

Mediating the Meeting of Model and Data: Statistical Issues for PBPK Modeling

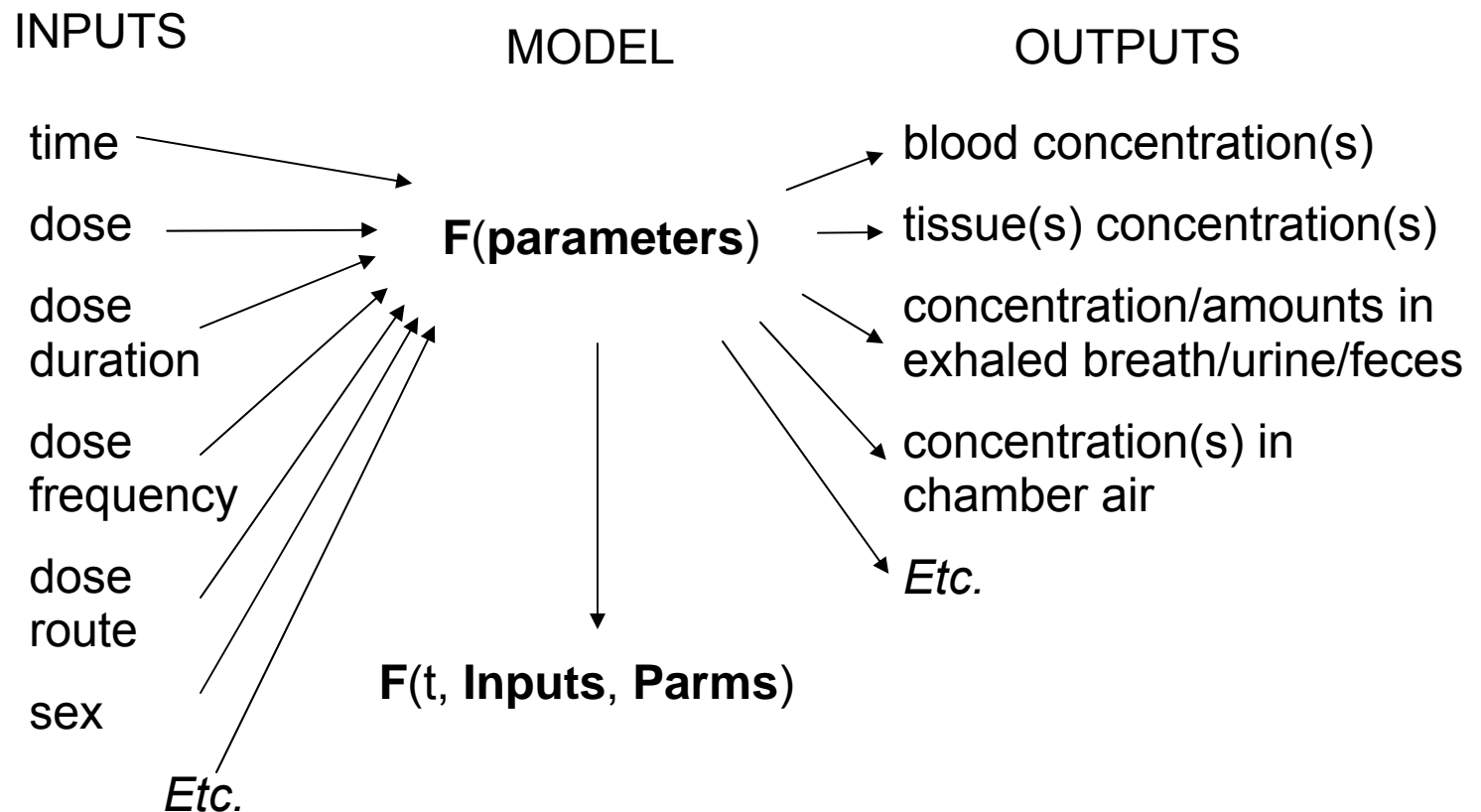
R. Woodrow Setzer
National Center for Computational Toxicology
US EPA
Research Triangle Park, NC
31 October 2006

Disclaimer: This presentation does not present official Agency Policy.

Goals

- Outline some statistical concepts fundamental to the analysis of PBPK models
- Identify some issues raised by statistical analysis of PBPK models.

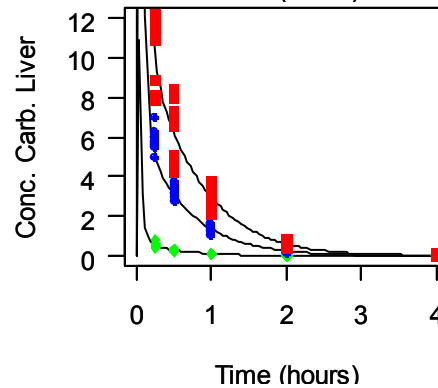
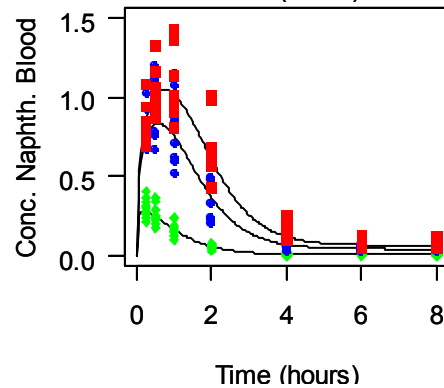
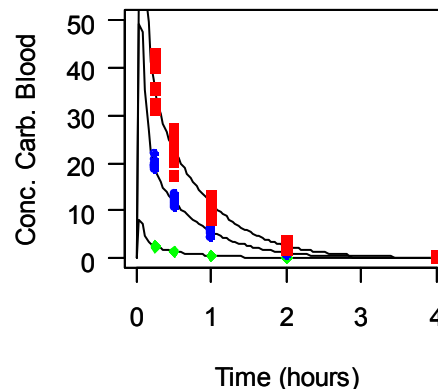
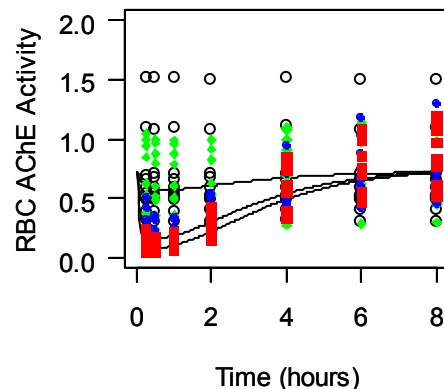
An Abstraction of a PBPK Model: An Abstraction of an Abstraction



A Meeting Between Model and Data is Inevitable

- Does the model describe reality (as revealed in our data)?
- Which (of several) model works better?
- What sets of parameter values are consistent with the data?
- Context of analysis
 - Model development (*i.e.* , generally, on our own model)
 - model evaluation (*i.e.* , generally, on someone else's model).

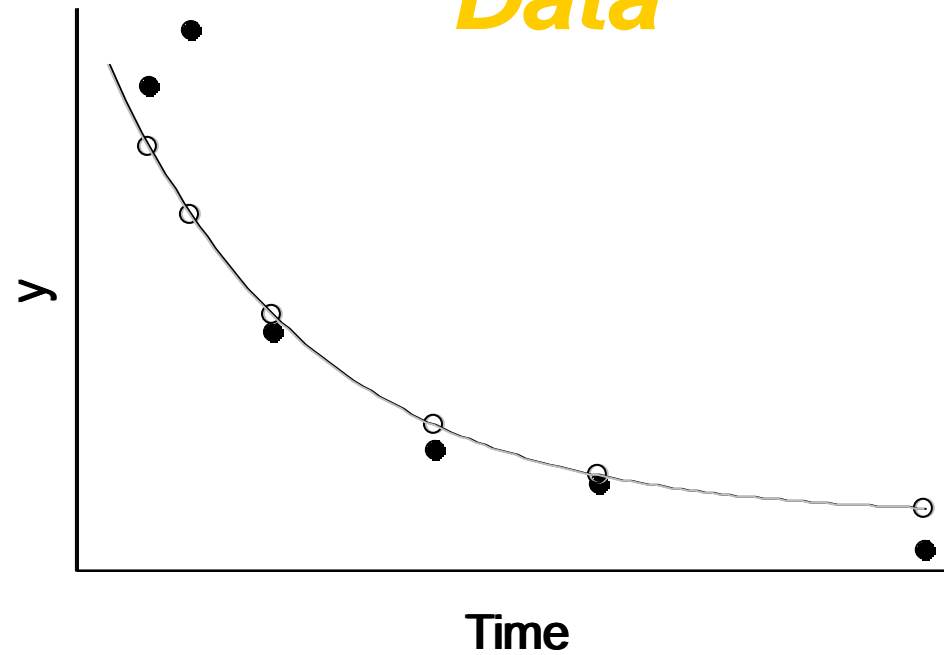
What Happens When We Try?



The Model is Incomplete!

- The PBPK model specification typically characterizes MEAN CHANGES in the tissue concentration-time relationship.
- We have neglected all the factors that add variation to the data, e.g.:
 - measurement error
 - usually the result of a complex set of operations
 - variation among subjects:
 - inherent differences among subjects
 - differences from subject to subject in dosing
 - consumption
 - variation among observations at different times within a subject
 - change in ventilation rates
 - change in cardiac output
 - variation from experiment to experiment
 - ...

Stochastic Models Link Deterministic Model and Data

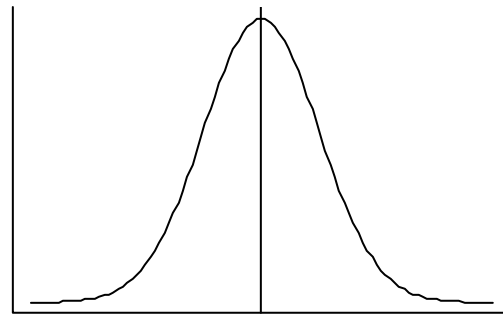


Stochastic Models Link Deterministic Model and Data

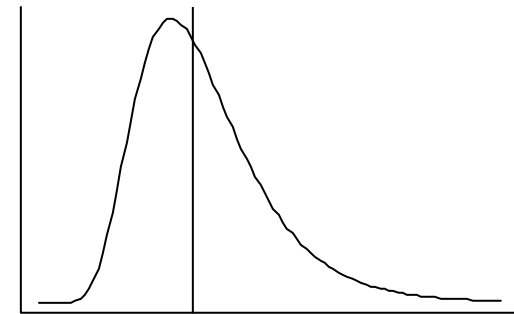
Measurement error e.g.:

$$y = F(t, \text{Inputs}, \text{Parms}) + \varepsilon$$

$$y = F(t, \text{Inputs}, \text{Parms}) \times \varepsilon$$



ε



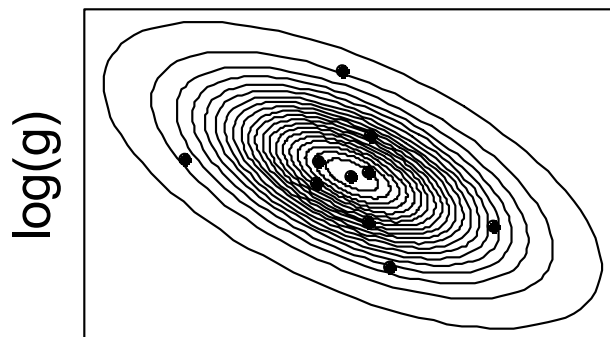
ε

Parameters Varying among Subjects

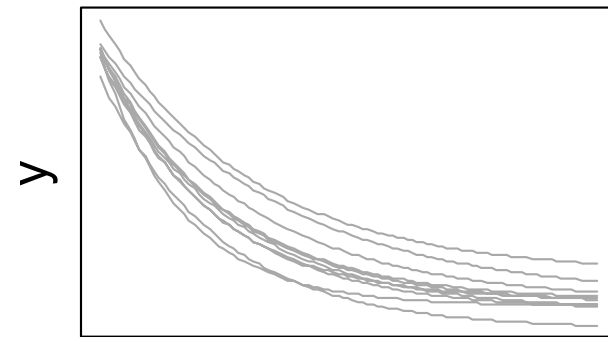
$$y = F(t, \text{Inputs}, [A, g, \text{Parms}]) + \varepsilon$$

$$[\ln(A), \ln(g)] \sim N(\mu, \Sigma)$$

$$\varepsilon \sim N(0, \sigma^2)$$



$\log(A)$



Time(hours)

The Likelihood Function

- Combines model and data in a single expression.
- Basis for both major approaches to inference about models.
- Example (lognormal error model):

$$L(\mathbf{Parms} \mid \mathbf{t}, \mathbf{Inputs}, \mathbf{y}) = \sigma^{-n} \prod_{i=1}^n \frac{1}{y_i} e^{\frac{-(\ln(y_i) - \ln(F(t_i, \mathbf{Inputs}_i, \mathbf{Parms})))^2}{2\sigma^2}}$$

How Do We Incorporate Prior Information?

- *E.G.*, physiological parameters.
- On the one hand we generally know their values to a reasonably small ballpark
- On the other, we generally do not know them for the subjects at hand, and it may matter.
- So... we either:
 - *Treat such parameters as known and fixed*
If the model does not fit, would we come to a different conclusion if we knew the true values?
 - *Estimate them along with other parameters*
Does the data have enough information to determine them all uniquely?

Bayesian Inference

- Based on Bayes's Rule from probability theory:

$$p(\beta | y) = \frac{p(y | \beta) p(\beta)}{\int p(y | \beta) p(\beta) d\beta}$$

Diagram annotations:

- posterior (points to $p(\beta | y)$)
- Likelihood (points to $p(y | \beta)$)
- prior (information before data) (points to $p(\beta)$)

The denominator is the difficulty!

- The posterior summarizes all the information we have (both from the data, and prior information) about the parameters
- We often use Markov Chain Monte Carlo (MCMC) to sample from posterior without evaluating the integral.

Nomenclature

- MCMC is a method for generating samples from a posterior distribution.
- Bayesian analysis is an approach to statistical inference that allows information from data to be combined with prior information.
- Bayesian analysis is NOT synonymous with MCMC analysis.

Inference Summary

- Likelihood:
 - Fit: Evaluate relevant functions comparing model to data (e.g., residuals, overall GOF measures like χ^2 , etc.)
 - Test: for consistency between data and particular parameter values.
 - Estimates are parameter values that maximize the likelihood. CI based on Test (above)
 - Compare models: nested models – test above; non-nested models -- AIC
- Bayes:
 - Fit: Distribution of relevant functions comparing model to data; do posteriors for parameters with informative priors differ substantially from priors?
 - Test: Compare parameter values of interest with posterior dist.
 - Estimates: Characteristics of posterior dist. of parameters, e.g., mean.
 - Compare models: Several methods, incl. Bayesian Information Criterion, BIC; Deviance Information Criterion, DIC.

PBPK Models Are Not So Special

Formally, the statistical analysis of PBPK models follows the lines of the statistical analysis of any other non-linear model.

But, there may be some practical difficulties!

Identifiability

- A system is identifiable if the set of inputs and outputs uniquely determine the parameter values for the system.
- Identifiability may fail for all sets of inputs and outputs (structural identifiability) or for some subsets (statistical identifiability).
- Failure of structural identifiability:

$$y = e^{(a+b)x}$$

- Failure of statistical identifiability:

$$y = \frac{Vx}{K + x},$$

$$x \ll K$$

Identifiability (cont)

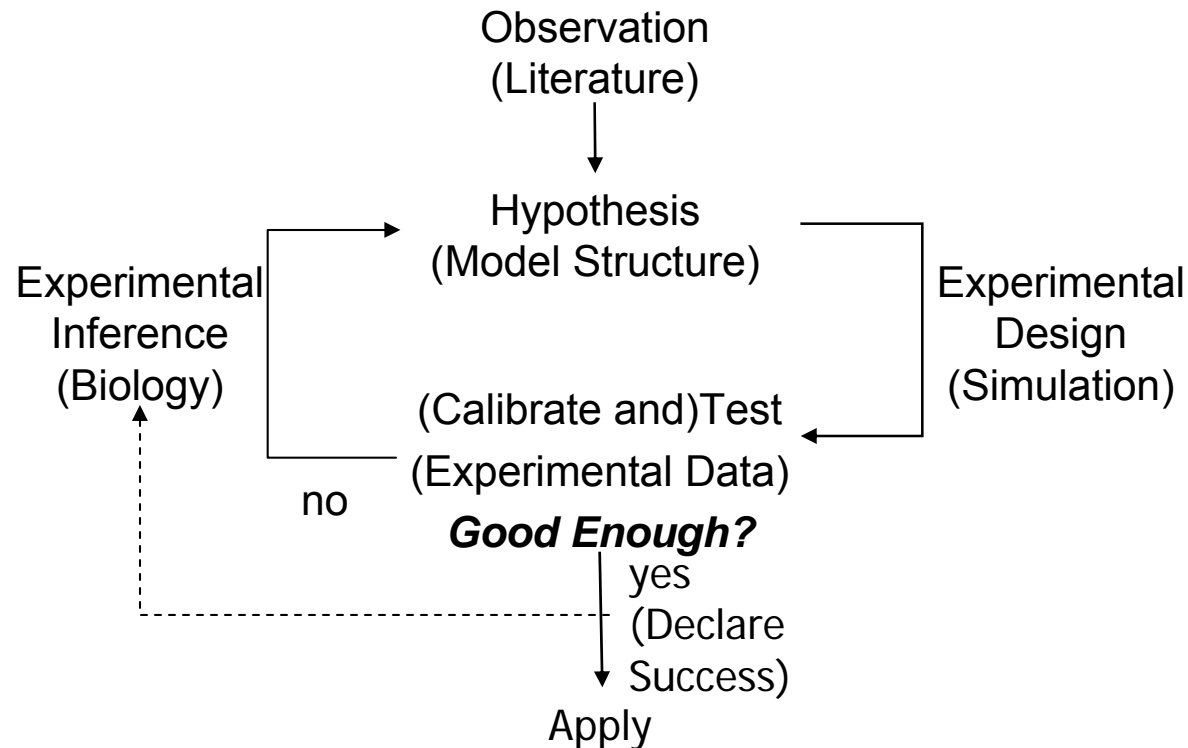
- Just because parameters are not identifiable, does not mean we cannot get “estimates”.
- For example, in the “structural identifiability” example for the previous slide, both gradient-free and gradient-based optimization methods are quite happy to report convergence.
- But, what do they mean?
- Need to be able to diagnose failures of identifiability. (Slob et al., 1997)

Multiple Datasets

- Statistical analysis of PBPK models generally involves analyzing multiple datasets.
- Typically, this complicates the stochastic model and inference:
 - Need another submodel for experiment to experiment variation
 - Limited or no replication of experiments limits the ability to estimate among-experiment error (but does not generally make it go away!)
 - Experiments look at different aspects of the model
- Much data are available only as summaries, complicating checking of the statistical model.

Time/Effort

Ideally, statistical analysis is an integral part of the model development cycle, but the time/effort can be a disincentive!



Software

- Need software that is:
 - flexible, so that alternative experimental designs and probability models are easily accommodated.
 - numerically sophisticated, to be able to handle the complicated combination of ODEs, often with time-varying inputs, and algebraic equations that make up a PBPK model.
 - computationally sophisticated, to take advantage of, for example, cluster computing to reduce total computation time.
 - easy to learn and use, so modelers spend their time thinking about modeling and data, not about programming and computational details!

Summary

- Comparing a PBPK model and data requires a model for the variability.
- If this is a stochastic model, the likelihood is the link between the PBPK model and data.
- Likelihood-based and Bayesian methods provide approaches for answering the questions we have about the relationship of the model to the world (though Bayesian methods have some advantages).
- PBPK models are not unique in their statistical analysis, but there are some features of PBPK models that complicate their statistical analysis.