Slide 1	# Model Evaluation **Guangda Ma** **Auckland Pharmacometrics Group** Department of Pharmacology & Clinical Pharmacology The University of Auckland MEDSCI 719 2020	
Slide 2	## Recommended Reading • Model Evaluation of Continuous Data Pharmacometric Models: Metrics and Graphics. *CPT Pharmacometrics Syst Pharmacol.* 2017 Feb; 6(2): 87–109. 	The figures used in this presentation are based upon this paper.

Slide 3	## Model Evaluation
	• In model building we fit a models to a dataset.
	• In model evaluation we examine:
	  • Goodness-of-fit between the model and dataset
	  • The appropriateness of the underlying model assumptions
	• Numerical Diagnostics
	  • Fit statistics
	  • Parameter estimates & Imprecision estimates
	• Graphical Diagnostics
	  • Prediction based
	  • Residual based
	  • Simulation based (visual predictive checks)
	
		Once a model is built we want to assess how good the model is, or compare one model to another.
		
		Approaches to model diagnostics can be classed as numerical or graphical.

Slide 4

Fit Statistics

- Objective Function Value
  - When fitting the model to the data, many software packages minimise the -2 log likelihood (-2LL). This is called the objective function value in NONMEM.
  - The likelihood describes the likelihood of the observations being observed given the current parameters and model.

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(Y_i - \widehat{Y_i})^2}{2\sigma_i^2}}$$

  - The OFV (-2LL) provides an overall summary of how a model with a given set of parameter values fit the data.
  - The model with the lowest OFV is of best fit.
- Information Criteria (AIC & BIC)

Objective functions are statistical criterions applied to nonlinear regression models as an objective measure of the differences between the observed and predicted values of parameters and the dependant variable. The objective function minimized in NONMEM is the -2 log likelihood.

The likelihood (L) describes the likelihood of all the observations under the current model, structural and variance parameters. There are two parts to this equation. First is the likelihood of an observation, here we describe this as the i^{th} observation; this is related to the observed value (Y_i), the model predicted value ($\hat{Y}_i$) and the variance of the model (σ_i). Second, the likelihood n observations is the product of the individual observations, thus we multiply the probability of the first (i=1) to the n^{th} (last) observation.

Rather than use the likelihood, which requires multiplication of n probabilities, we may take the log of both sides of the equation, as well as multiply this by -2. This results in the -2 log likelihood. When fitting the model to the data, we wish to find the structural model, and parameter values which minimise the -2LL, thus maximises the likelihood.

Note that OFV is dependent on the method of parameter estimation and the data set, thus should not be used for comparison across data sets.

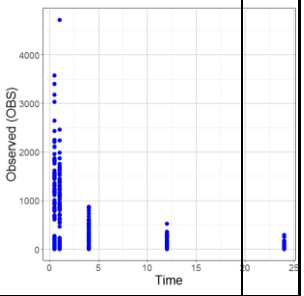
Increasing the number of parameters in a model increases the degrees of freedom and can artificially inflate goodness of fit. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to rank the goodness of fit of models taking into account improved model fit due to increased model complexity.

See Mould & Upton. CPT Pharmacometrics Syst Pharmacol. 2013 Apr; 2(4): e38.

Slide 5

Graphical Diagnostics
PK Dataset

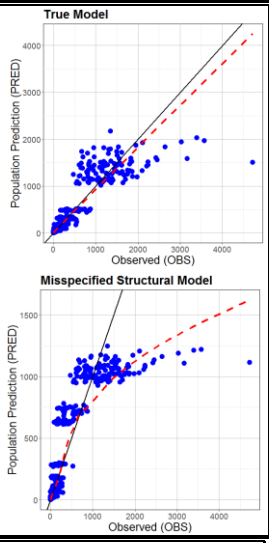
- Graphical prediction and residual based diagnostic tools is illustrated using a PK dataset.
- This data generated by simulation under a two-compartment model, first-order elimination and single IV bolus input with combined additive & proportional error.
- The model used to simulate the data was fitted to the data (True Model)
- A one-compartment (rather than two compartment) model also fitted to the data (Misspecified Model)



Slide 6

Population Prediction Based Diagnostics

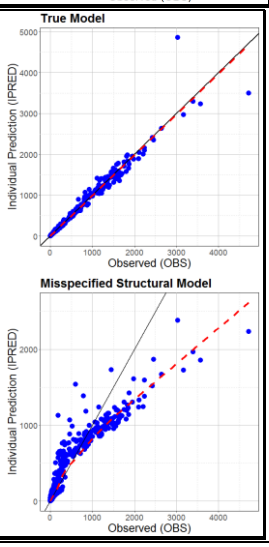
- The population prediction (PRED) assumes all random effects equal zero
- Under a correct model:
  - Data points are scattered about the line of identity (black)
  - The trend line (red dashed) lies close to the identity line
- Trends may suggest a misspecification of the structural, or the parameter variability model.

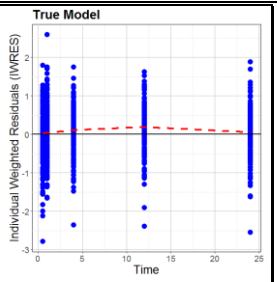
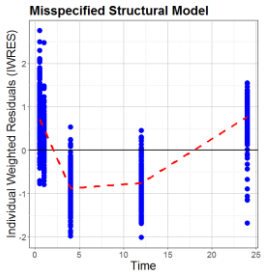


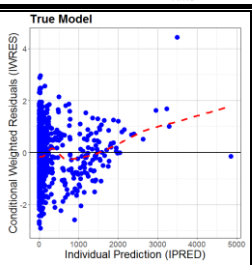
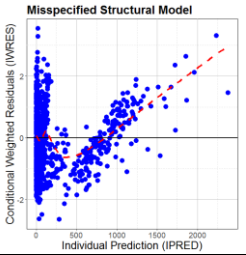
Slide 7

Individual Prediction Based Diagnostics

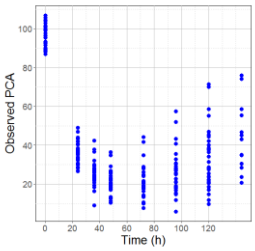
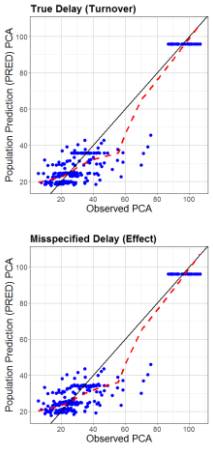
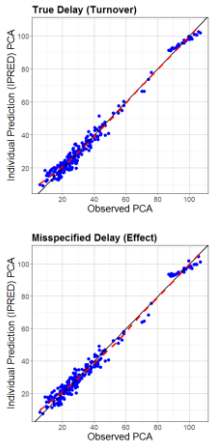
- The individual prediction (IPRED) is based upon the individual parameter estimates.
- IPRED v OBS allows evaluation of the individual fit for all patients
- Under a correct model:
  - Data points are scattered about the line of identity
  - Data points should scatter closer to the line of identity than with a plot of PRED v OBS, particularly when between-subject variability is large
- Trends may suggest a modification of the structural or residual error model.

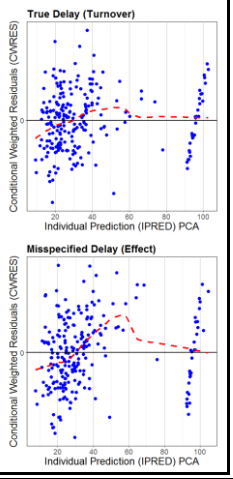
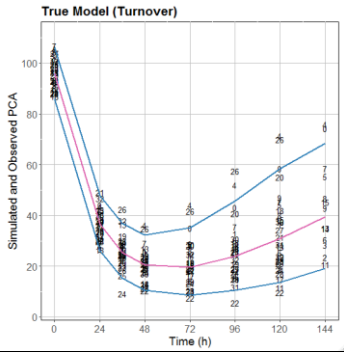


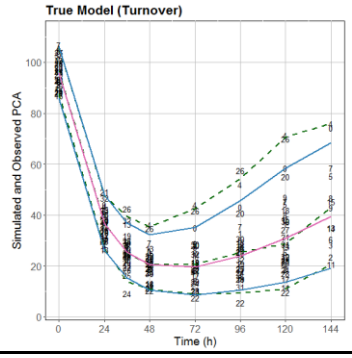
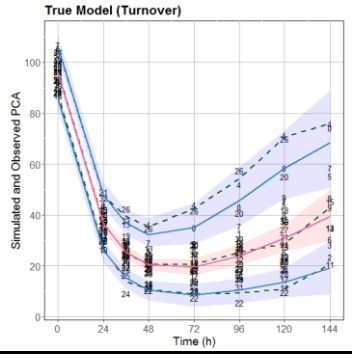
Slide 8	**Residual Based Diagnostics**
	• Weighted residuals (e.g. IWRES, CWRES) describe the weighted difference between the model prediction and data
	• Under a correct model:
	  • Data points are scattered symmetrically about the horizontal zero-line.
	  • Most points lie between -1.96 & 1.96
	• Trends may suggest a modification of the structural or residual error model.
	 
	 
	
		Weighted by the square root of the variance of the data given the model.

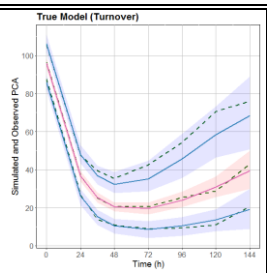
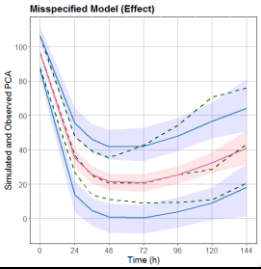
Slide 9	**Residual Based Diagnostics**
	• Weighted residuals (e.g. IWRES, CWRES) describe the weighted difference between the model prediction and data
	• Under a correct model:
	  • Data points are scattered about the horizontal zero-line (more or less evenly).
	  • Most points lie between -1.96 & 1.96
	• Trends may suggest a modification of the structural or residual error model.
	 
	 
	
		Residuals v individual prediction

Slide 10	**Simulation Based Tools**
	• Simulation based evaluation tools rely on the principle that if a model correctly describes the data, the data simulated under the model should match the observations.
	• Evaluation Tools include
	  • Visual Predictive Checks (VPC)
	  • Numerical Predictive Check (NPC)
	  • Prediction Discrepancies
	  • Normalised Prediction Distribution Errors (NPDE)
	10
	
		A drawback of using simulation for model evaluation is that it may not be able to fully capture all characteristics of the design under which the data was observed. For example subjective choices in behaviour (dose adherence or time of dose) or protocol violations may not be fully captured in simulation.
		
		Note Numerical Predictive Check, Prediction Discrepancies, and Normalised Prediction Distribution Errors are beyond the scope of this course.

Slide 11	Simulation Based Diagnostics **PD Dataset** • A simulated dataset is used to illustrate the visual predictive check. • Data generated by simulation under a PK model (one-compartment, first-order input & output) and turnover model (delayed PD) to describe the effect of plasma concentration on prothrombin complex activity (PCA). • The model used to simulate the data is fitted to the data (True Model) • An effect compartment (rather than turnover) model is also fitted to the data (Misspecified Model)  	
Slide 12	**Evaluation Requires Many Tools** • For the PD dataset, the prediction and residual based diagnostics appear visually similar for the true and misspecified models. • Plots do not show important disagreement between the data (observations) and model predictions.  	Though the effect compartment model is misspecified, misspecification is not apparent when graphical prediction and residual based diagnostics are examined.
Slide 13	**Evaluation Requires Many Tools** • For the PD dataset, the prediction and residual based diagnostics appear visually similar for the true and misspecified models. • Plots do not show important disagreement between the data (observations) and model predictions.  	Though the effect compartment model is misspecified, misspecification is not apparent when graphical prediction and residual based diagnostics are examined.

Slide 14	**Evaluation Requires Many Tools** • For the PD dataset, the prediction and residual based diagnostics appear visually similar for the true and misspecified models. • Plots do not show important disagreement between the data (observations) and model predictions.  	
Slide 15	**Visual Predictive Checks** • In a VPC the proposed model is used to simulate a number of samples (100-1000) under the design of the observed data. • Distribution statistics from the observations can be compared to the distribution statistics of predictions and their associated confidence interval. 15	Simulated predictions include fixed and random effects. (e.g between-subject, between occasion variability), whereas evaluation using population predictions (PRED) are based upon fixed effects (without random effects) and evaluation based upon individual predictions (IPRED) are based on parameter estimates (empirical Bayes estimates) which may be subject to shrinkage (perfect fit phenomenon). Holford NHG. The visual predictive check – superiority to standard diagnostic (Rorschach) plots www.page-meeting.org/?abstract=738. Last accessed 13 Feb 2019. PAGE. 2005;14.
Slide 16	**Scatter VPC** • The prediction intervals of the simulated data (blue and pink lines) can be plotted against the observations.  16	A scatter VPC plots the prediction intervals (e.g. 5th, 50th, 95th percentiles) over the observations. We expect 50% of the observations to lie above and below the median, 5% above the 95th percentile, and 5% below the 5th percentile. When there are few data points, this can be visually evaluated, however, when there is a large number of observations, it is difficult to visually compare the prediction intervals with the observed data.

Slide 17	**Percentile VPC** • In a percentile VPC the percentiles of the observed (green dashed) can be plotted with the percentiles of that simulated (blue & pink).  	In a percentile VPC, we can summarise the percentiles of the observed data and plot this with that simulated. This allows direct comparison between the percentiles of the observed and simulated (predicted).
Slide 18	**Confidence Interval VPC** • In a confidence interval VPC the 95% confidence interval associated with the percentiles are also shown.  18	A confidence interval VPC extends the percentile VPC and plots a 95% confidence band around the predicted prediction intervals. Addition of the confidence interval gives an indication of the uncertainty associated with the predictions. It allows appreciation of whether the differences between observed and simulated percentiles arise by chance. Generation of confidence interval VPCs requires the data to be binned, where the independent variable (e.g. time) is grouped into a bin such that each bin contains approximately equal numbers of observations. Thus the choice of bins can influence the interpretation of a VPC; often observations from a clinical source are collected *ad hoc* rather than in an orderly fashion.

Slide 19	**Evaluating a VPC**
	• Under a correct model:
	  • No systematic difference between observed and predicted percentiles
	  • Predicted percentiles lie within the corresponding confidence interval.
	• Trends may suggest misspecification in the structural, or the parameter variability model.
	 
	 
	
		The data for the VPC plots was generated using a PK model which described warfarin concentration using one-compartment model with first-order absorption, a lag-time, and first-order elimination, and PD model which used a turnover model to describe the relationship between concentration and the rate of prothrombin complex activity (PCA) production.
		
		The VPC using the true model fits the uses the same turnover model (that was used to simulate the observed data). The VPC using the misspecified model uses an effect compartment model to describe the relationship between concentration and PCA; this is misspecified as the wrong delay mechanism is used.

		Simulation may not be able to fully capture all characteristics of the design under which the data was observed. For example subjective choices in behaviour (dose adherence or time of dose) or protocol violations may not be fully captured in simulation.