Slide 1

Model Evaluation

Visual Predictive Checks

PAGE 2008 Marseille

Nick Holford
University of Auckland

Mats Karlsson
University of Uppsala

www.page-meeting.org/?abstract=1434

Slide 2

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

Outline

- What is a Visual Predictive Check?
- What choices are there in presentation?
- What can it help to show?
- What may it fail to show?

Slide 3

UPPSALA UNIVERSITET

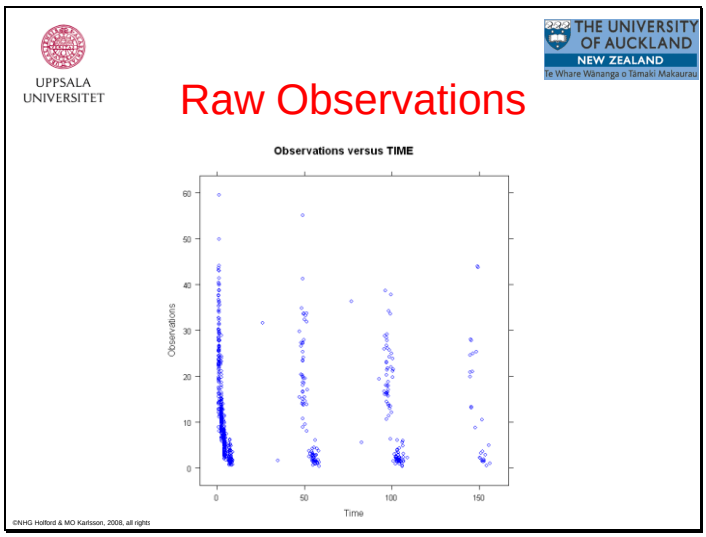
THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

What is a VPC?

- Comparison of Observations and Simulated Predictions (SPRED)
  - Simulated predictions include fixed and random between subject variability as well as residual error
  - They are different from 'population' predictions (PRED) (fixed effects without random effects) and individual predictions (IPRED)(shrinkage)
- VPC compares statistics derived from the distribution of observations and the distribution of SPRED at specific times or time intervals
  - E.g. median and 90% intervals at 1 h after the dose
  - Intervals can be joined together in time sequence to create bands (but most often the bands are called 'intervals')

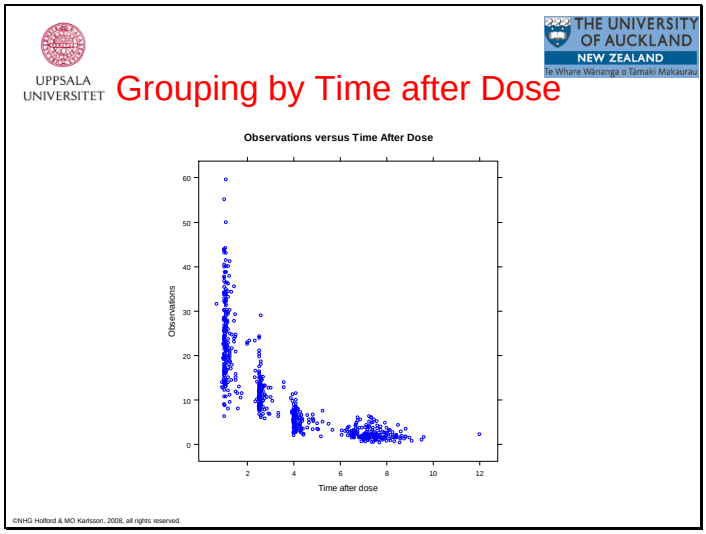
VPCs use a different kind of prediction compared with traditional diagnostic plots. They are based on simulations of model predictions including random effects (especially between subject variability (BSV)). Summary measures of the distribution of predictions and observations are compared visually. Typical summary measures are the median and an interval defined by the lower 5% and upper 5% of the values.

Slide 4



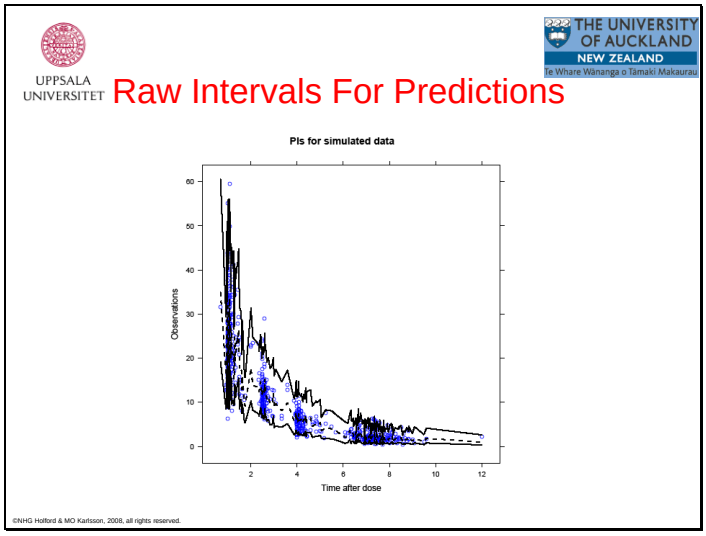
Data on the real time axis, but that may not be the most informative independent variable for a vpc.
[To look for long-term trends, we may stratify on occasion. – more on stratification later]

Slide 5



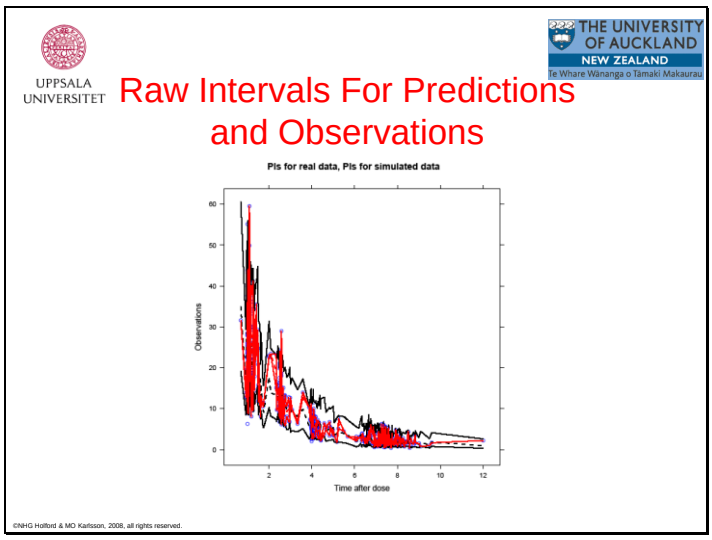
For now it may be better to start inspecting concentration versus time after dose.

Slide 6



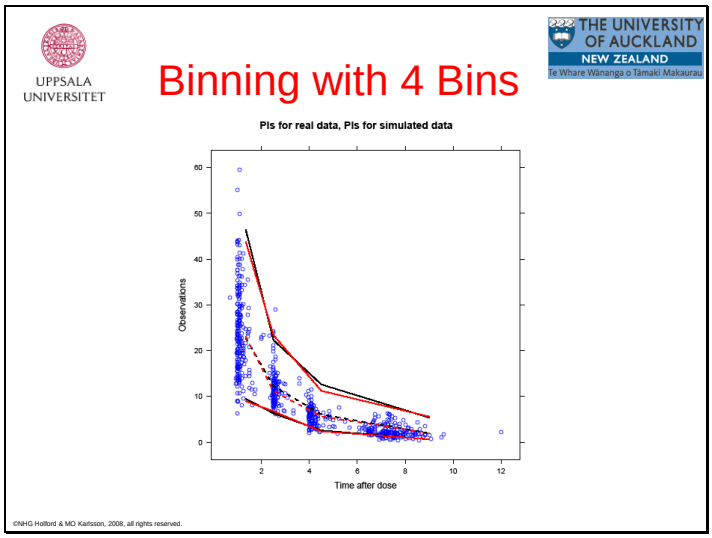
Without binning, the predicted intervals (PIs) based on simulated data will display erratic patterns, precluding or making difficult a suitable judgement.

Slide 7



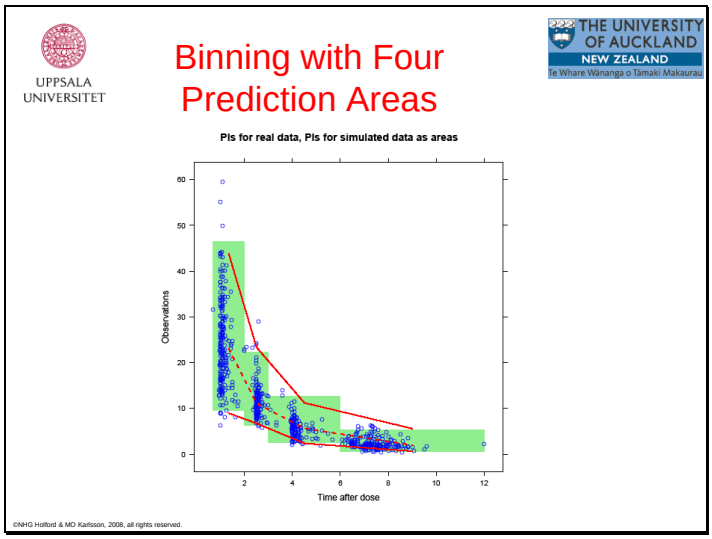
It doesn't help if PIs are added also for the real data (these are in red)

Slide 8



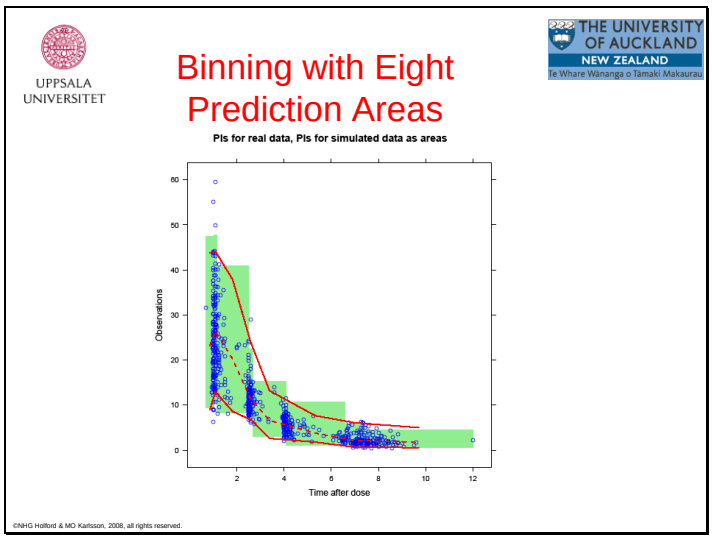
Instead we bin data. In this case in 4 bins (with boundaries between bins at 2, 3 and 6 hours). Displying both real (red) and simulated (black) data as PIs make the plot helpful in interpretation. However, how crucial is the binning strategy? When the PI at the midpoint of each time interval is connected with the next midpoint time, we actually don't display how we really are binnning and what the true PIs are over time.

Slide 9



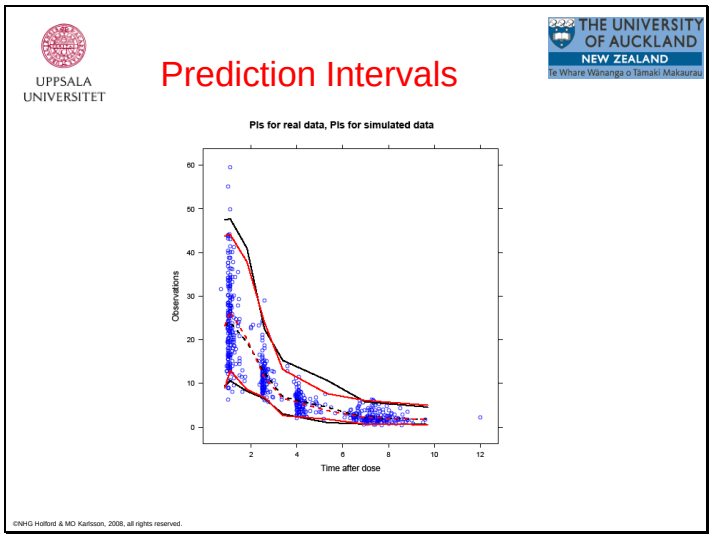
Showing PIs as areas (PIs of simulated data in green) is in some sense more "honest". The true PI for each binning interval is displayed (in this case only for simulated data).

Slide 10



This graph shows 8 binning intervals (binned to have similar amount of data in each bin), compared to the previous graph's 4 binning intervals. Does it matter how you bin?

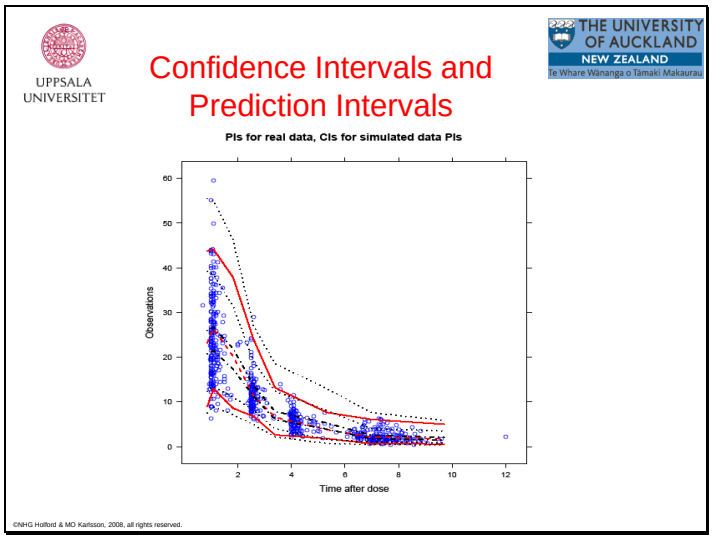
Slide 11



Let's go back to comparing real vs simulated data. This may probably easier done based on a graph with lines rather than areas. (choice of binning may easier be based on the previous graph).

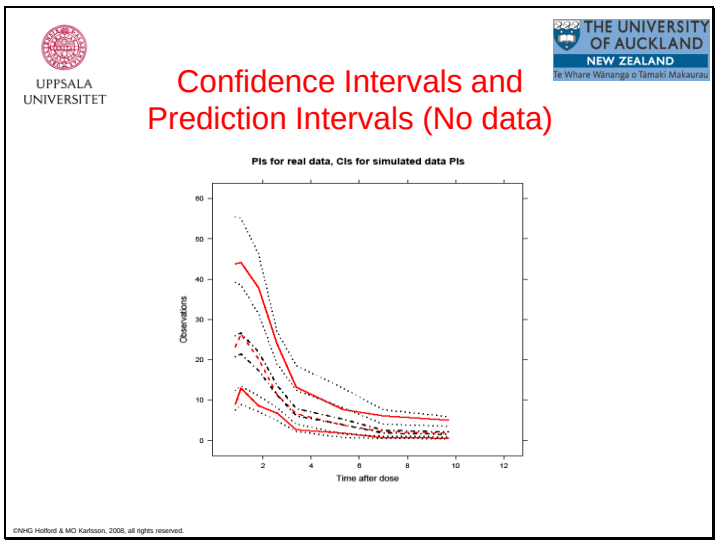
This vpc does look very promising for the model, but how can we be sure that the differences we see are not major? Let's look at the confidence intervals around simulated PIs.

Slide 12



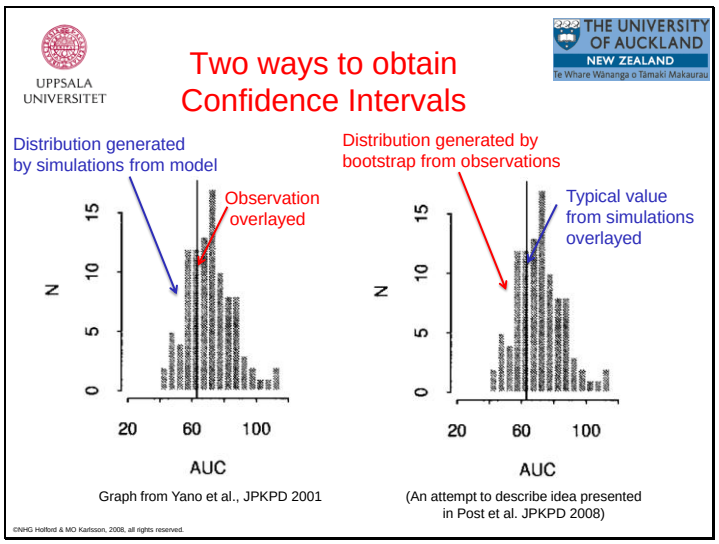
When adding confidence intervals around the PIs, it is better to drop the PIs for simulated data themselves. Otherwise there will be too many lines to keep track of. Thus red lines are PIs for real data, black dotted lines for CIs based on simulated data. This look quite OK, but maybe we could improve further.

Slide 13



Some may find it easier without the observations, whereas others would miss them. With CIs, as opposed to PIs only, there is less need for displaying the observations as the CI will reveal where the information in data is rich and where it is sparse.

Slide 14



We can either create a confidence interval based on simulated data and compare the PI for the real data to this confidence interval. This is the same idea as what Yano et al. Used in their outline of the posterior predictive check. Note that each simulated data set must have the same structure as the original data i.e. same covariates, same number of predicted observations.

An alternative strategy is to create a confidence interval based on the real data, by bootstrapping, and compare the PI for the simulated data to this. The confidence interval generated based on the bootstrap is then compared to the median value generated by simulations. Also in this case must each simulated data set have the same structure as the original data. For this approach one will have to recognise that there are limitations to the bootstrap when data becomes sparse. This will be particularly pronounced if one tries to obtain confidence intervals in the tails of a distribution, e.g. The 10th or 90th percentile. It should be noted that Post et al. only did recommend this bootstrap procedure for the median, not extreme PIs.

Slide 15

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

How to choose Prediction Intervals?

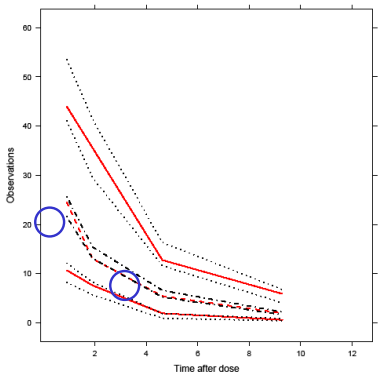
- Different parts of the model (structural, variability, etc) are likely to be maximally informed by different PIs and depend on data richness (Wilkins et al. PAGE 2006)

Slide 16

UPPSALA UNIVERSITET

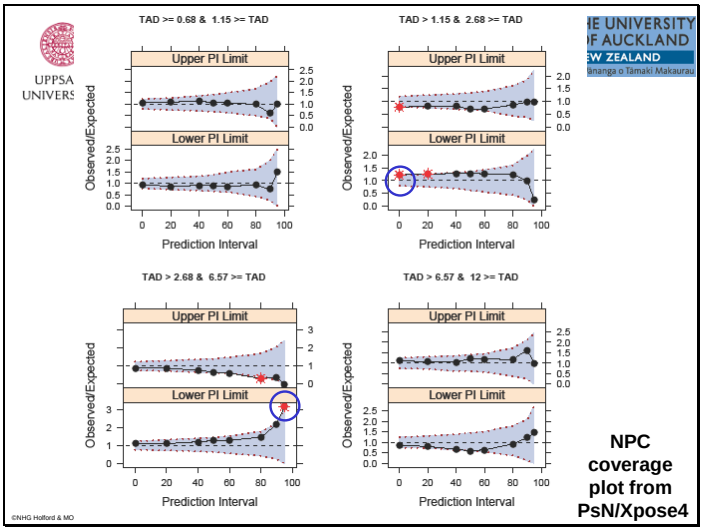
THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

Is this a suitable prediction interval?



For illustration, let's return to the plot with only 4 intervals. If we look at this plot, we can identify two apparently significant deviations between model and data. The blue circles show places where the PI for real data are outside the 95% CI based on the simulations.
Can we feel certain that these are suitable choices for PIs? Would other PIs reveal more pronounced deviations? Let's look at that (next slide).

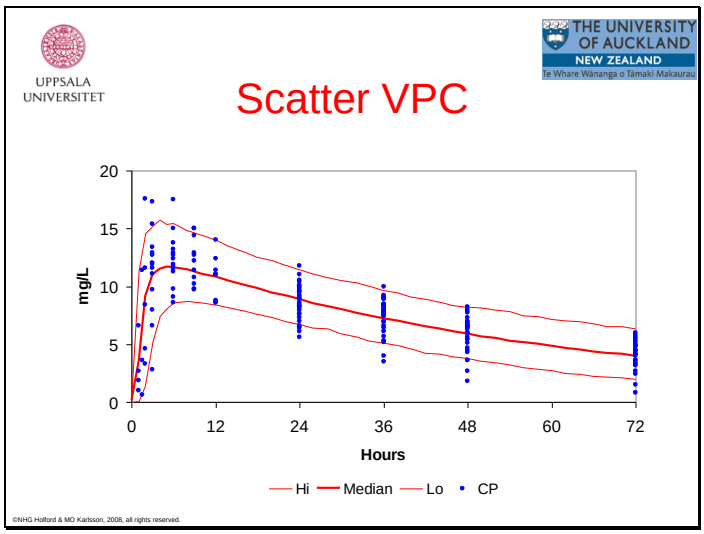
Slide 17



NPC coverage plot from PsN/Xpose4

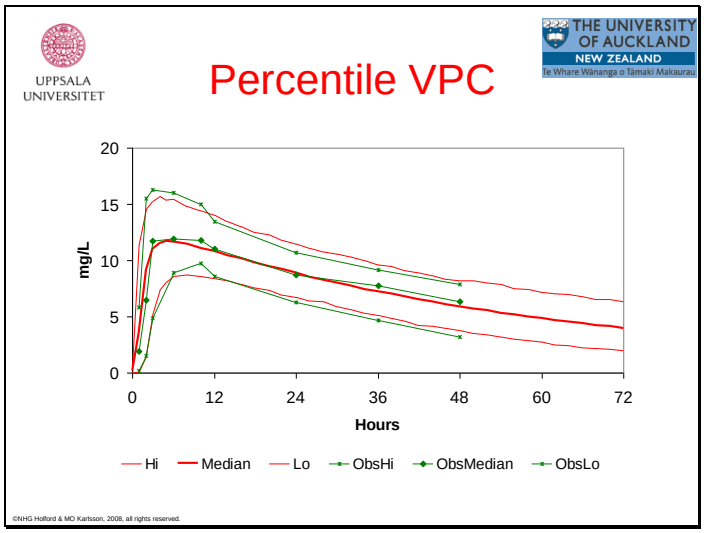
Coverage plots provide information across many PIs. In a VPC only 2 or 3 are possible to show. Covarage plots can provide information about other PIs that may indicate more or less misspecification. The four sets of panels correspond to the 4 binning groups in the previous plot (0-2h, 2-3h, 3-6h and >6h).
In this plot we can identify, in red with blue circles, the two significant differences seen in the previous plot (actually it is three because the 0% PI (which is the median), will turn up twice). We can also see that at other PIs, there seems not to be any major differences. The selected intervals actually is the most "unfavourable" choice of PIs for the model.

Slide 18



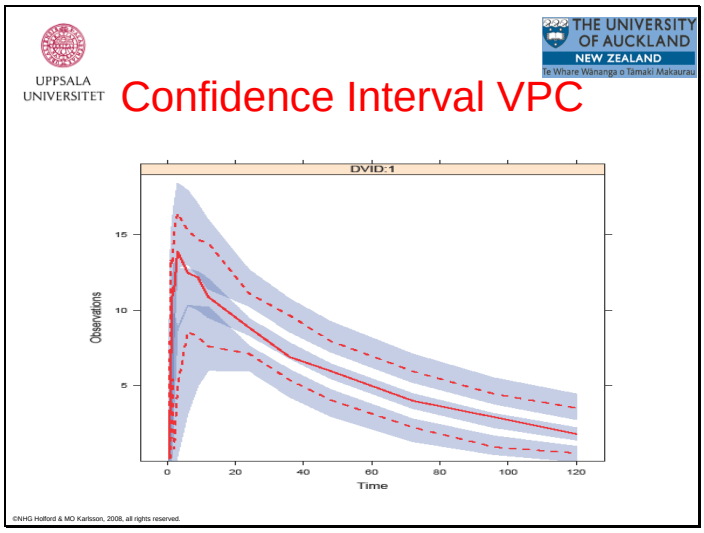
As you have seen there are many ways of creating VPCs with increasing complexity. In summary there are three basic kinds of VPC. The first is the scatter plot VPC which shows the observations along with some simple prediction intervals. This is a useful starting point for connecting observations with prediction intervals. However when there is a lot of data the actual distribution of the observations can be hard to appreciate.

Slide 19



The second kind of VPC summarises the distribution of observations with observation intervals so they can be compared directly with the prediction intervals and the medians of the observed and predicted values. The percentile VPC is easier to interpret when there are lots of observations.

Slide 20



The third type shows the 95% confidence interval around each of the prediction intervals obtained by simulation. The red lines are the observation intervals.

Slide 21

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

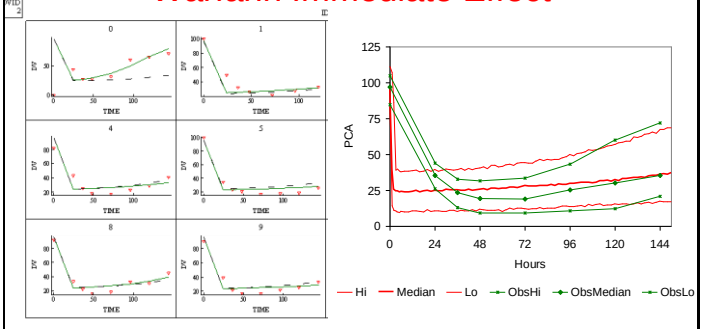
What Can a VPC Show?

Slide 22

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

Warfarin Immediate Effect



Simulated Using Turnover Model

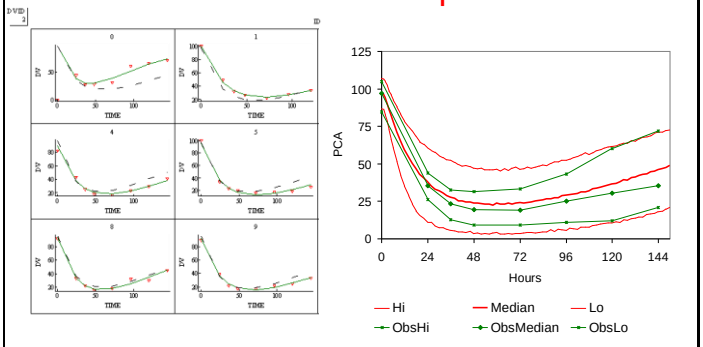
Data has been simulated from a warfarin PKPD model involving turnover of prothrombin complex activity (PCA) after a single oral dose of warfarin. The PK model is first order absorption and first order elimination from one compartment.
The data has been fitted with the same PK model used to simulate the data but the PD model assumes an immediate effect of warfarin plasma concentration on PCA.
The left hand plots show individual predictions obtained from empirical Bayes estimates and the corresponding observations. The right hand plot is a percentile VPC with 90% intervals.
Both the individual plots and the VPC show poor predictions.

Slide 23

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

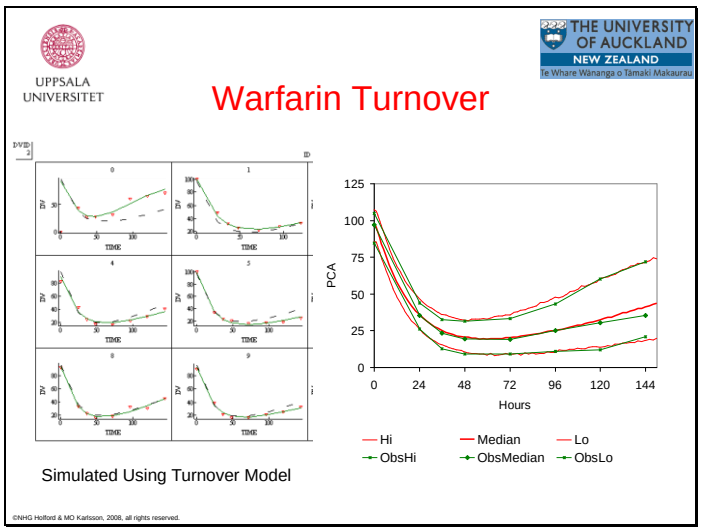
Warfarin Effect Compartment



Simulated Using Turnover Model

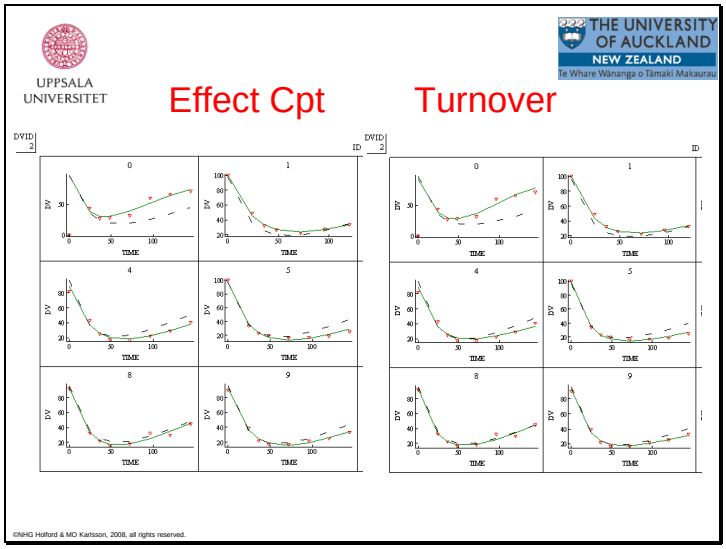
The PD model now assumes a delayed onset of warfarin effect using an effect compartment model for concentrations driving the change in PCA.
The individual predictions look very good but the VPC median prediction lies above the median observation from 24 h onwards and the 90% interval is clearly much wider than the observations. This suggests the model is not properly describing the data despite the very good individual predictions.

Slide 24



Finally we can see what happens when the true model is used to fit the data. When a turnover model is used the individual predictions remain good and the VPC percentile plot looks good as well.

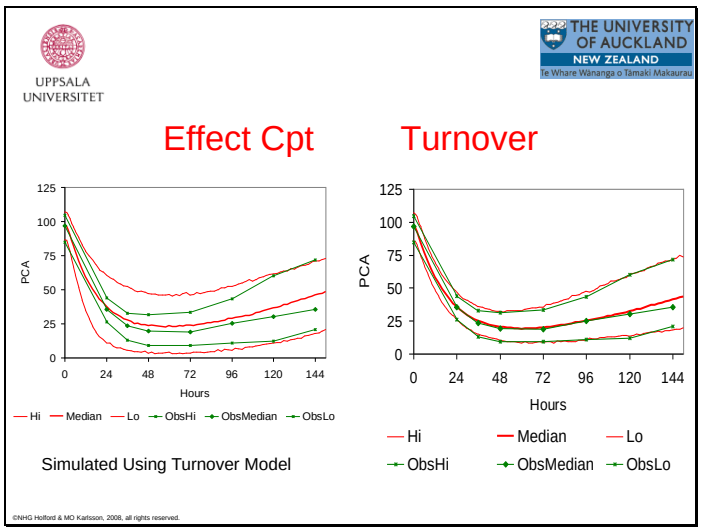
Slide 25



Notice that plots using empirical Bayes estimates for predictions are essentially the same for both the effect compartment and turnover model. Yet the VPCs show the predictions of the effect compartment model are a poorer description of the observations.

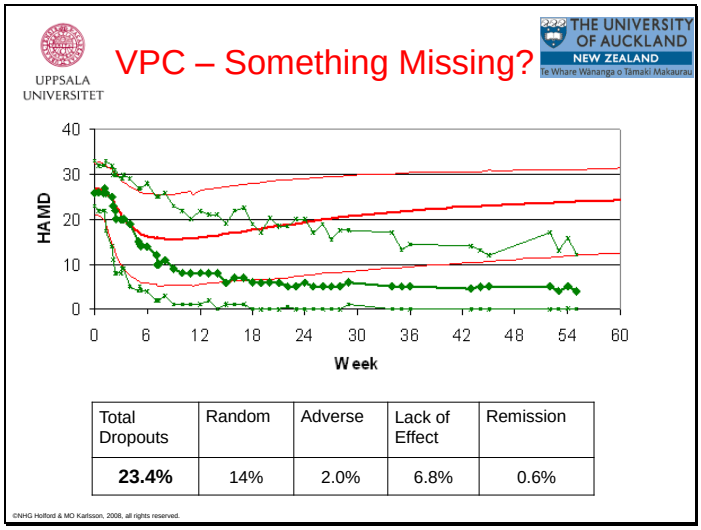
This is a consequence of shrinkage. The shrinkage for the effect cpt model was 14-44% and for the turnover model was 11-40%. It has been suggested that all parameters must have less than 20% shrinkage in order to draw reliable conclusions from individual plots. However, shrinkage estimates do not take account of the correlation between parameters and it is the correlated set of parameters that determines the prediction. Even if one or more parameters have relatively high shrinkage the individual prediction may be quite reliable e.g. for sequential PKPD models using the IPP approach.

Slide 26



This slide compares the VPC using the incorrect effect compartment model with the VPC obtained from the true turnover model. It is reassuring to see that the VPC with the true model has good agreement with the observations.

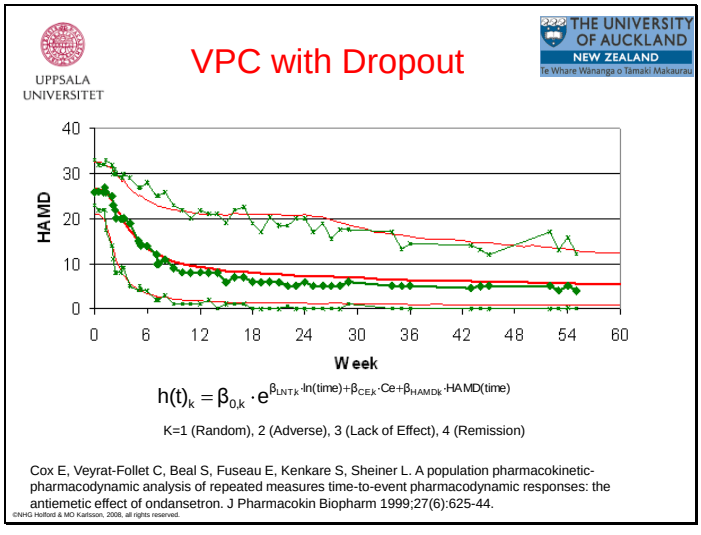
Slide 27



Total Dropouts	Random	Adverse	Lack of Effect	Remission
23.4%	14%	2.0%	6.8%	0.6%

Here is a VPC from a large study of patients in an anti-depressant drug trial. The predicted median and 90% PIs do not agree well with the observed values. This is because there are patients who drop out and the pattern of dropout is influenced by the treatment and patient response. This is known as informative missingness.

Slide 28



When it is possible to predict missing values from the data e.g. patients whose HAMD score remains high may dropout because they are not getting better, then this is known as missing at random. A model for the dropout process can be constructed by combining the model predictions for HAMD with a time to event analysis. When the VPC is performed using the dropout model to include a realistic pattern of dropout then the VPC predictions match more closely with the observations.
The previous VPC is probably makng good predictions if patients continue with treatment. In that case the observations are 'wrong' because they has been censored by dropouts.

Slide 29

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

How to handle censored data?

- Use planned design and include model for censoring and
  - Show only uncensored data, or
  - Replace censored values, both observed and simulated, with e.g. LOCF
- Perform separate predictive check on censored data
  - For example drop-out frequency over time

Slide 30

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

A VPC may fail to identify model misspecification

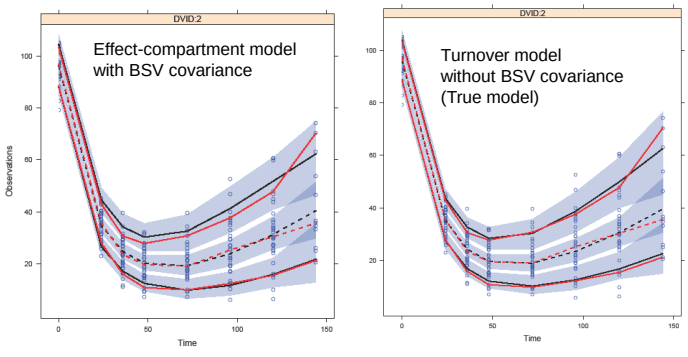
Here will follow three examples where the VPC looks OK, but a model misspecification is present in the underlying model.

Slide 31

UPPSALA UNIVERSITET

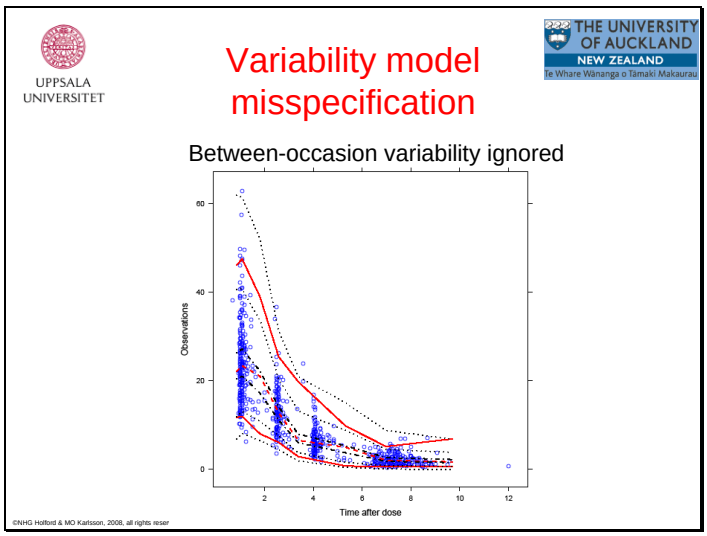
THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

Structural model misspecification



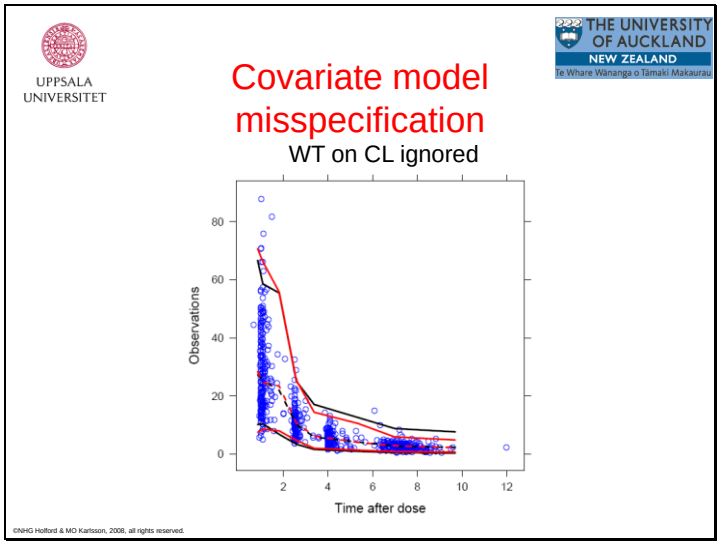
The vpc for the effect-compartment model indicated something was wrong, but not what was wrong. If one were to take the route of trying a different, more flexible, Between-Subject Variability (BSV) model, a choice that also was supported if one looked at the EBEs, then one would find that such a model very well described the data according to a VPCdespite using the wrong structural model.

Slide 32



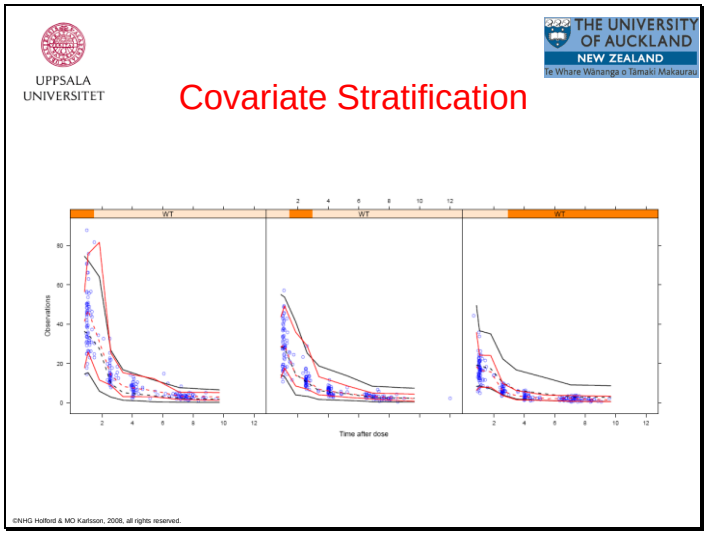
There was pronounced BOV in the model that simulated these data, but the model that was used to generate the above VPC ignored BOV. Despite this misspecification the VPC looks ok. This illustrates that VPCs may not always show up variability model misspecifications.

Slide 33



This is a VPC of a model without covariates describing pediatric PK data over a wide age and weight range. Despite ignoring an important covariate, it looks fine. However, as opposed to the two presvious examples, here we can make the VPC considerably more informative. This we accomplish by stratification of the VPC by weight (next slide).

Slide 34



The left, middle and right panels show a VPC for the lightest, medium-weight and heaviest children (not the color bar at the top of each panel). The misspecifications are evident.

Slide 35

What to stratify on?

- Consider for example to stratify on different:
  - response variables
  - trial arms (active / placebo)
  - doses
  - dose intervals (BID vs QD)
  - studies (in meta-analyses)
  - routes of administration
  - covariate values
  - occasions (if TAD is independent variable)

Slide 36

Pros and cons of stratification

- Pro
  - Allows subset of data/model to be inspected
  - Can increase resolution of model misspecification
- Con
  - Can dilute the signal
- Alternative
  - Correct for variability in predictions in a bin due to covariates, design and observation time

$OBS_{ij} = OBS_{ij} * PRED_{av,bin} / PRED_{ij}$

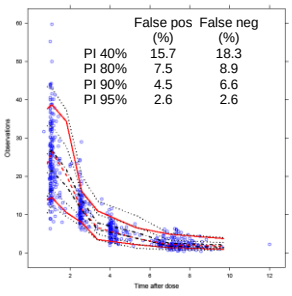
$SPRED_{ij} = SPRED_{ij} * PRED_{av,bin} / PRED_{ij}$

When there are considerable variability in the expected values due to differences in design (e.g. Different doses) or covariate values of an important covariate relation, much of the variability defining the PIs will not come from the unexplained variability. This may make the VPC less sensitive (examples to follow). PRED-correction is similar to dose-normalisation of concentration data. It allows a normalisation, in this case based on the PRED value of an individual observation/SPRED compared to the average PRED in the bin. [A similar correction could be made also for the variance component but that is not included here]

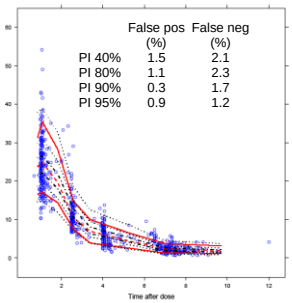
Slide 37

PRED-correction

Standard VPC



PRED-corrected VPC



In this case there is no model misspecification, but we can see that part of the observed variability was due to differences in design, covariates and binning. The % false+ and false- decrease considerable with correction.

Definition of false +ve is an observation that compared to the directly corresponding simulated SPREDs is within the simulated PI, but compared to the binned PI is outside. False –ve are of course the opposite. False +ve and –ve are created by binning (across single and multiple doses; across different times within one binning interval) and lack of stratification (by covariate values, doses).

The fact that it does not decrease to zero is because correction only is made with respect to the typical prediction not to differences in variability between data. The fact that for the standard VPC false + and false – are of similar size can be a sign

that they take eachother out (although to be sure of that we would like to see the number of false + and false negative rate per bin and per PI, upper and lower).

Slide 38

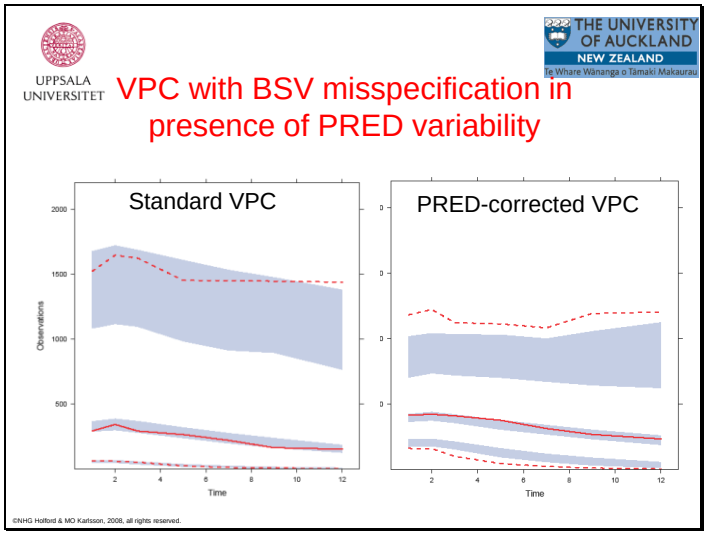
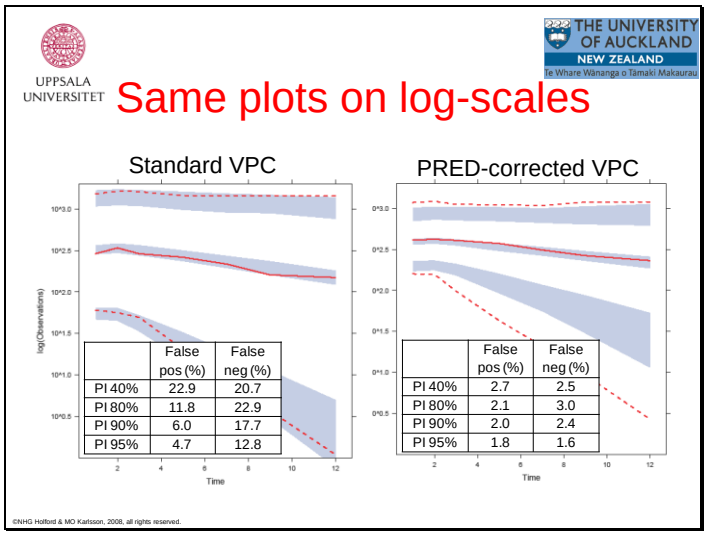


Illustration of concentration-time data PIs for a steady-state dosing interval. The standard VPC is not as sensitive as the PRED-corrected VPC to pick up the misspecification of the variability in this case (variability in CL is underestimated), since there is much variability due to differences in covariate values (genotype influencing CL/F). PRED-correction reveals more clearly the misspecification.

Slide 39



	False pos (%)	False neg (%)
PI 40%	22.9	20.7
PI 80%	11.8	22.9
PI 90%	6.0	17.7
PI 95%	4.7	12.8

	False pos (%)	False neg (%)
PI 40%	2.7	2.5
PI 80%	2.1	3.0
PI 90%	2.0	2.4
PI 95%	1.8	1.6

Same PIs as previous slide but just displayed on log-scale. Also displayed are the false + false –ve rates. For the standard VPC these should heighten our suspicion that something is wrong with the way the vpc represents the model. The false +ve rate is much lower than the false –ve rate. Thus, the standard vpc may make the model look too good. This indeed is what becomes evident when we look at the PRED-corrected VPC.

Slide 40

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

How to do a VPC

- Simulate Data
  - Can be the hardest part
  - Simulation times (binning)?
  - How to simulate covariates?
- Group Simulated Predictions at each time
  - Needs some programming
- Group Observations at each time
  - Needs some programming

There are 3 steps involved in creating a VPC. The first step is to simulate from the model to produce predictions. This step typically requires user intervention for every dataset that is being studied. The next two steps can usually be automated with procedures that are the same for all problems. A convention is needed to identify the independent and dependent variables (especially when there is more than one type of observation e.g. Concentrations and efffects).

Slide 41

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

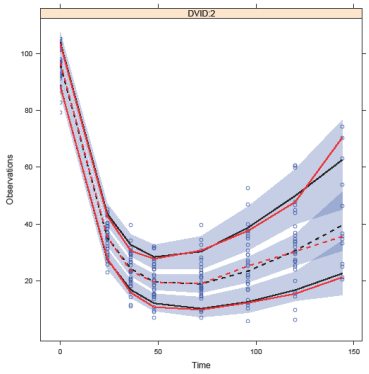
Available Freeware Packages

- PSN / Xpose4
  - psn.sf.net
  - xpose.sf.net
- R for NONMEM
  - www.metruminstitute.org/downloads/mitools/index.html
- Others?

Slide 42

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Whare Wānanga o Tāmaki Makaurau



vpc run1.mod –lst=run1.lst –samples=1000

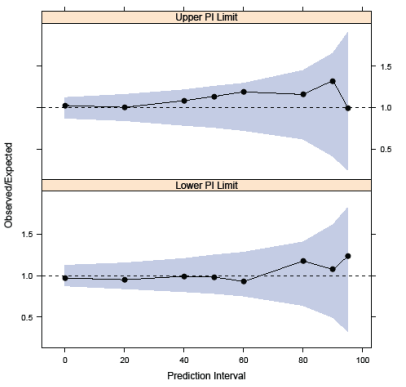
xpose.VPC(PI=”lines”, PI.real=”lines”,PI.ci=”area”)

This is a VPC graph generated from the PsN/Xpose programs. It assumed that a model has been run (run1.mod, run1.lst). The command
” vpc run1.mod –lst=run1.lst –samples=1000 ”
Will automatically generate all files that are necessary to generate a vpc based on 1000 simulated data sets. The vpc tool has many binning, stratification and other options.

By opening the R program and giving the command
” xpose.VPC(PI=”lines”, PI.real=”lines”,PI.ci=”area”)”
the displayed plot will be generated.

There are many options for cusomization of Xpose plots. All plots shown in this presentation except 18-19 and 22-28 were generated with Xpose.

Slide 43



npc run1.mod –lst=run1.lst –samples=1000
npc.coverage()

This is an example of a coverage plot where Pis for observed data can be compared to those of simulated data and including Cis for the simulated data. The advantage being that many PIs (in the above graph for 2.5, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 97.5) can be inspected at the same time (not only 10, 50, 90) as is often the case in a VPC)

The command
" npc run1.mod –lst=run1.lst –samples=1000 "
is given to PsN to automatically create and run NONMEM and post-process output generated by NONMEM. Following this, the command
"npc.covarage()"
is given in the R program and Xpose will generate the displayed plot. There are several options to customize the graph.

Slide 44

Acknowledgement

- Programming of VPC functionality in PSN and Xpose4 by **Kajsa Harling** and **Andrew Hooker**

The VPC functionality in PsN and Xpose has been developed by Mats Karlsson, Andrew Hooker and Kajsa Harling. Valuable input has also been provided by Rada Savic and Justin Wilkins.

Slide 45

UPPSALA UNIVERSITET

THE UNIVERSITY OF AUCKLAND NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Discussion Points

- User-selected aspects
  - Number of simulations
  - Level of stratification
  - Level of binning
  - PI to use
- VPC for what
  - Drive model building vs final quality-check
  - Structural vs variability vs covariate model
  - Apply to all final models?