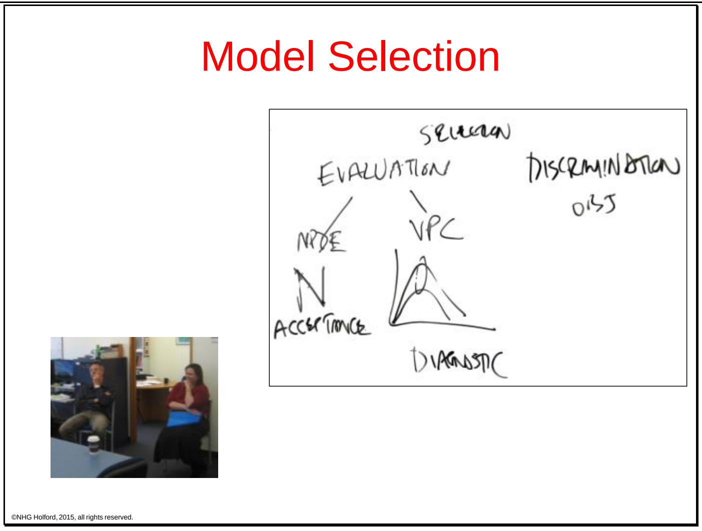
Slide 1

Model Evaluation

Nick Holford

Dept Pharmacology & Clinical Pharmacology
University of Auckland, New Zealand

Slide 2

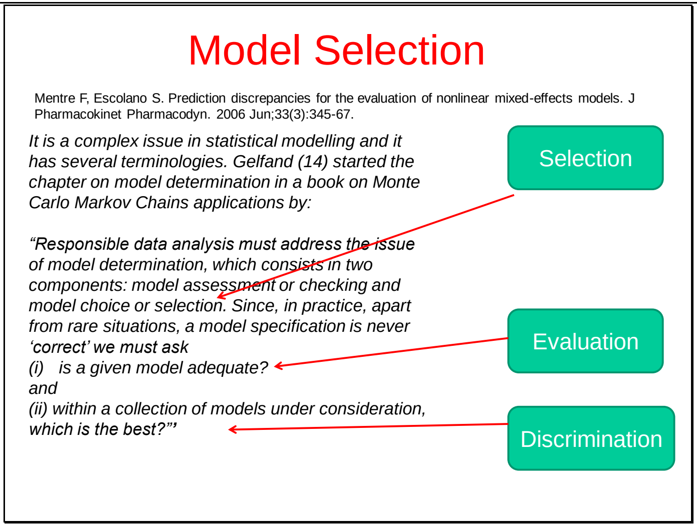


Model Selection

Steve Duffull, Nick Holford and Catherine Sherwin sorted out how to select models over a cup of coffee in the Modelling and Simulation Lab at the University of Otago School of Pharmacy in Dunedin, NZ (12 Nov 2008)

Model selection typically involves an initial Discrimination step using goodness of fit (e.g. objective function value) to find a candidate model for evaluation. Evaluation may use a diagnostic process (e.g. VPC) for learning about the model weaknesses ("all models are wrong") which may suggest how to improve the model then an acceptance process (e.g. NPDE) for confirming a model ("some models are useful").

Slide 3



Model Selection

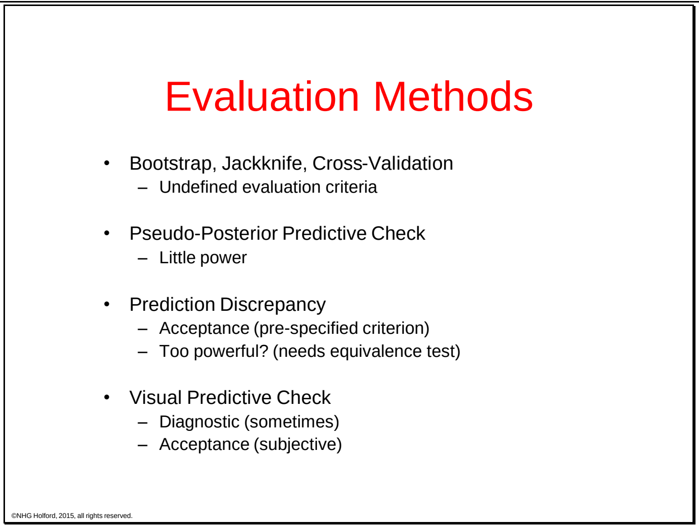
Mentre F, Escolano S. Prediction discrepancies for the evaluation of nonlinear mixed-effects models. J Pharmacokinet Pharmacodyn. 2006 Jun;33(3):345-67.

It is a complex issue in statistical modelling and it has several terminologies. Gelfand (14) started the chapter on model determination in a book on Monte Carlo Markov Chains applications by:

"Responsible data analysis must address the issue of model determination, which consists in two components: model assessment or checking and model choice or selection. Since, in practice, apart from rare situations, a model specification is never 'correct' we must ask
(i) is a given model adequate?
and
(ii) within a collection of models under consideration, which is the best?"'

Several authors have tried to describe the fundamental processes of model building and decisions about models. Gelfand proposed 'determination' but 'selection' seems to capture the overall picture more clearly.

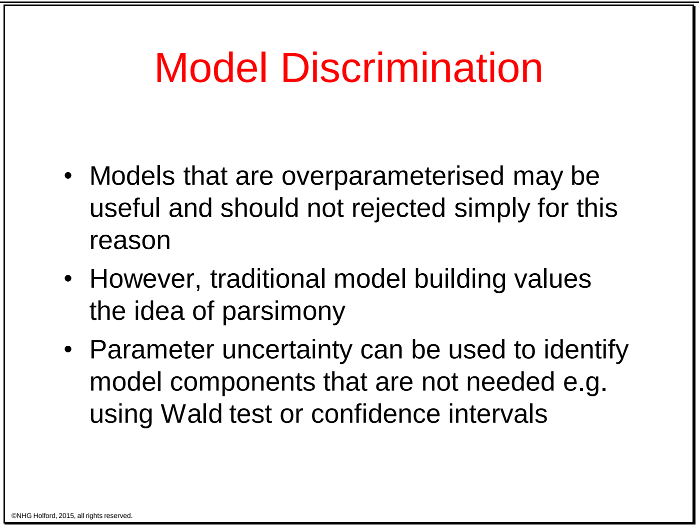
Slide 4



Evaluation Methods

- Bootstrap, Jackknife, Cross-Validation
  - Undefined evaluation criteria
- Pseudo-Posterior Predictive Check
  - Little power
- Prediction Discrepancy
  - Acceptance (pre-specified criterion)
  - Too powerful? (needs equivalence test)
- Visual Predictive Check
  - Diagnostic (sometimes)
  - Acceptance (subjective)

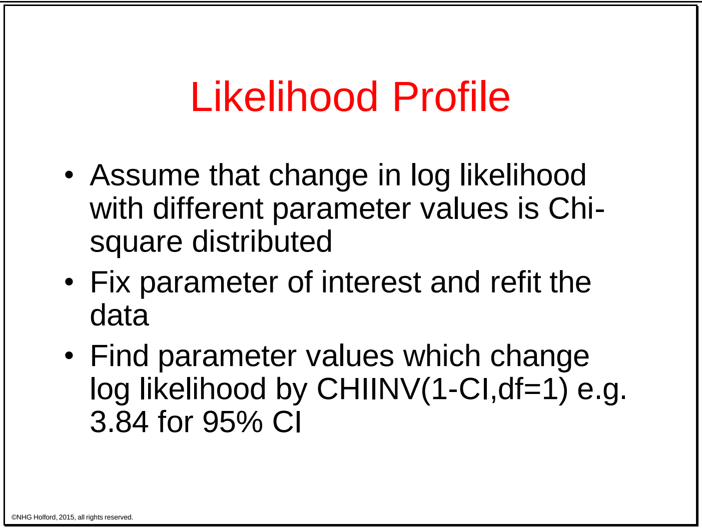
Slide 5



Model Discrimination

- Models that are overparameterised may be useful and should not rejected simply for this reason
- However, traditional model building values the idea of parsimony
- Parameter uncertainty can be used to identify model components that are not needed e.g. using Wald test or confidence intervals

Slide 6

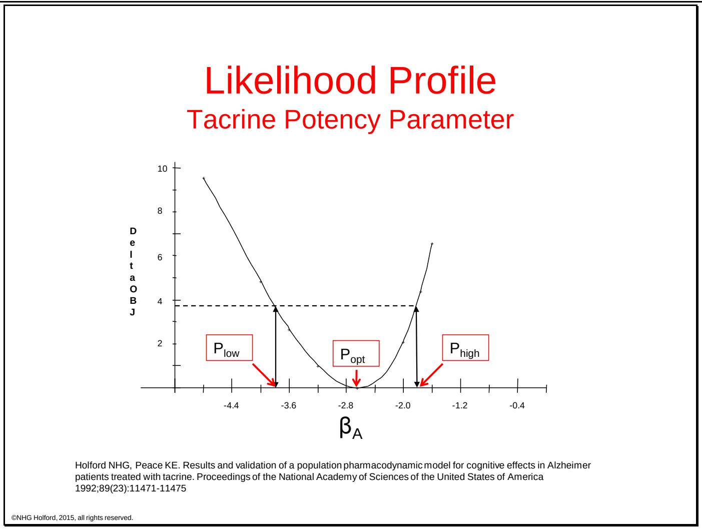


Likelihood Profile

- Assume that change in log likelihood with different parameter values is Chi-square distributed
- Fix parameter of interest and refit the data
- Find parameter values which change log likelihood by CHIINV(1-CI,df=1) e.g. 3.84 for 95% CI

The log likelihood profile method does not assume symmetry of the parameter uncertainty but it does use the likelihood ratio test (LRT) based on the change in NONMEM objective function value to predict the probability of the confidence interval. This assumption is known to be only approximately true (see discussion of the randomization test).

Slide 7



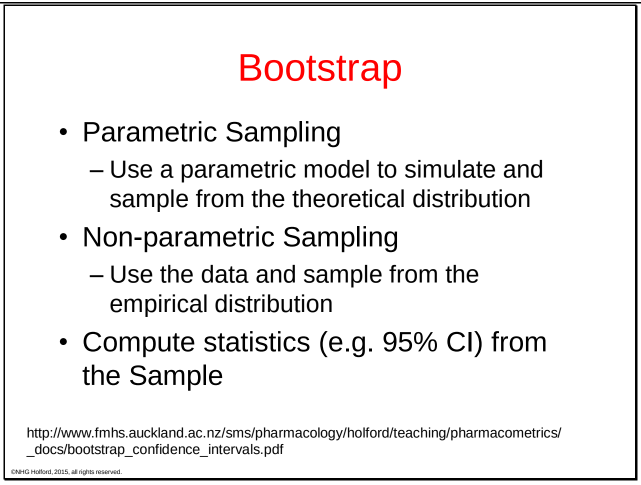
Likelihood Profile

Tacrine Potency Parameter

Holford NHG, Peace KE. Results and validation of a population pharmacodynamic model for cognitive effects in Alzheimer patients treated with tacrine. Proceedings of the National Academy of Sciences of the United States of America 1992;89(23):11471-11475

A log likelihood profile (LLP) is illustrated here. The parameter is BetaA the potency parameter for the effect of tacrine at a dose of 80 mg/day. The approximate 95% confidence interval is shown under the assumption of the chi-square distribution. This LLP was obtained using the FO method and therefore the actual 95% CI is almost certainly wider than shown here.

Slide 8



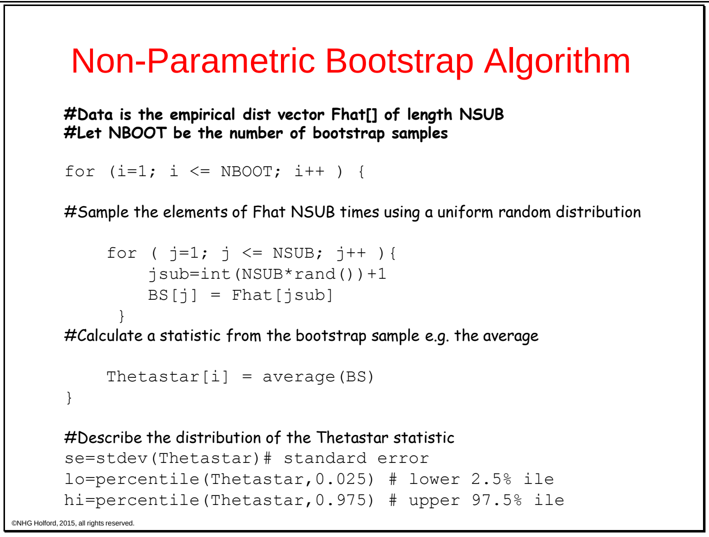
Bootstrap

- Parametric Sampling
  - Use a parametric model to simulate and sample from the theoretical distribution
- Non-parametric Sampling
  - Use the data and sample from the empirical distribution
- Compute statistics (e.g. 95% CI) from the Sample

http://www.fmhs.auckland.ac.nz/sms/pharmacology/holford/teaching/pharmacometrics/_docs/bootstrap_confidence_intervals.pdf

Both the parametric and non-parametric bootstrap procedures can be used to generate samples from their respective distributions. The parametric method requires a full parametric model (e.g. PK model with population parameter variability and residual unidentified variability) while the non-parametric method only requires an original data set.

Slide 9



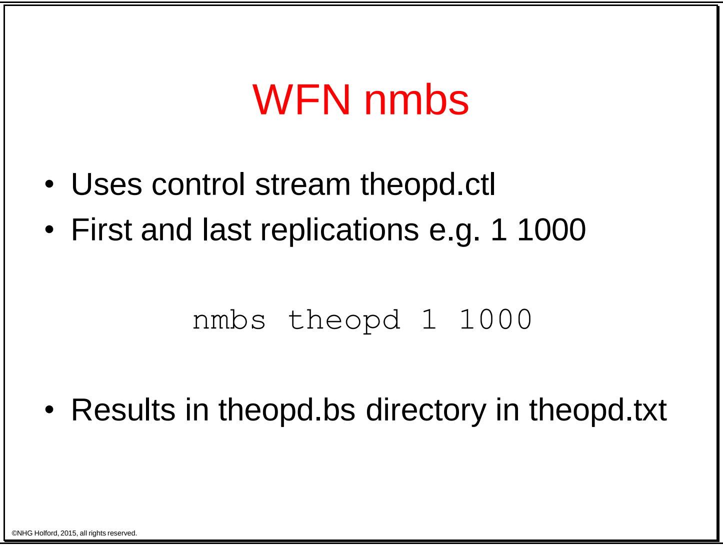
Non-Parametric Bootstrap Algorithm

```
#Data is the empirical dist vector Fhat[] of length NSUB
#Let NBOOT be the number of bootstrap samples

for (i=1; i <= NBOOT; i++ ) {

#Sample the elements of Fhat NSUB times using a uniform random distribution

    for ( j=1; j <= NSUB; j++ ){
        jsub=int(NSUB*rand())+1
        BS[j] = Fhat[jsub]
     }
#Calculate a statistic from the bootstrap sample e.g. the average

    Thetastar[i] = average(BS)
}

#Describe the distribution of the Thetastar statistic
se=stdev(Thetastar)# standard error
lo=percentile(Thetastar,0.025) # lower 2.5% ile
hi=percentile(Thetastar,0.975) # upper 97.5% ile
```

The basic bootstrap algorithm is shown using awk code. NBOOT is the number of bootstrap samples requested. This would typically be 1000 or more to obtain an estimate of the 95% confidence interval. Fhat is the empirical distribution i.e. the original data set. BS is a bootstrap data set obtained by resampling from Fhat. Nsub is the number of subjects. Thetastar is a vector of estimates. It is an empirical distribution of the statistic. In this case the average is computed for each BS sample data set. This step in the algorithm can be much more complex e.g. a NONEMM run using the BS data set can be used to estimate a full set of parameters.

In the last line of the algorithm a meta-analysis procedure is used to examine the results in Thetastar. In this case the standard deviation of the average values in Thetastar is used to estimate the standard error. The same Thetastar array can be used to find the 95% confidence interval by looking for the values of Thetastar that are less than the 2.5%centile and greater than the 97.5%centile.

Slide 10



WFN nmbs

- Uses control stream theopd.ctl
- First and last replications e.g. 1 1000

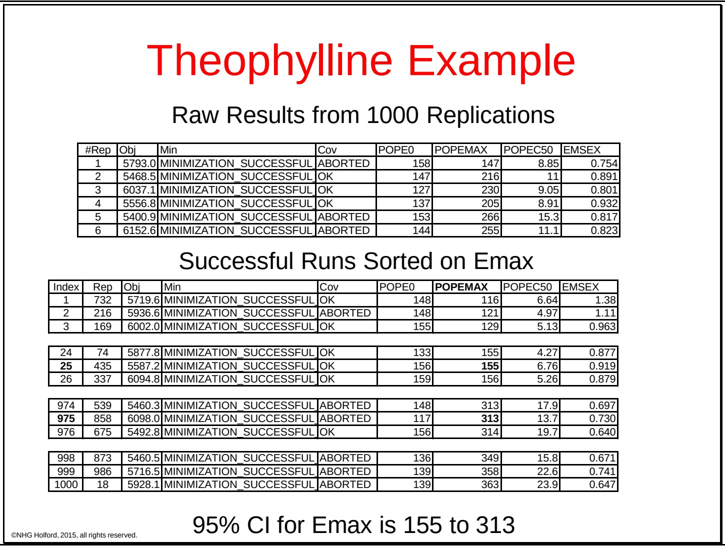
```
nmbs theopd 1 1000
```

- Results in theopd.bs directory in theopd.txt

Wings for NONMEM has an nmbs command to automatically create bootstrap data sets and run NONMEM models. The only restriction is to be sure that any paths that exist in $SUB recognize that the bootstrap NONMEM run is two levels down from the parent directory. It is usually easier to give a fully qualified path for any $SUB user defined subroutines.

The bootstrap results are found in the a *.bs folder in a *.txt file. The *.txt file has the parameter estimates for each bootstrap replicate on one line of the file. They are tab delimited and can be easily read into Excel for further analysis.

Slide 11



Theophylline Example

Raw Results from 1000 Replications

#Rep	Obj	Min	Cov	POPE0	POPEMAX	POPEC50	EMSEX
1	5793.0	MINIMIZATION_SUCCESSFUL	ABORTED	158	147	8.85	0.754
2	5468.5	MINIMIZATION_SUCCESSFUL	OK	147	216	11	0.891
3	6037.1	MINIMIZATION_SUCCESSFUL	OK	127	230	9.05	0.801
4	5556.8	MINIMIZATION_SUCCESSFUL	OK	137	205	8.91	0.932
5	5400.9	MINIMIZATION_SUCCESSFUL	ABORTED	153	266	15.3	0.817
6	6152.6	MINIMIZATION_SUCCESSFUL	ABORTED	144	255	11.1	0.823

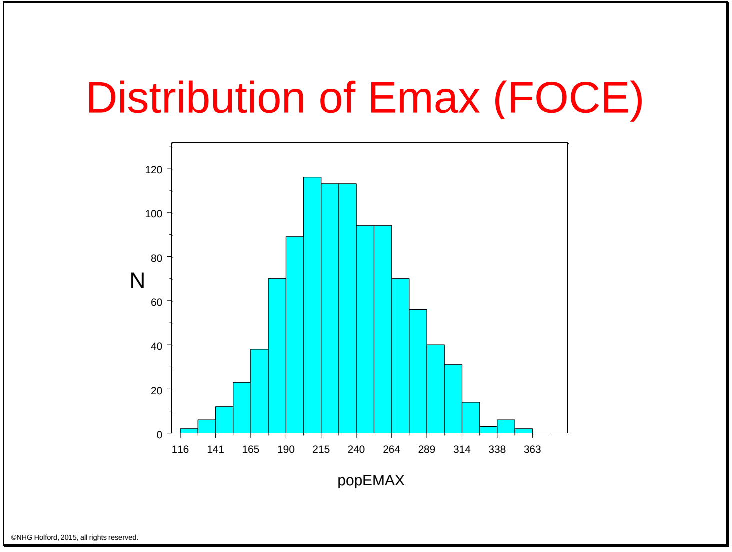
Successful Runs Sorted on Emax

Index	Rep	Obj	Min	Cov	POPE0	**POPEMAX**	POPEC50	EMSEX
1	732	5719.6	MINIMIZATION_SUCCESSFUL	OK	148	116	6.64	1.38
2	216	5936.6	MINIMIZATION_SUCCESSFUL	ABORTED	148	121	4.97	1.11
3	169	6002.0	MINIMIZATION_SUCCESSFUL	OK	155	129	5.13	0.963
24	74	5877.8	MINIMIZATION_SUCCESSFUL	OK	133	155	4.27	0.877
25	435	5587.2	MINIMIZATION_SUCCESSFUL	OK	156	**155**	6.76	0.919
26	337	6094.8	MINIMIZATION_SUCCESSFUL	OK	159	156	5.26	0.879
974	539	5460.3	MINIMIZATION_SUCCESSFUL	ABORTED	148	313	17.9	0.697
975	858	6098.0	MINIMIZATION_SUCCESSFUL	ABORTED	117	**313**	13.7	0.730
976	675	5492.8	MINIMIZATION_SUCCESSFUL	OK	156	314	19.7	0.640
998	873	5460.5	MINIMIZATION_SUCCESSFUL	ABORTED	136	349	15.8	0.671
999	986	5716.5	MINIMIZATION_SUCCESSFUL	ABORTED	139	358	22.6	0.741
1000	18	5928.1	MINIMIZATION_SUCCESSFUL	ABORTED	139	363	23.9	0.647

95% CI for Emax is 155 to 313

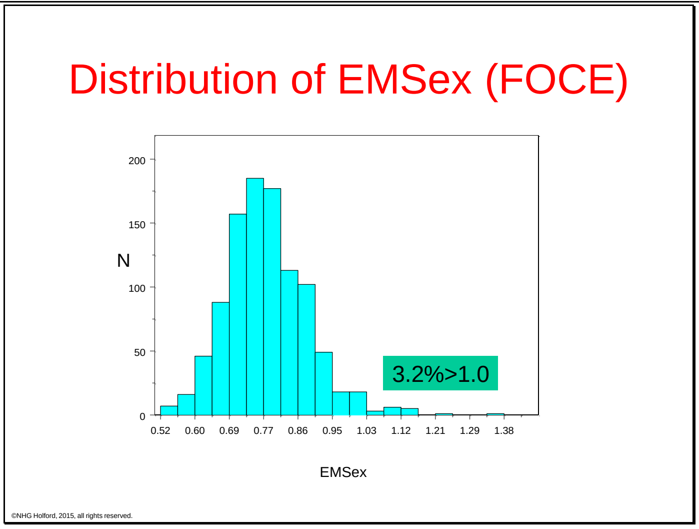
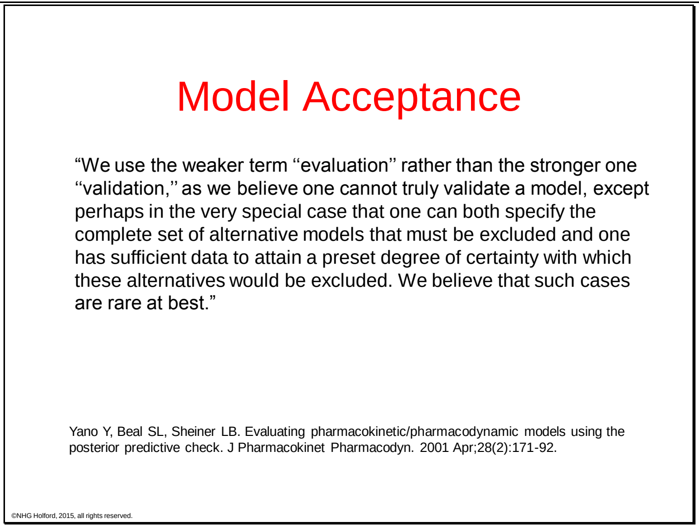
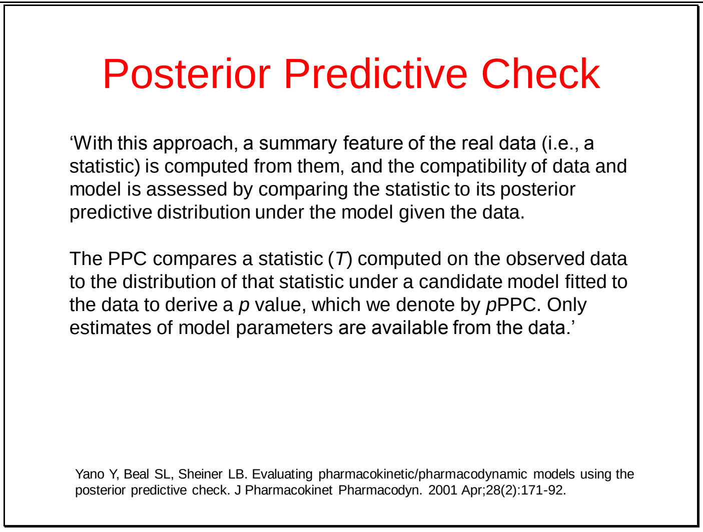
The theopdsex example is shown here. The Raw Results table shows the first 6 replications. The Successful Runs tables shows the same results sorted on the POPEMAX value. The lower 2.5% centile and upper 97.5%centile can be identified from their index in the table and the corresponding POPEMAX estimates used to define the 95% confidence interval for Emax.

Slide 12

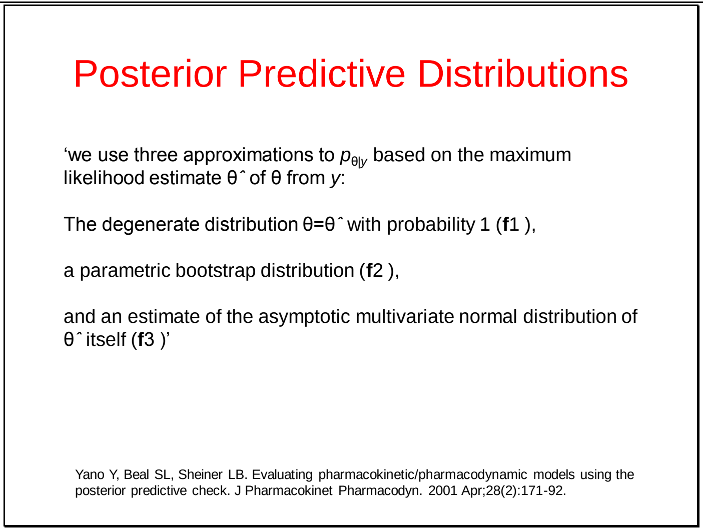


Distribution of Emax (FOCE)

The bootstrap distribution of Emax is shown here. It looks reasonably symmetrical and even normal in shape.

Slide 13	 Distribution of EMSex (FOCE) 	When the sex on Emax model is used the estimate of the reduction of Emax in females is shown above. The mode is about 0.75 which means the typical Emax is 25% lower in females. Only 3.2% of estimates are greater than 1 which provides support that this parameter is different in females.
Slide 14	 Model Acceptance “We use the weaker term “evaluation” rather than the stronger one “validation,” as we believe one cannot truly validate a model, except perhaps in the very special case that one can both specify the complete set of alternative models that must be excluded and one has sufficient data to attain a preset degree of certainty with which these alternatives would be excluded. We believe that such cases are rare at best.” Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001 Apr;28(2):171-92. 	Yano, Beal and Sheiner used simulation based methods to test if a model was evaluate models. They were primarily interested in an acceptance test rather than looking at the model and data for diagnostic purposes.
Slide 15	 Posterior Predictive Check ‘With this approach, a summary feature of the real data (i.e., a statistic) is computed from them, and the compatibility of data and model is assessed by comparing the statistic to its posterior predictive distribution under the model given the data. The PPC compares a statistic (*T*) computed on the observed data to the distribution of that statistic under a candidate model fitted to the data to derive a *p* value, which we denote by *p*PPC. Only estimates of model parameters are available from the data.’ Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001 Apr;28(2):171-92.	

Slide 16



Posterior Predictive Distributions

‘we use three approximations to $p_{\theta|y}$ based on the maximum likelihood estimate θˆ of θ from *y*:

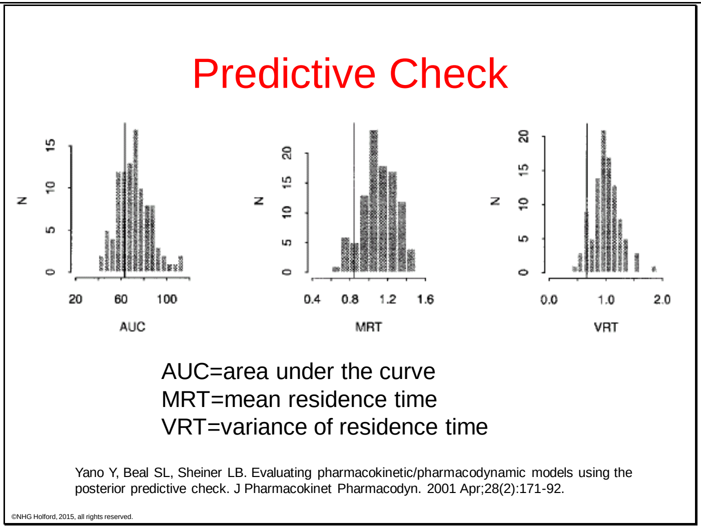
The degenerate distribution θ=θˆ with probability 1 (**f**1),

a parametric bootstrap distribution (**f**2),

and an estimate of the asymptotic multivariate normal distribution of θˆ itself (**f**3)’

Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001 Apr;28(2):171-92.

Slide 17

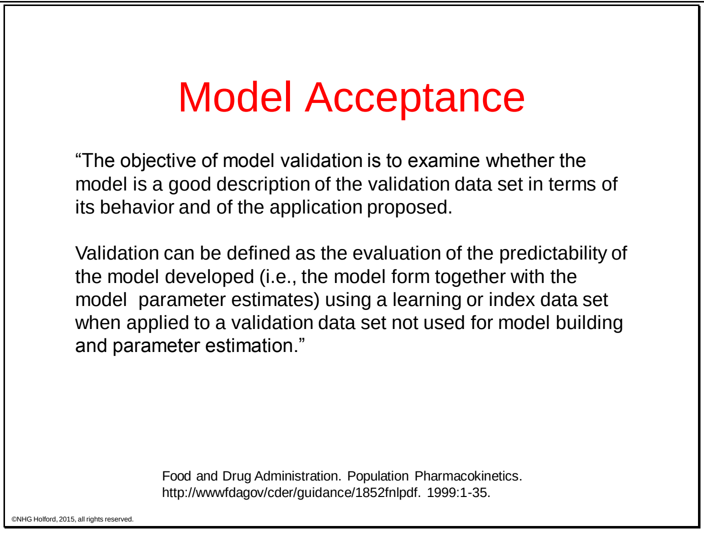


Predictive Check

AUC=area under the curve
MRT=mean residence time
VRT=variance of residence time

Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001 Apr;28(2):171-92.

Slide 18



Model Acceptance

“The objective of model validation is to examine whether the model is a good description of the validation data set in terms of its behavior and of the application proposed.

Validation can be defined as the evaluation of the predictability of the model developed (i.e., the model form together with the model parameter estimates) using a learning or index data set when applied to a validation data set not used for model building and parameter estimation.”

Food and Drug Administration. Population Pharmacokinetics. http://wwwfdagov/cder/guidance/1852fnlpdf. 1999:1-35.

The FDA Guidance on Population Pharmacokinetics was one of the first to grapple with model selection for mixed effect models applied to pharmacokinetics. They proposed an acceptance method based on comparing the predictions from a model developed in a learning data set with the observations in a separate ‘validation’ data set. This is an external acceptance method because it relies on having a second test data set. No criteria for acceptance were proposed. Typical attempts rarely if ever fail because criteria are not specified a priori and often they appear to be underpowered as a consequence of using a test data set that is small (e.g. 1/3) compared to the learning data set.

Slide 19

Prediction Discrepancy

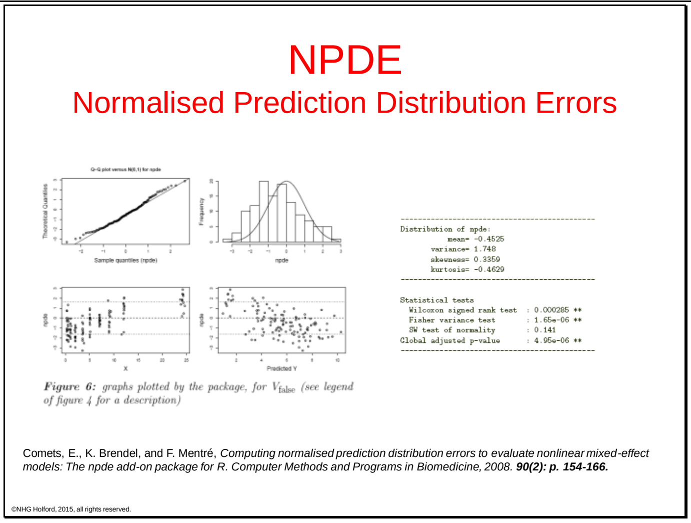
‘We evaluate what we call the “prediction discrepancy” (pd) which is defined as the percentile of an observation in the whole marginal predictive distribution under *H*0.’

Mentre F, Escolano S. Prediction discrepancies for the evaluation of nonlinear mixed-effects models. J Pharmacokinet Pharmacodyn. 2006 Jun;33(3):345-67.

The prediction discrepancy method uses stochastic simulation to generate a distribution of predictions for each observation. The percentile of each observation in this distribution is called the prediction discrepancy. The distribution of prediction discrepancies is expected to be uniform if the model correctly predicts the distribution from which the observations came.

The prediction discrepancy distribution can be ‘normalized’ and also take into account correlations of observations within an individual. The resulting normalized prediction discrepancy distribution (NPDE) should have a mean of zero and a standard deviation of 1. Estimates of these parameters can be computed from the NPDE and tested against the null hypothesis that the distribution is ~N(0,1).

Slide 20



NPDE

Normalised Prediction Distribution Errors

```
-----------------------------------------
Distribution of npde:
          mean= -0.4525
      variance= 1.748
      skewness= 0.3359
      kurtosis= -0.4629
-----------------------------------------

Statistical tests
  Wilcoxon signed rank test  : 0.000285 **
  Fisher variance test       : 1.65e-06 **
  SW test of normality       : 0.141
Global adjusted p-value      : 4.95e-06 **
-----------------------------------------
```

Figure 6: graphs plotted by the package, for V_{false} (see legend of figure 4 for a description)

Comets, E., K. Brendel, and F. Mentré, *Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: The npde add-on package for R. Computer Methods and Programs in Biomedicine, 2008.* ***90(2): p. 154-166.***

The NPDE tests for differences from a perfect fit of the model to the data. Because all models are wrong it is unrealistic to expect a perfect fit. When there is a lot of data the NPDE is sensitive to differences that have no practical relevance. This means it can be considered overpowered and will lead to rejection of the null hypothesis when the model is in fact adequate for purpose.

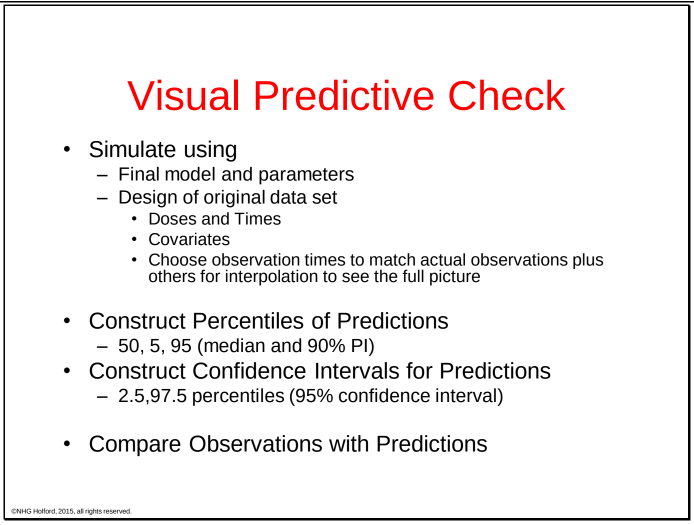
An equivalence type of hypothesis test (such as that used for bioequivalence) is an obvious extension of the method to make it more practically useful as an acceptance method.

Slide 21

Model Diagnosis

- Traditional model diagnostics based on residuals and empirical Bayes estimates can be misleading
- Simulation based diagnostics are more robust

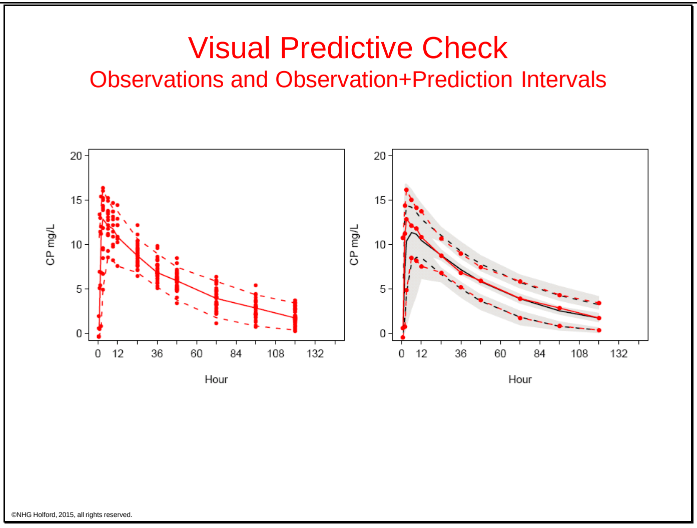
Slide 22



Visual Predictive Check

- Simulate using
  - Final model and parameters
  - Design of original data set
    - Doses and Times
    - Covariates
    - Choose observation times to match actual observations plus others for interpolation to see the full picture
- Construct Percentiles of Predictions
  - 50, 5, 95 (median and 90% PI)
- Construct Confidence Intervals for Predictions
  - 2.5,97.5 percentiles (95% confidence interval)
- Compare Observations with Predictions

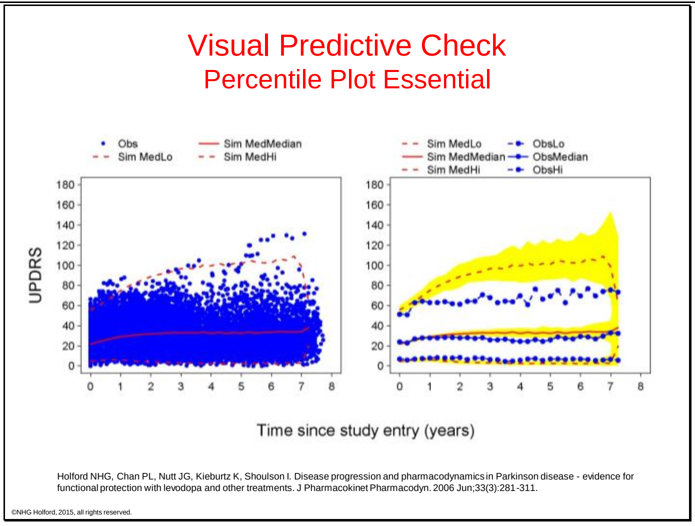
Slide 23



Visual Predictive Check

Observations and Observation+Prediction Intervals

Slide 24

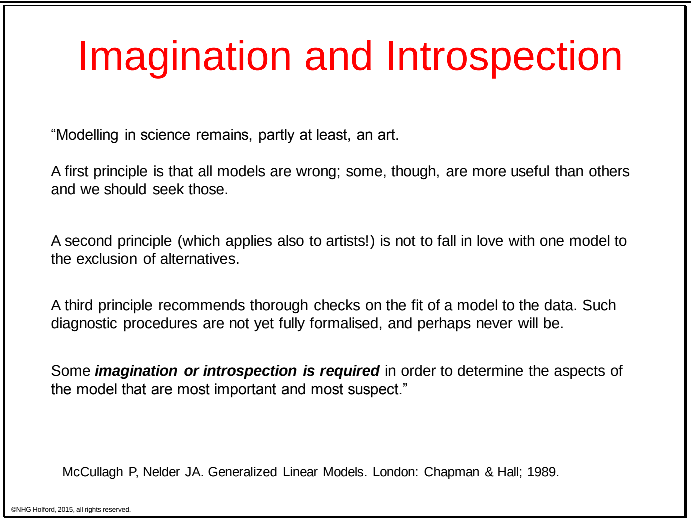


Visual Predictive Check

Percentile Plot Essential

Holford NHG, Chan PL, Nutt JG, Kieburtz K, Shoulson I. Disease progression and pharmacodynamics in Parkinson disease - evidence for functional protection with levodopa and other treatments. J Pharmacokinet Pharmacodyn. 2006 Jun;33(3):281-311.

Slide 25



Imagination and Introspection

"Modelling in science remains, partly at least, an art.

A first principle is that all models are wrong; some, though, are more useful than others and we should seek those.

A second principle (which applies also to artists!) is not to fall in love with one model to the exclusion of alternatives.

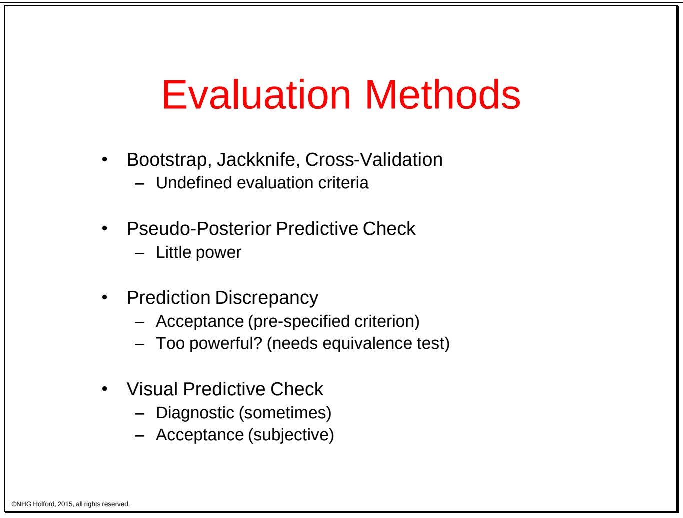
A third principle recommends thorough checks on the fit of a model to the data. Such diagnostic procedures are not yet fully formalised, and perhaps never will be.

Some ***imagination or introspection is required*** in order to determine the aspects of the model that are most important and most suspect."

McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman & Hall; 1989.

Slide 26

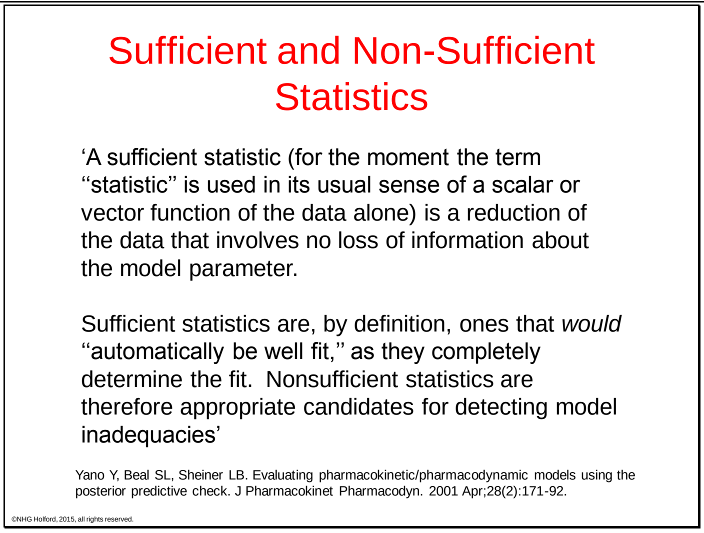
Slide 27



Evaluation Methods

- Bootstrap, Jackknife, Cross-Validation
  - Undefined evaluation criteria
- Pseudo-Posterior Predictive Check
  - Little power
- Prediction Discrepancy
  - Acceptance (pre-specified criterion)
  - Too powerful? (needs equivalence test)
- Visual Predictive Check
  - Diagnostic (sometimes)
  - Acceptance (subjective)

Slide 28



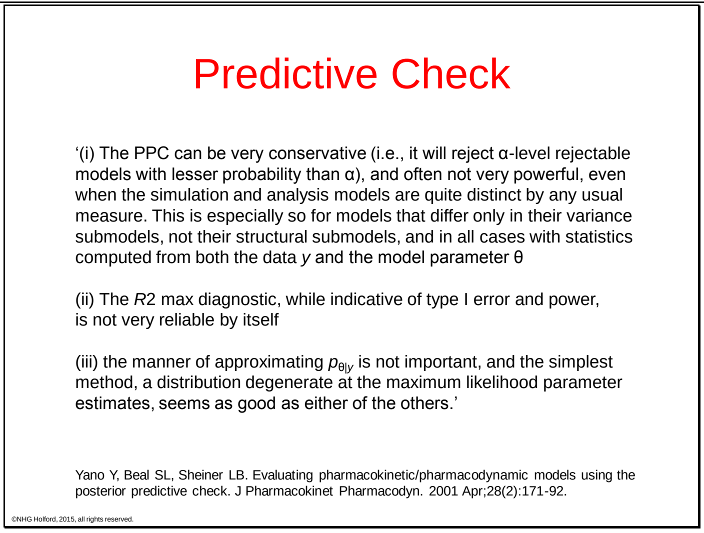
Sufficient and Non-Sufficient Statistics

‘A sufficient statistic (for the moment the term “statistic” is used in its usual sense of a scalar or vector function of the data alone) is a reduction of the data that involves no loss of information about the model parameter.

Sufficient statistics are, by definition, ones that *would* “automatically be well fit,” as they completely determine the fit. Nonsufficient statistics are therefore appropriate candidates for detecting model inadequacies’

Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001 Apr;28(2):171-92.

Slide 29



Predictive Check

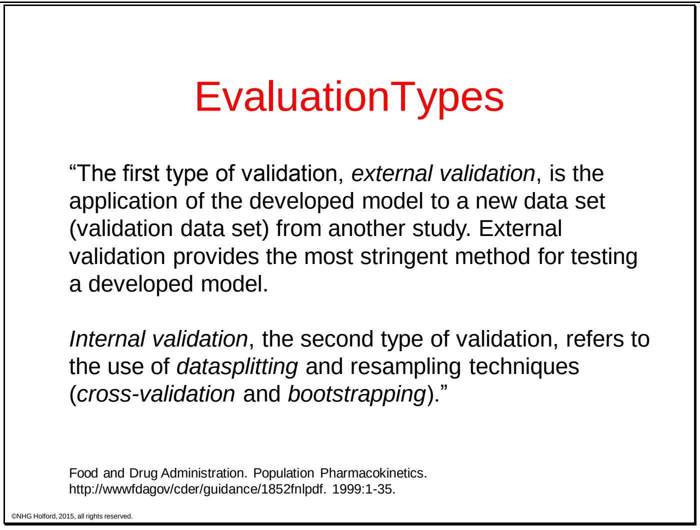
‘(i) The PPC can be very conservative (i.e., it will reject α-level rejectable models with lesser probability than α), and often not very powerful, even when the simulation and analysis models are quite distinct by any usual measure. This is especially so for models that differ only in their variance submodels, not their structural submodels, and in all cases with statistics computed from both the data *y* and the model parameter θ

(ii) The *R*2 max diagnostic, while indicative of type I error and power, is not very reliable by itself

(iii) the manner of approximating $p_{\theta|y}$ is not important, and the simplest method, a distribution degenerate at the maximum likelihood parameter estimates, seems as good as either of the others.’

Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001 Apr;28(2):171-92.

Slide 30



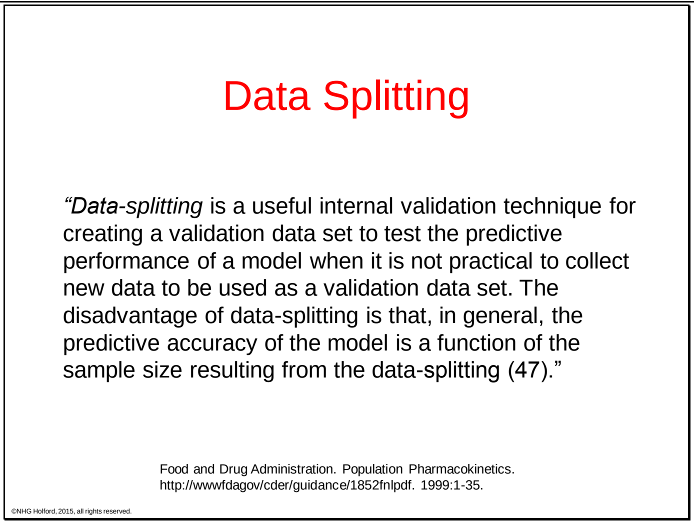
EvaluationTypes

“The first type of validation, *external validation*, is the application of the developed model to a new data set (validation data set) from another study. External validation provides the most stringent method for testing a developed model.

Internal validation, the second type of validation, refers to the use of *datasplitting* and resampling techniques (*cross-validation* and *bootstrapping*).”

Food and Drug Administration. Population Pharmacokinetics. http://wwwfdagov/cder/guidance/1852fnlpdf. 1999:1-35.

Slide 31

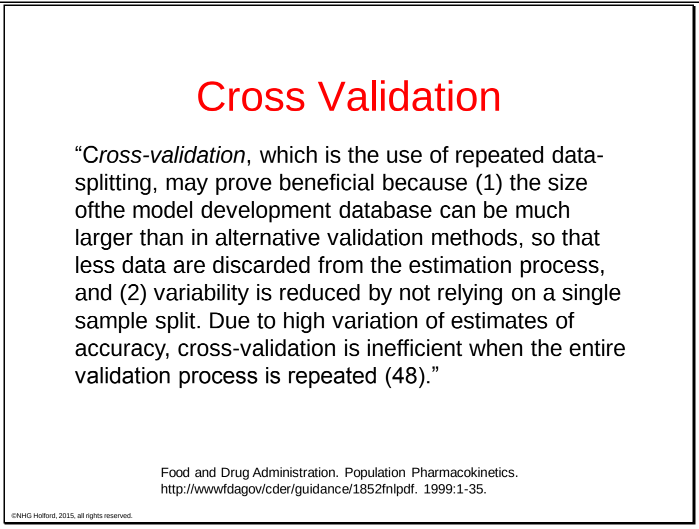


Data Splitting

“Data-splitting is a useful internal validation technique for creating a validation data set to test the predictive performance of a model when it is not practical to collect new data to be used as a validation data set. The disadvantage of data-splitting is that, in general, the predictive accuracy of the model is a function of the sample size resulting from the data-splitting (47).”

Food and Drug Administration. Population Pharmacokinetics. http://wwwfdagov/cder/guidance/1852fnlpdf. 1999:1-35.

Slide 32

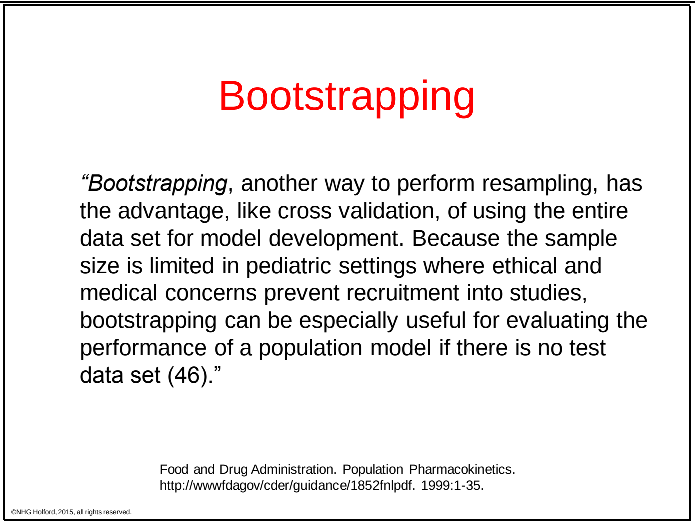


Cross Validation

“C*ross-validation*, which is the use of repeated data-splitting, may prove beneficial because (1) the size ofthe model development database can be much larger than in alternative validation methods, so that less data are discarded from the estimation process, and (2) variability is reduced by not relying on a single sample split. Due to high variation of estimates of accuracy, cross-validation is inefficient when the entire validation process is repeated (48).”

Food and Drug Administration. Population Pharmacokinetics. http://wwwfdagov/cder/guidance/1852fnlpdf. 1999:1-35.

Slide 33



Bootstrapping

“Bootstrapping, another way to perform resampling, has the advantage, like cross validation, of using the entire data set for model development. Because the sample size is limited in pediatric settings where ethical and medical concerns prevent recruitment into studies, bootstrapping can be especially useful for evaluating the performance of a population model if there is no test data set (46).”

Food and Drug Administration. Population Pharmacokinetics. http://wwwfdagov/cder/guidance/1852fnlpdf. 1999:1-35.