Automation of Structural Pharmacokinetic Model Search in NONMEM: Evaluation with Preclinical Datasets

Jeroen Schaap [1], Stefan Verhoeven [2], Gerard Vogel [3], Martijn Rooseboom [3,4] and Rene van Schaik [2]

(1) PK-PD/M&S, Clinical Pharmacology and Kinetics, Organon N.V., The Netherlands; (2) Molecular Design and Informatics, Pharmacology Oss, Organon N.V., The Netherlands; (3) DMPK & Safety, Dept. Pharmacology Oss, Organon N.V., The Netherlands; (4) Dept. Toxicology and Drug Disposition Oss, Organon N.V., The Netherlands.

Objective

- **Implement an automated structural PK model search with NONMEM**

Background

- Availability of modelers limits application of PK-PD
- At the same time, computational power of hardware is increasing

=> Automation of model development only logical!

- A patented hybrid genetic algorithm exists for PK-PD modeling [1]
  - more or less black-box approach
  - preference for model search that more closely resembles human method
- Staged approach
  - supervised selection: after each stage, human approval
- First stage:
  - screen structural PK models only
  - use simple absorption models only
  - application: preclinical PK datasets
  - test with database of preclinical datasets

Software

- NONMEM V
- ifort 9.1 or g77
- PsN v2.1.10
- Sun Grid Engine
- Perl scripts
  - data management
  - execution logic
- R scripts
  - generate plots
  - derived parameters
- MySQL with phpMyAdmin
  - result collection & analysis

Model space

- Distribution and elemination
  - 1, 2 or 3-comparmental
- Absorption
  - none, zero order, first order or combination
  - with or without lagtime
- Proportional error model
- 24 models in total
- Expected to handle majority of datasets
  - certainly not all
  - e.g. complex absorption profiles
  - e.g. non-linear pharmacokinetics
  - e.g. metabolite profiles
  - e.g. problematic bioanalysis
  - ....

Initial parameters and reruns

- How to avoid local minima?
- How to increase succesful estimation?

- **Solution:**

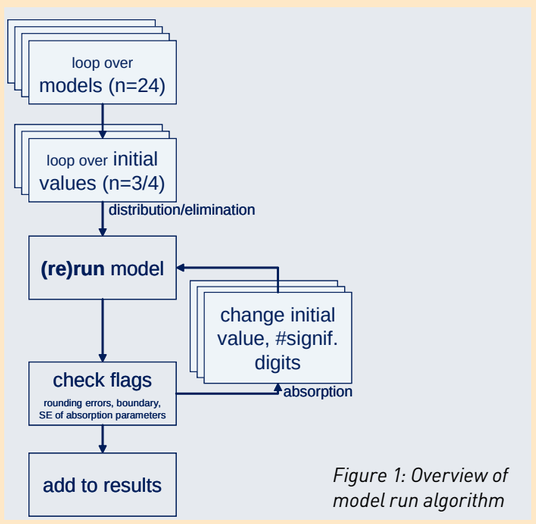


Figure 1: Overview of model run algorithm

  - specify parameters lognormal
  - bioavailability: logistic
  - use 3-4 different sets of initial parameters
  - relative position most important
  - only for distribution and elimination
  - in total 88 runs per dataset
  - absorption:
  - loop over initial sets not feasible (4752 runs per dataset)
  - rules developed for screening per parameter
  - check for rounding errors, parameter near boundary or large SE associated with parameter
  - rerun with different initial value if necessary
  - increase SIGDIG upon mild rounding errors
  - rerun model if NONMEM termination had rounding errors up to one lower than the number requested

Dataset

Current stage: simple & clean

- 56 preclinical PK experiments
  - i.v. + other route(s) (po, sc, ...)
  - manually selected
  - range from easy to impossible
  - see where approach fails
- Combine any non-i.v. with i.v.
  - => 73 datasets

Model selection

- Perfect selection statistic?
  - does not exist!
- Bootstrap-derived methods?
  - computational burden prohibitive
- Keep it practical
  - hierarchy of termination
  - with covariance step
  - successful
  - iteration terminated
  - Akaike information criterium (AIC) as main selection criterium
  - tolerate deviation of 2 from lowest AIC
  - within tolerated, select lowest residual error and than fewest parameters

Overview model runs

- On average 197 model runs per dataset
  - 12141 normal and 15059 reruns (rerun: initial parameter changed)
  - much less than the 4752 runs per dataset needed for a full initial value screen
  - 138 datasets with results over 2 platforms
  - clearly more model runs than an experienced modeler would execute
- 0.88 % of reruns were selected as optimal model
  - 0.11 % of models without reruns were selected
  - 1.24 rerun per modelfit
  - 38 % of models without reruns
  - in only 1 occasion a rerun of lagtime was selected
- the 3-compartmental distribution submodel was selected most frequently (n=83)
- the first order absorption submodel was selected most frequenly (n=65)
- lagtime submodels were selected 18 times

Relative acceptance[1] (%) of reruns per parameter and platform

Parameter	Reason for rerun	Itanium	Xeon	average
Lagtime	large SE	0.4	2.5	1.5
	parameter rounding error	0.6	1.6	1.1
	parameter near boundary	2.1	4.1	3.1
	average	1.0	2.8	1.9
Zero-order duration	large SE	32.1	33.2	32.6
	parameter rounding error	68.5	67.3	67.9
	parameter near boundary	49.9	48.7	49.3
	average	50.1	49.7	49.9
Fbio	large SE	29.5	32.0	30.7
	parameter rounding error	64.1	65.4	64.8
	parameter near boundary	48.7	47.5	48.1
	average	47.4	48.3	47.9
ka	large SE	27.0	26.9	27.0
	parameter rounding error	62.6	69.6	66.1
	parameter near boundary	42.4	40.7	41.5
	average	44.0	45.7	44.9
SIG	"2 <# significant digits < 3"	52.0	56.9	54.4
all	**average**	**38.9**	**40.7**	**39.8**

1) acceptance: rerun resulted in model improvement (OFV, termination status) compared to original

Evaluation

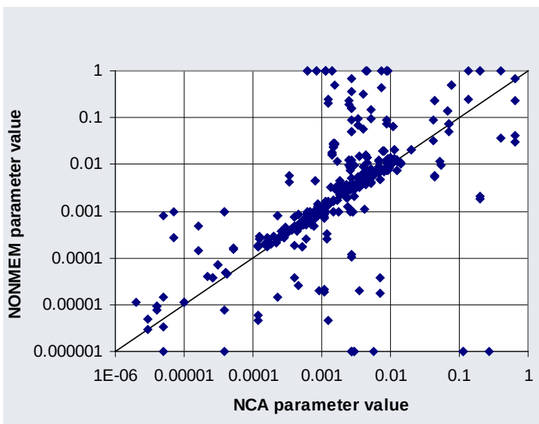
- **No real external reference available!**

1. Compare between platforms (Itanium and Xeon)
2. Compare with noncompartmental analysis (NCA) estimates
   - Estimates obtained with WinNonlin: terminal half-life (t1/2), clearance (CL), volume of distribution at steady-state (Vss) and absolute bioavailablity (Fbio)
3. Manual check on each dataset screen (subjective) (Itanium platform)
   - diagnostic plots OK?
   - model selection OK?

Comparison between platforms

1. n=8 datasets: technicalities
   - time-outs caused by extremy long or hanging NONMEM runs
2. n=29 exactly same objective function value (OFV)
3. n=23 OFV: difference < 3.7
   - n=11 apparent rounding errors as difference was smaller than 0.1
4. n=6 OFV: 3.7 < difference < 10
5. n=7 OFV: difference >= 10
   - 6/7 clear problems in GOF and datasets => easy to detect with manual check

Comparison with NCA parameters



Summary ratio NCA/NONMEM

Parameter	Median
CL	0.91
Fbio	0.93
t1/2	0.78
Vss	0.68
all	**0.87**

Figure 2: Comparison of parameters as calculated by NCA and the automatic NONMEM search method. Individual parameter values were plotted against each other after normalisation, and in the case of NONMEM parameters, truncation at a 10^6-fold range.

Fraction of NONMEM-parameters that differ 2- fold or more with NCA parameters

	Hardware platform		
Parameter	**Xeon**	**Itanium**	**average**
CL	0.27	0.29	0.28
Fbio	0.26	0.25	0.25
t1/2	0.41	0.39	0.40
Vss	0.41	0.40	0.41

Occurence of classified[1] reasons for more than 2-fold differences with NCA outcomes [1]

	Difference in objective function between Itanium and Xeon platforms						
Reason	**==**	**< 0.1**	**< 3.8**	**< 10**	**>= 10**	**n.a.[2]**	**all**
Fbio > 1		1					1
High IIV i.v. route				1			1
Low absorption rate	1		1				2
NM results preferable	2		3				5
Noisy profiles i.v. route		1	1				2
Outlying point		1			2		3
Sampling terminal phase	1	1		1		1	4
Scale, scintilation counts	3	2			2		7
Unknown[3]		1		1			2
Variable data p.o. route	1	1	1			1	4
all	**8**	**8**	**6**	**3**	**4**	**2**	**31**

1) classification based upon manual check of model (multiple models) diagnostics and observed versus time plots; evaluation per experiment rather than dataset

2) n.a.: not available as model for one platform did not result in estimatable parameters

3) unknown: reason for difference with NCA was not obvious from the evaluation under 1)

Occurence of classified[1] reasons for more than 2-fold differences with NCA outcomes [2]

	Difference in objective function between Itanium and Xeon platforms					
	==	**< 0.1**	**< 3.8**	**< 10**	**>= 10**	**all**
NCA <= 2-fold different	17	2	2	2	1	24
NCA > 2-fold different	8	8	6	3	4	29
all	**25**	**10**	**8**	**5**	**5**	**53**

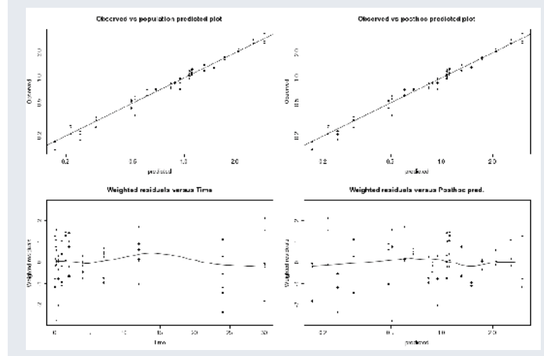
1) classification based upon manual check of model (multiple models) diagnostics and observed versus time plots; evaluation per experiment rather than dataset

Example of diagnostic plots, case 1:

Difference between itanium and xeon platforms without clear reason

run on Xeon platform

zero- & first-order; OFV: -6.5



run on Itanium platform

first-order

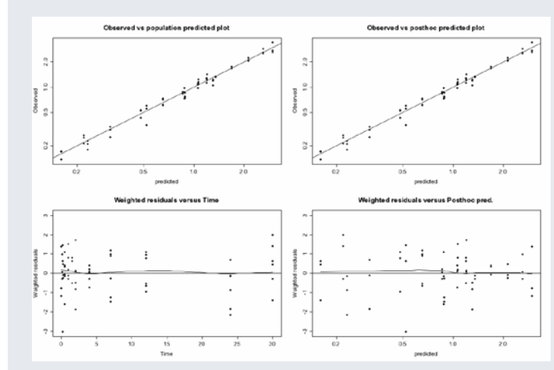


Figure 3: Example of goodness-of-fit plots of a dataset that yielded undoubtedly different results between platforms, although no clear problems appeared in the observed or WRES against predicted or time plots.

Example of diagnostic plots, case 2:

Difference between NONMEM and NCA estimates without clear reason

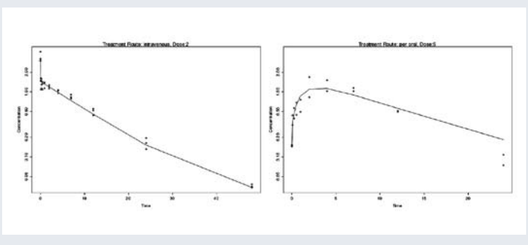


Figure 4: Observed and predicted against time plot of an example dataset that yielded clear differences between NCA and NONMEM parameter estimates without any obvious large problems in the fit.

Example of diagnostic plots, case 3:

Effect of outlier on model-predicted PK profile

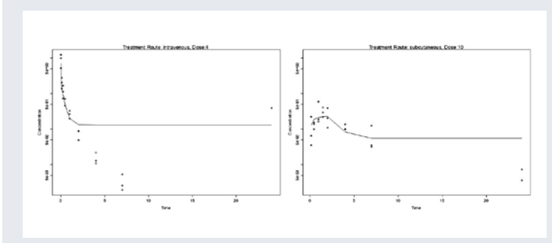


Figure 5: Observed and predicted against time plot of an example dataset with an outlier that resulted in a problematic modelfit.

Manual checks

- Only 16 out of 67 selected models had trend-free diagnostic plots
  - diagnostic plot classification did not correlate (eye-ball) with NCA/NONMEM or platform difference
- All model selections were jugded to be defendable
  - i.e. no other screened model provided improvement compared to selected model, including models without successful termination

Key findings

- Model selection method (nested sorting, AIC with tolerance) performs outstanding (see also [2])
- 25 out 67 model selections resulted in differences (larger than attributable to rounding errors) between two hardware platforms
- NCA and NONMEM parameter estimates were mildly biased relative to each other (NCA estimates 10% lower) (see also [3])
- Differences between NCA and NONMEM estimates occurred more often with Vss and t1/2; the NONMEM method frequently resulted in very high parameters values
- Model results are rather sensitive for scaling, i.e. observations and dosages should lie within a limited range
- Only mild correlation (eye-ball) between NCA/NONMEM estimate differences and platform differences

Conclusions

- **Automaton of structural model search seems feasible**
  - **runtime not prohibitive for preclinical datasets (~1 day of processor time for the complete database discussed)**
- **Human supervision remains necessary**
  - **many types of encountered model problems hard to detect automatically**

Discussion

- Trend finder for model selection worthwhile?
  - run number too global
- An external reference is hard to find
  - NCA estimates not perfect
  - simulations not reflective of real life problems
- Concept also seems useful as model building start
  - staged approach allows seemless integration with manual modeling
- Staged approach if variable non-i.v. data?
  - screen & select model i.v. data
  - fix i.v.-thetas and fit non-i.v. data

Next steps

- Screen for additive error
- Combine >2 routes
- Primary covariate screen
- Mixed effect screen
- Sequential PK-PD modeling

References

[1] R. R. Bies, M. F. Muldoon, B. G. Pollock, S. Manuck, G. Smith, and M. E. Sale. A genetic algorithm-based, hybrid machine learning approach to model selection. J.Pharmacokinet.Pharmacodyn. 33 (2):195-221, 2006.

[2] T. M. Ludden, S. L. Beal, and L. B. Sheiner. Comparison of the Akaike Information Criterion, the Schwarz criterion and the F test as guides to model selection. J Pharmacokinet.Biopharm. 22 (5):431-445, 1994.

[3] J. P. Hing, S. G. Woolfrey, D. Greenslade, and P. M. Wright. Is mixed effects modeling or naive pooled data analysis preferred for the interpretation of single sample per subject toxicokinetic data? J Pharmacokinet.Pharmacodyn. 28 (2):193-210, 2001.

Organon

0721010