(2):171-192.

# References (continued)


- EI Ette. Population Model Stability and Performance. *J Clin Pharmacol* 37:486-495 (1997).
- FDA. Guidance for Industry: Population Pharmacokinetics (1997).
- Gelman et al. Model Checking and Sensitivity Analysis. In *Bayesian Data Analysis.* Chapman and Hall: New York (1995).
- T Hastie, R Tibshirani. Modern Regression and Classification Short Course (1998) (http://www-stat.stanford.edu/~trevor/mrc.html)
- http://davidmlane.com/hyperstat/B163479.html (randomization test)


**Validation Applications in Population PK/PD**


- R Bruno et al. *JPB* 24:153-172 (1996).
- JP Mandema et al. *Br J Clin Pharmacol* 42: 747-756 (1996).
- Grasela et al. *Eur J Clin Pharmacol* 45:123-128 (1993).
- Fattinger et al. *Br J Clin Pharmacol* 31:279-286 (1991).
- Aarons et al. *Br J Clin Pharmacol* 28:305-314 (1989).