Selecting the right model across modelling platforms - a PopPKPD perspective.

Stephen Duffull
University of Queensland

Aim

To describe model selection techniques during model development

This talk does not cover model evaluation that may be performed after model development

Platforms considered in this talk

- Parametric maximum likelihood (e.g. NONMEM)
- Non-parametric maximum likelihood (e.g. NPEM)
- Markov chain Monte Carlo (e.g. WinBUGS)

Model appropriateness

- All models are wrong – but some are useful

  [GEP Box, 1979]

- Do the deficiencies in the model have a noticeable effect on its substantive inferences?

  [A Gelman, 1995]

- Checking the appropriateness of a model therefore requires the purpose for which the model was developed to be known *a priori*

Global / Local?

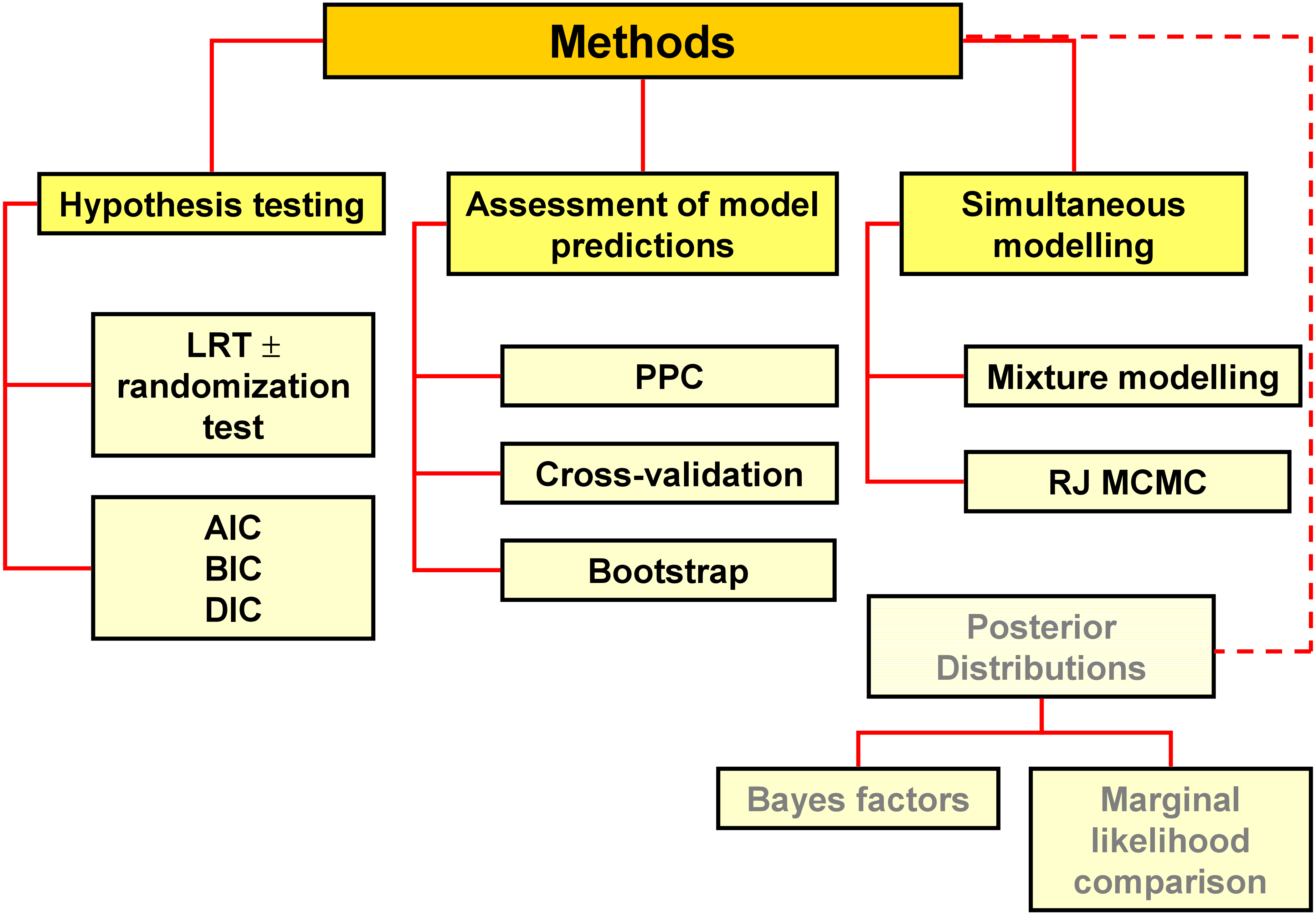
- Global
  - Refers to methods that assess the global fit of the model to the data, without reference to any particular features of the model or data
    - sum of squares
    - Cross-validation
- Local
  - Refers to methods that assess local features of the model, e.g. how well does the model describe Cmax
    - PPC

Model selection during model building

- Structural Model e.g. PK model
  - Input model
  - Disposition model
- Statistical Model
  - Between subject variability
  - Between occasion variability
  - Correlations between parameters (covariance matrix)
  - Residual variability
- Addition of covariates

What constitutes a good model selection method?

1. Accuracy
   - The method will have appropriate statistical properties
2. Relevance
   - The method tests the relevant features of the model
3. Ease of use
   - You can perform the method in a real time setting without requiring excessive custom written code and preferably on-the-fly


Methods
Hypothesis testing
Assessment of model predictions
Simultaneous modelling
LRT ± randomization test
AIC
BIC
DIC
PPC
Cross-validation
Bootstrap
Mixture modelling
RJ MCMC
Posterior Distributions
Bayes factors
Marginal likelihood comparison

Hypothesis testing

- Parametric maximum likelihood
  - Likelihood ratio test (LRT)
    - ± randomization test
- Non-parametric maximum likelihood
  - Likelihood ratio test (LRT)
- MCMC
  - Deviance Information Criterion (DIC)

Assessing Goodness of Fit
Likelihood ratio test (LRT) for nested models

- For **full** (k parameters) **and reduced models** (k-r parameters) the difference in OBJF is approximately χ^2 distributed
- A model is a reduced model of the full model if one or more parameters (r) of the full model can be fixed (usually to 0) to exactly then match the reduced model
- Likelihood ratio test (LRT)

Degrees of Freedom	5% (p<0.05)	1% (p<0.01)	0.1% (p<0.001)
1	3.84	6.63	10.83
2	5.99	9.21	13.82

(Adapted from Lynn McFadyen, Model building and hypothesis testing, PAGANZA 2002)

Why χ^2?

- The LRT can be shown to be asymptotically χ^2 distributed for all likelihoods (normal, binomial, Poisson etc)
  - Must be nested
- For mixed effect models asymptotic requires that both n_patients $\rightarrow \infty$ and n_samples_i $\rightarrow \infty$ (for i = 1:n_patients)
  - When these asymptotes are not reached the LRT is said to be approximately χ^2 distributed

Problems with assumption of χ^2 under null hypothesis

- Wahlby, Jonsson and Karlsson, 2001

  Get better agreement with nominal and actual p values if:

  - Use FOCE for additive residual error including log transformed both sides
  - or FOCE with Interaction (FOCEI) for proportional or slope intercept residual error models
- Gobburu, Lawrence, 2002
  - Can use FO method for sparse data but it is NOT better than FOCE or FOCEI

Example:
Weight on central volume

- M1
  - TVV1=THETA(2)
- M2
  - TVV1=THETA(2)*WT/70 + THETA(3)
- For both M1 & M2
  - V1=TVV1*EXP(ETA(2))

The LRT

- The data were modelled with the FO method (FOCEI had convergence problems)
- It is known that the LRT can be inaccurate with the FO NONMEM method
- The ΔOBJF = 18.57 from 2 runs in NONMEM (under FO)

- 1000 data sets were created and analysed using NONMEM where weight was permuted amongst the individuals (n=806 runs successful)
- The P-value can be computed as the number of runs that provide a more extreme difference than 18.57

Randomization test results

	REP#	ΔOBJF	CHI2	QTLE	OBJF	TERM MSG
39	475	19.66	9.0E-6	0.048	3118.50	MINIMIZATION_SUCCESSFUL_
40	**614**	**19.62**	**9.0E-6**	**0.050**	**3118.54**	**MINIMIZATION_SUCCESSFUL_**
41	870	19.53	1.0E-5	0.051	3118.63	MINIMIZATION_SUCCESSFUL_

Conclusions about WT on V1

- The addition of WT to V1 was not statistically significant (p=0.057)
- However, sufficient scientific evidence warrants its inclusion in any case

Non-parametric use of LRT

- Current publications using NPEM have used the LRT as you would do for the parametric case
- A single publication using NPML used the LRT but set the dof = ΔP^*N
- Problems
  - Are non-parametric likelihoods χ^2 distributed?
  - How do you determine the dof?

Bayesian Information Criterion (BIC)

$$BIC = \log P(\mathbf{Y} \mid \hat{\boldsymbol{\theta}}, M_x) - \frac{p}{2}\log(n)$$

- The difficulty with implementing this criterion for hierarchical models is that the true dimensionality (*p*) is not known
- How many parameters are influential in a hierarchical population model?
  - population parameters
  - residual variance parameters
  - *n* x *p* sets of individual parameters
- 1 cpt model with 100 patients = 311
- 2 cpt model with 100 patients = 522

Congdon. Bayesian Statistical Modelling. John Wiley & sons Ltd, New York 2003

Deviance Information Criterion (DIC)

- The DIC is computed as:

$$DIC = \overline{D(\boldsymbol{\theta})} + p.eff$$

- Where *p.eff* is the "apparent" number of parameters that enters the model
  - WinBUGS 1.4 provides the DIC value
  - WinBUGS 1.3 requires some coding to modify the model and then run 4001 iterations to produce 1 evaluation of the model at the mean parameter values to compute

Spiegelhalter et al. JRSS 2002;64:583-639

Assessment of model predictions

- Parametric maximum likelihood
  - Bootstrap
  - Predictive distribution check (not automatic)
  - Cross-validation (not automatic)
- Non-parametric maximum likelihood
  - Predictive distribution check (not automatic)
  - Cross-validation (not automatic)
- MCMC
  - Posterior predictive distribution check (PPC)
  - Cross-validation (not automatic)

Posterior predictive check

- Does the observed data look plausible under the posterior distribution?
- The replicated data ($\mathbf{y}^{rep}$) generated under the model (M_1 or M_2) should look similar to the observed data ($\mathbf{y}$)
- The test statistic may be
  - an observation (e.g. Cmax) $T(y)$
  - a joint function of the observations and model (e.g. ME or MSE) $T(y, \theta)$

Gelman et al. Bayesian Data Analysis, Chapman & Hall 1995

PPC (*P*–values)

- A *P*-value is computed by summing up the number of times that a prediction is more extreme than an observation.
- If the observation is, e.g. the median Cmax, then a good model should produce as many more extreme values of Cmax as less extreme values
- The *P*-value from this example should be close to 0.5
- A *P*-value of > 0.9 or < 0.1 might indicate poor model performance.

Example – Enoxaparin

- Enoxaparin is a low molecular weight heparin used in the treatment of
  - acute coronary syndromes,
  - pulmonary embolism
  - deep vein thrombosis
- Its use is characterised by a reduction in the complications arising from these conditions – but at a risk of increasing the risk of bleeding if the dose is not selected appropriately

Prior PKPD

- Previous data supported a strong relationship between Cmax and the risk of bleeding and a significant but weaker relationship between Cmin and therapeutic failure

Enoxaparin - PPC

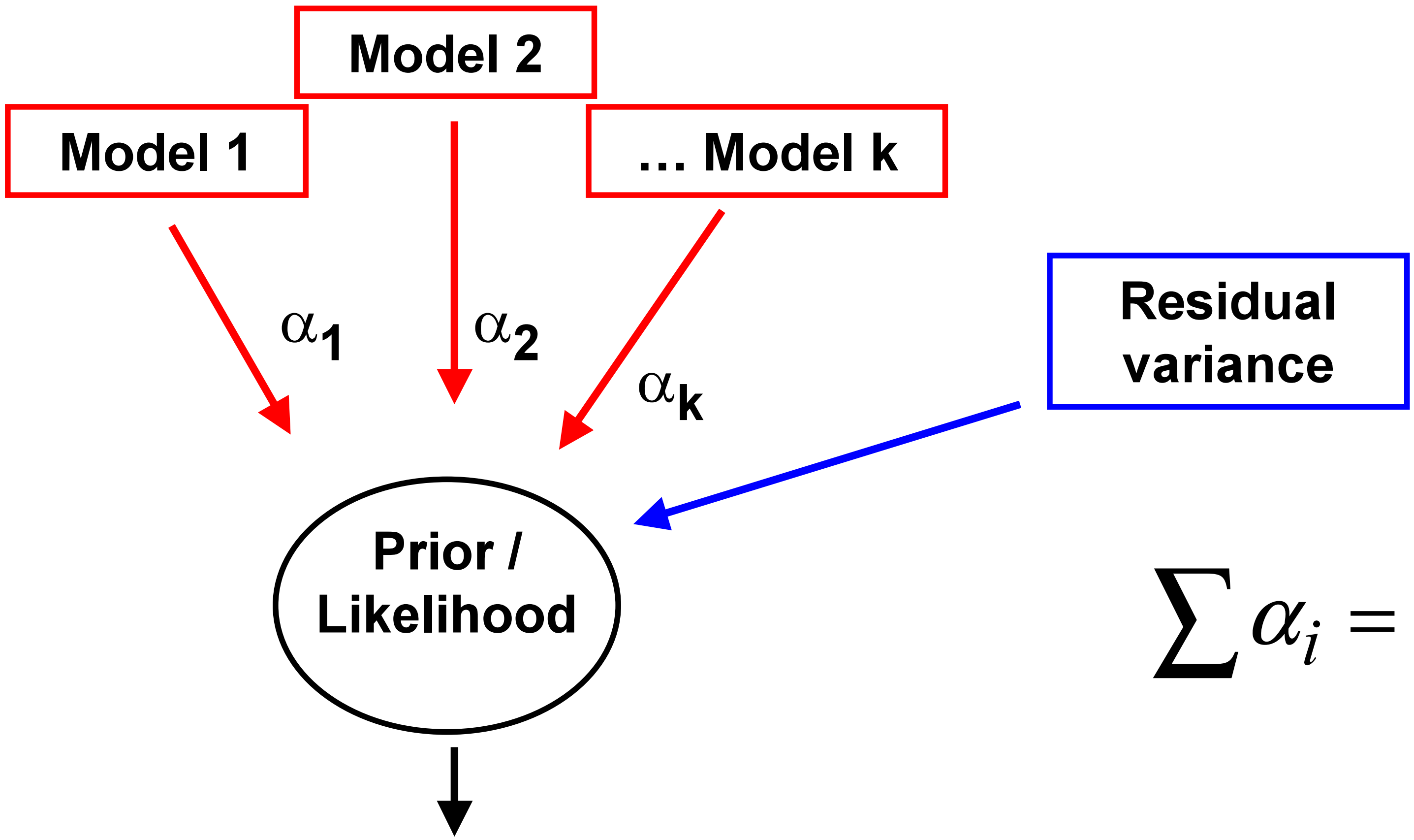
- $T(y)$ = {Cmax, Cmin}
- Cmax
  - $P(M_2) = 0.243$
  - $P(M_1) = 0.001$
- Cmin
  - $P(M_2) = 0.81$
  - $P(M_1) = 0.76$

- An hypothesis test (DIC) was unable to show a difference between the descriptive performance of the models
- PPC was performed during the process of generating the posterior distribution

Simultaneous modelling

- Parametric maximum likelihood
  - N/A
- Non-parametric maximum likelihood
  - N/A
- MCMC
  - Mixture modelling
  - Reversible Jump MCMC (cannot do automatically in BUGS)

Mixture modelling



$$\sum \alpha_i = 1$$

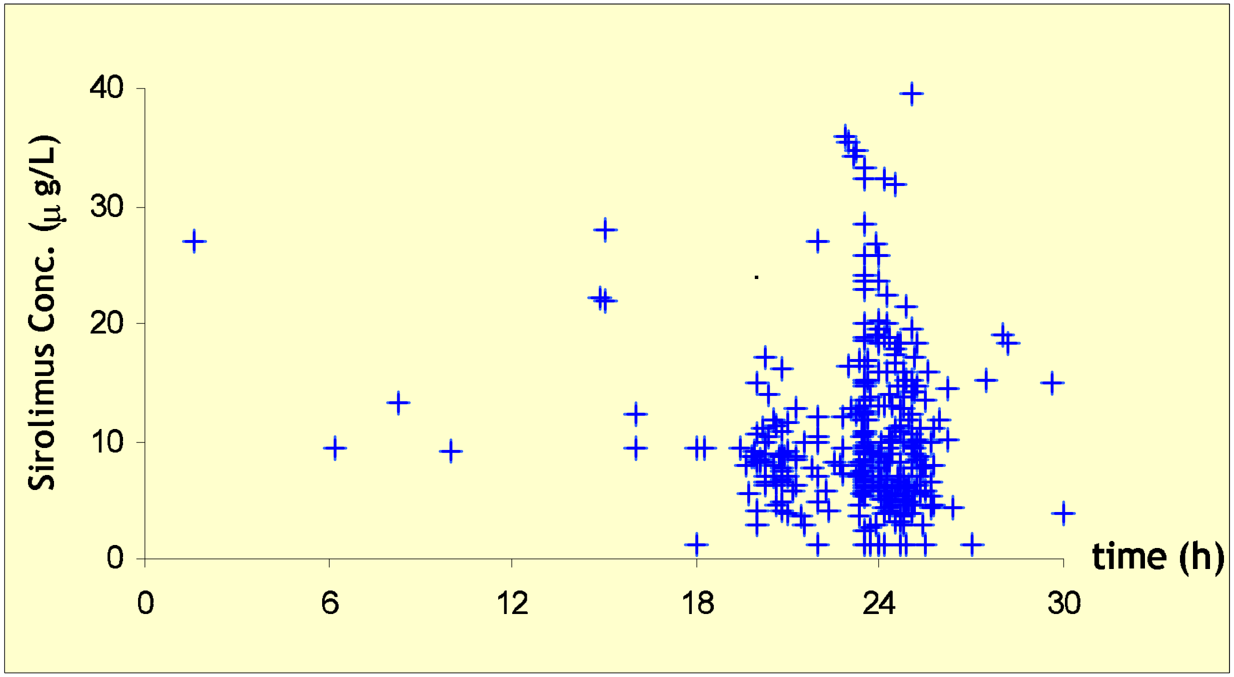
Posterior distribution (α)

The mixing parameter (α)

- The mixing parameter provides the support for the models in question
- In a simple case of 2 competing models the value α and (1- α) provide the weight for each model
- The mean of the posterior distribution of α provides the probability that one model is preferred over another
  - this can be shown as an odds
- The method can be performed on-the-fly

Example - sirolimus analysis

One or two compartment model?



- 315 observations
- 25 patients
- routine clinical care
- clustered at about 24 hours

The outcome

- The posterior odds for preferring a 2-compartment model was 8.1
- The DIC gave weak support for this decision

Checking the mixing parameter

- Data from the same design was simulated and fitted using MCMC with informative and non-informative priors 10 times for each prior under each model
- The mixing parameter supported the correct model 1 or 2 compartment (with mean odds ranging from 13 to 23) on each occasion
- The use of informative priors improved the ability to discriminate between models

Hypothesis tests

	Accuracy	Relevance	Ease of use
"NONMEM"	?	✓	✓
"NPEM"	?	X	✓
"BUGS"	?	✓	✓

Predictive distributions

	Accuracy	Relevance	Ease of use
"NONMEM"	✓	✓ / X	X
"NPEM"	✓	✓	✓
"BUGS"	✓	✓	✓

Simultaneous modelling

	Accuracy	Relevance	Ease of use
"NONMEM"	?	✓/?	X
"NPEM"	?	?/X	X
"BUGS"	✓	✓	✓

Conclusions

- Different purposes for model use affect the relevance of the model selection procedure
- Different platforms for model building affect the accuracy, relevance and ease of use of some procedures

- Generally simulation platforms and non-parametric methods “perform well” when assessing predictive distributions
- Parametric maximum likelihood methods can be linked easily with standard posthoc procedures such as randomization tests, bootstrap etc