The 6 Biggest Pharmacometrics Modelling Mistakes!

Alan Maloney PhD
Consultant Pharmacometrician

PAGE Meeting
9 June 2017

The 6 Biggest (technical) Pharmacometrics Modelling Mistakes!

Alan Maloney PhD
Consultant Pharmacometrician

PAGE Meeting
9 June 2017

3

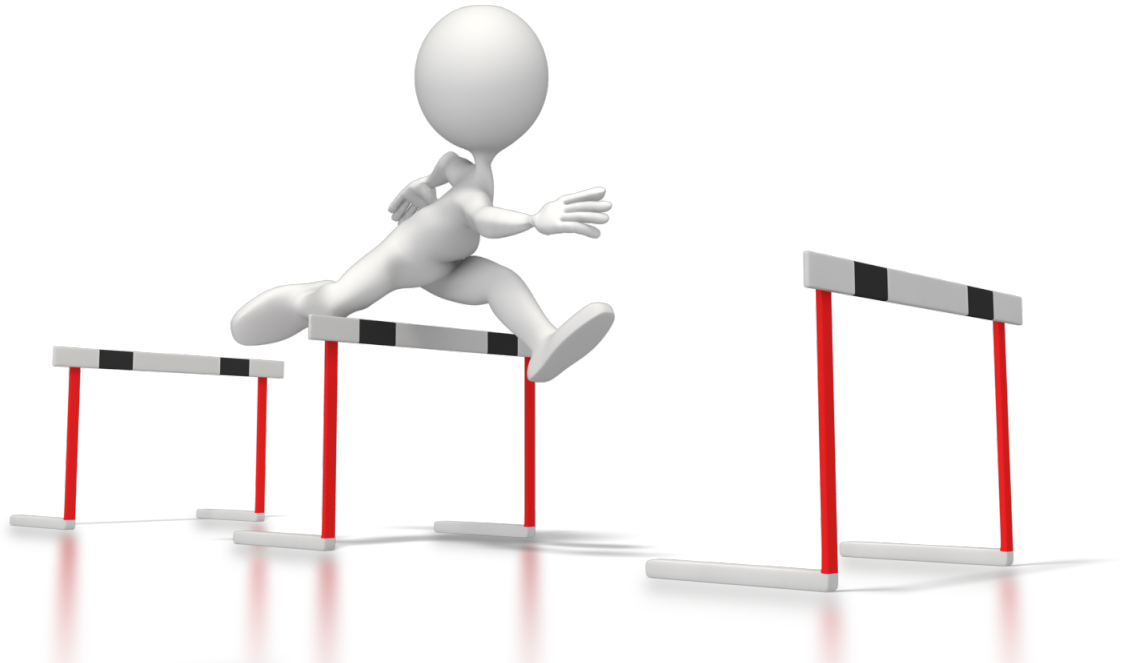
My goal today !

To help you avoid some of the mistakes I have made/seen.

…and to ask you to challenge your modelling "toolkit", so you can avoid every obstacle !

4

To help you avoid some of the mistakes I have made/seen.



We are all indebted to our educators, past and present…

5

“If I have seen further than others, it is by standing upon the shoulders of giants.”

Isaac Newton

…however we should not assume everything we were taught is faultless

6

"If I have seen further than others, it is by standing upon the shoulders of giants."

Isaac Newton

"Some years ago I was struck by the large number of falsehoods that I had accepted as true in my childhood, and by the highly doubtful nature of the whole edifice that I had subsequently based on them."

Rene Descartes

(Edifice = complex system of beliefs)

7

Advice

"If I have seen further than others, it is by standing upon the shoulders of giants."

Isaac Newton

"Some years ago I was struck by the large number of falsehoods that I had accepted as true in my childhood, and by the highly doubtful nature of the whole edifice that I had subsequently based on them."

Rene Descartes

(Edifice = complex system of beliefs)

Don't believe everything you were taught 100% (and, by extension, what I say today !!)

Advice Box

There are 3 areas we need to understand to be good modellers…

8

Design/Data Space

Controls the data we get…key!!

Model Space

What model(s) can we consider?

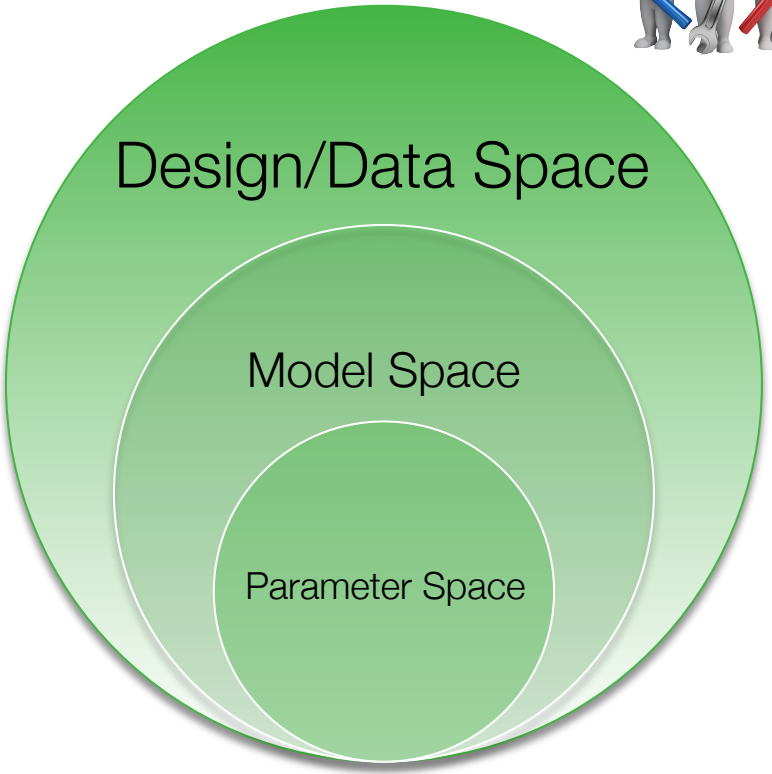
Conditional on the Design/Data Space

Parameter Space

What are the model parameters?

Conditional on the Design/Data Space

Conditional on the Model Space



9

Advice

Design/Data Space

Controls the data we get…key!!

Model Space

What model(s) can we consider?

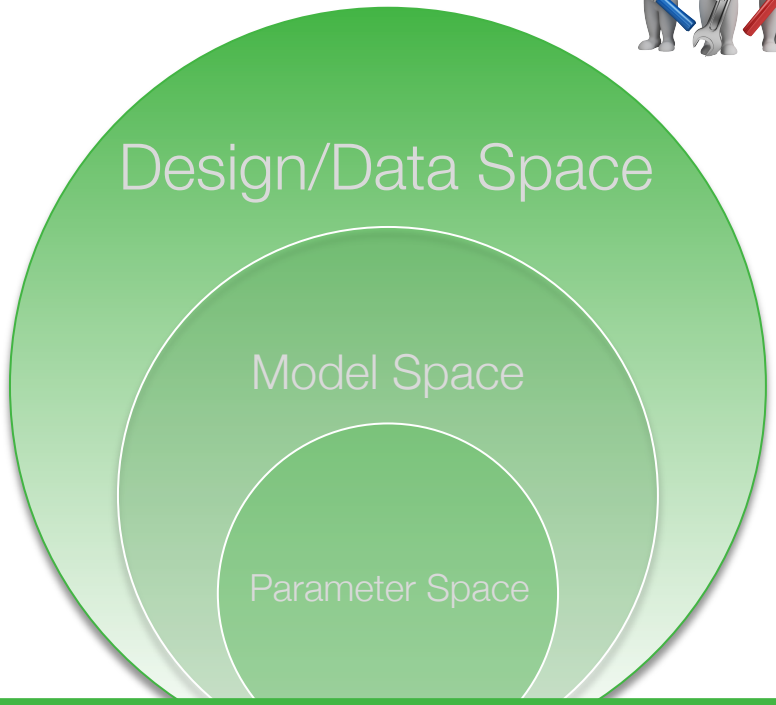
Conditional on the Design/Data Space

Parameter Space

What are the model parameters?

Conditional on the Design/Data Space

Conditional on the Model Space



Focus on the design most, then the models, then the parameters.

Advice Box

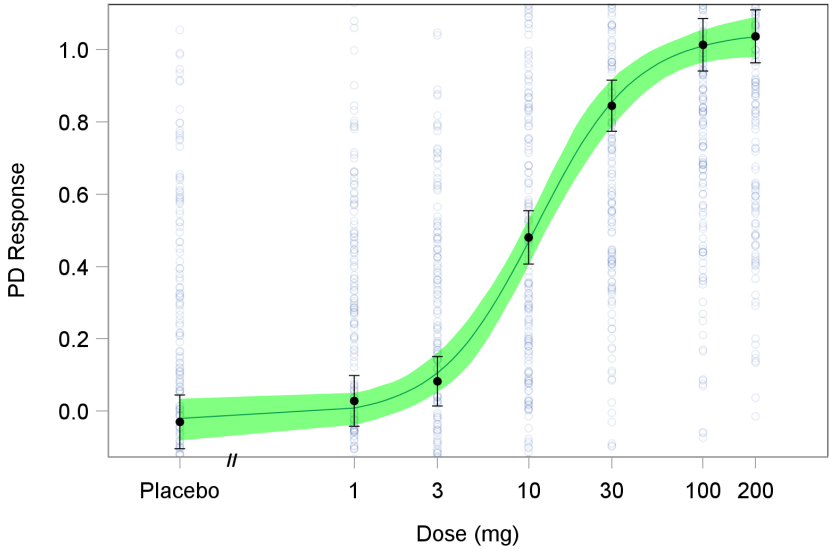
Error 1: When a poor design limits the model space

10

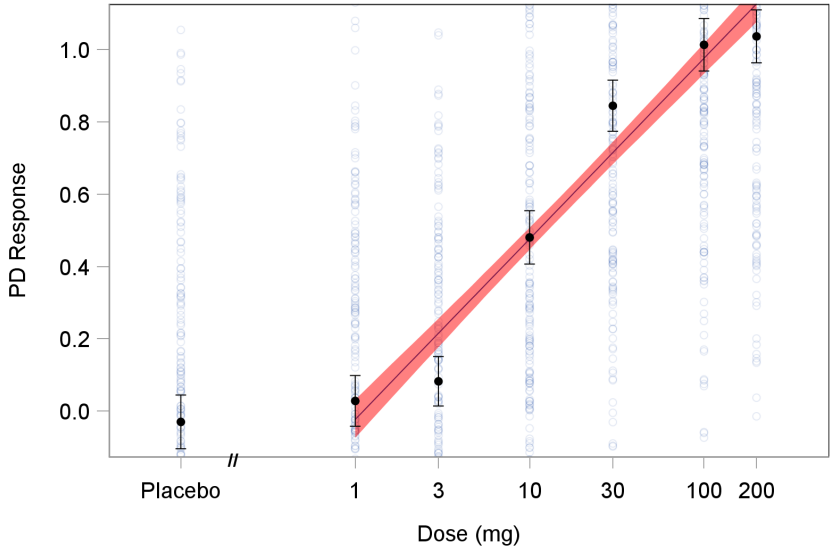
With a great design, we can clearly differentiate between good/bad models…

11

Sigmoidal Emax fit



Log-linear fit

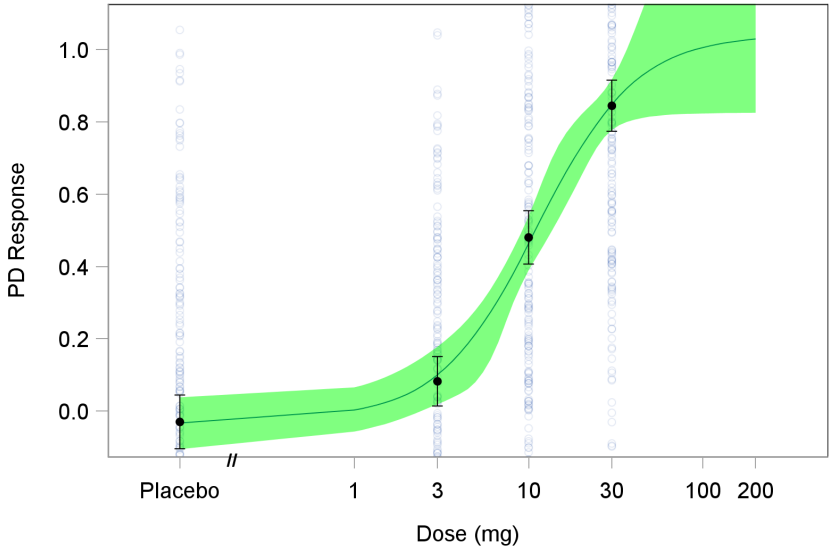


Log-linear model fit clearly worse than sigmoidal Emax fit (Δ-2LL = 35)

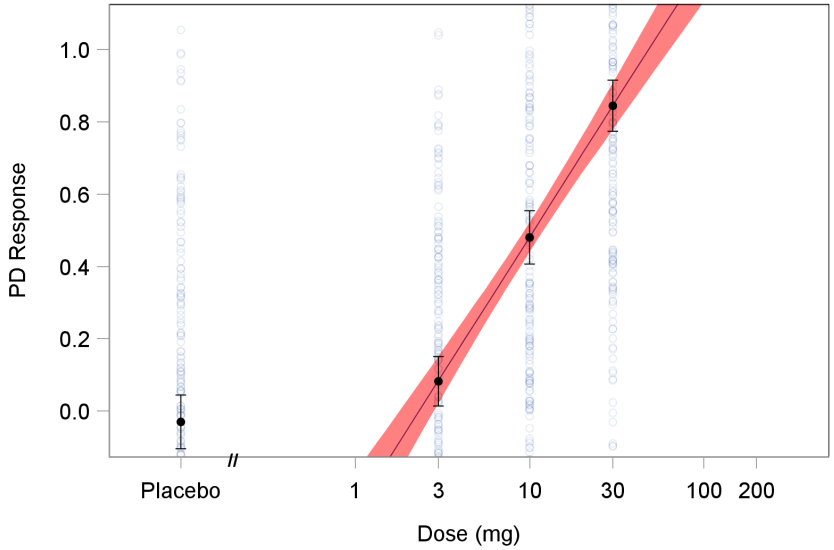
…however with a poor design, we may struggle to differentiate between good/bad models

12

Sigmoidal Emax fit



Log-linear fit



Log-linear model fit ***not*** clearly worse than sigmoidal Emax fit (Δ-2LL = 0)

Error 1: When a poor design limits the model space

13

Always consider how the design could:

- Limit your ability to fit the model(s) you wish to consider
- Limit your ability to differentiate between candidate models

Some consequences:

- Poor inference (e.g. when forced to fit overly simplistic models, or forced to fix/guess model parameters) => Poor decision making
- Biased estimates (e.g. IIV on PK absorption parameters when limited data collected in the absorption phase)

14

Advice

Always consider how the design could:

- Limit your ability to fit the model(s) you wish to consider
- Limit your ability to differentiate between candidate models

Some consequences:

- Poor inference (e.g. when forced to fit overly simplistic models, or forced to fix/guess model parameters) => Poor decision making
- Biased estimates (e.g. IIV on PK absorption parameters when limited data collected in the absorption phase)

Appreciate that fitting complex models to weak data may be a recipe for disaster.
At the design stage, assess if you will be able to fit/differentiate between candidate models.

Advice Box

Error 2: Model Selection: On the misuse of Hypothesis Testing !!

15

The principle of parsimony - "…the simplest model that adequately describes" - is poor (modelling) science. Simple, in most cases, does not equate to smart, sensible or sound (recall the log-linear model fit before)

Significance testing and the parsimony principle have become "twin devils". Just because something is not "statistically significant", does not mean it is not important.

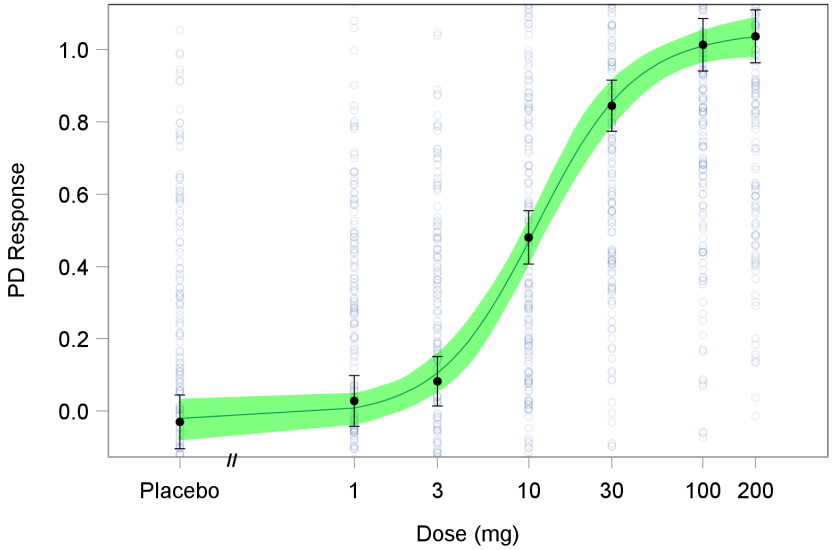
Significance testing depends on design and N (as well as the true effect).

We should try to avoid "weasel words" like "no evidence of a difference with respect to age, sex, race…" (from "non-significant" results)

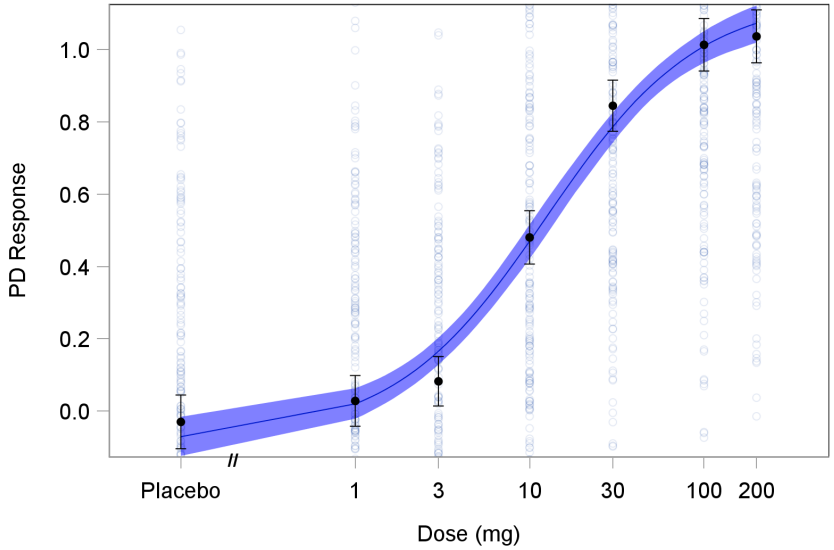
Hypothesis testing is useful for identifying weak (and hence not interesting) models

16

Sigmoidal Emax fit (N=200 per arm)



Simple Emax fit (N=200 per arm)



Hill = 1.54 (approx. SE=0.23)

Hill =1 (fixed). Δ-2LL = 9.2

…however we cannot used “non-significant” differences to rule out models

17

N (per arm)	Δ-2LL (expected)		Decision	Correct?
N = 200	*9.2*		*Estimate Hill*	*Yes*
N = 100	*4.6*	**> 3.84 ?**	*Estimate Hill*	*Yes*
N = 50	*2.3*		*Fix Hill = 1?*	*No*

As our information gets weaker (N decreases) it is an error to be “more sure” that Hill is exactly 1. Subsequent predictions intervals may be biased and/or too narrow (and, in this case, clearly wrong).

18

Advice

N (per arm)	Δ-2LL (expected)		Decision	Correct?
N = 200	*9.2*		*Estimate Hill*	*Yes*
N = 100	*4.6*	**> 3.84 ?**	*Estimate Hill*	*Yes*
N = 50	*2.3*		*Fix Hill = 1?*	*No*

As our information gets weaker (N decreases) it is an error to be "more sure" that Hill is exactly 1. Subsequent predictions intervals may be biased and/or too narrow (and, in this case, clearly wrong).

Use hypothesis testing to rule out "clearly worse" models.
Generate predictions (…effect on decisions) for all "similar fit" candidate models.

Advice Box

Error 3: Model Assessment: Using simple diagnostic plots as evidence a model is sound

19

NLME models are complex. The ability of random effects to generate good fits (even when the model is very poor) should always be appreciated.

Thus simple diagnostic plots, such as:

- Observed v Individual Prediction (DV v IPRED) and
- Observed v Population Prediction (DV v PRED)

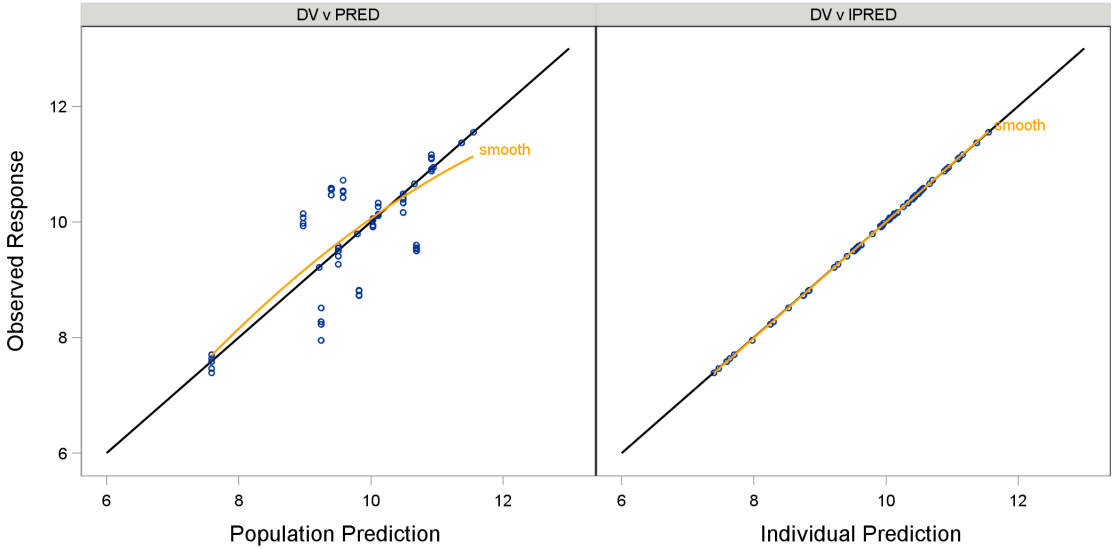
provide very limited evidence that a model is sound

(For me, they are archaic plots that provides no evidence the model is OK, and we can now leave them behind !)

It is straightforward to create examples where these plots are misleading…

20

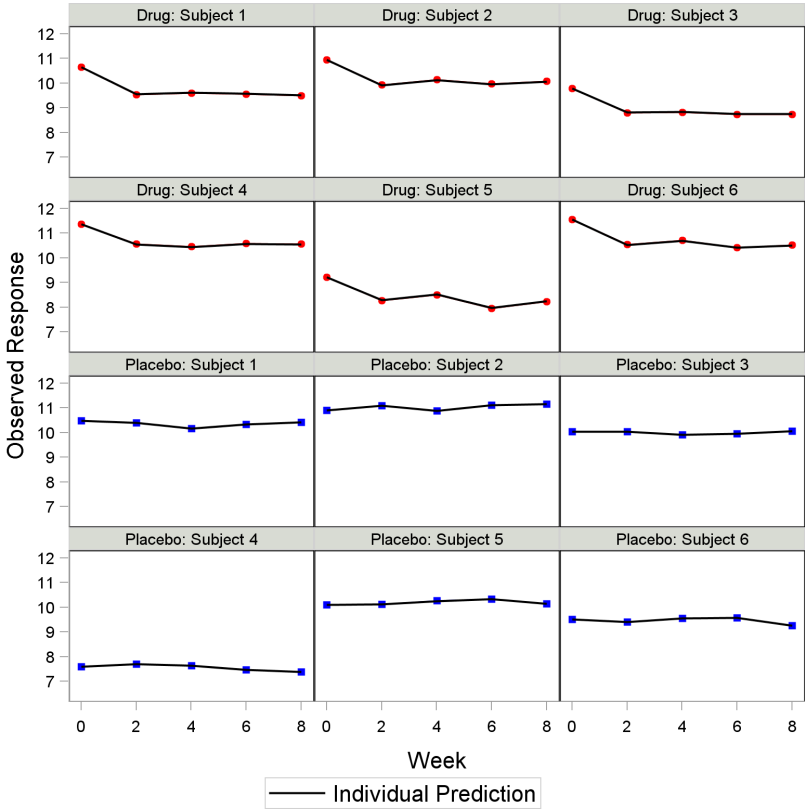
Example 1: A longitudinal model



The plot looks very good, but the model is not a good one!

In this longitudinal model, the individual fits look great…

21



The data behind this "perfect fit" model

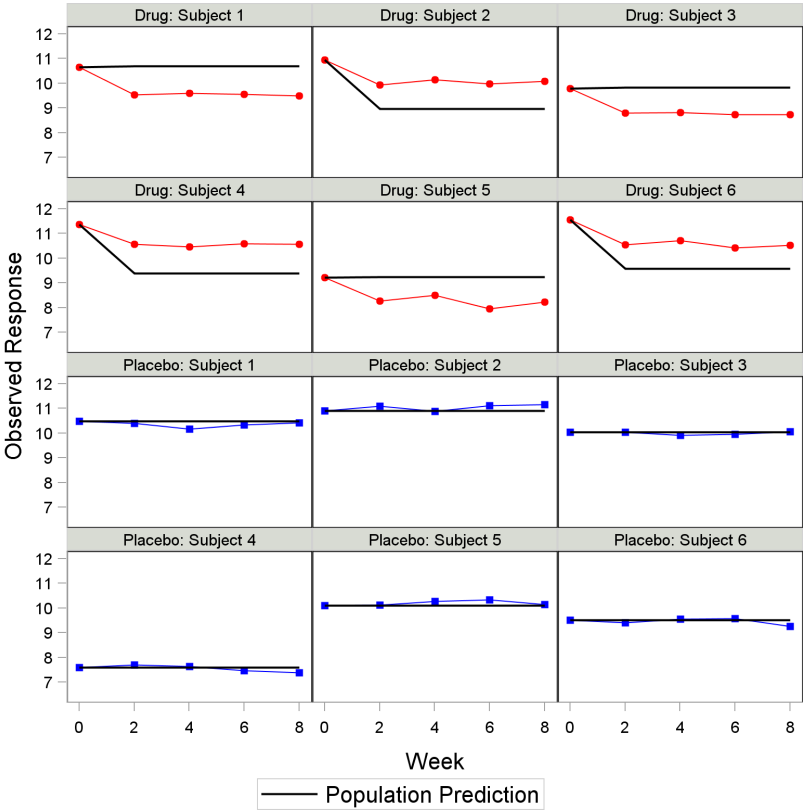
- Observed Drug data
- Observed Placebo data

The simulation used a drug effect (of -1) for all Drug subjects post baseline.

This model "appears" to be fine but…

22

…but the drug model is nonsense…



The drug effect depends on the subject ID number (even or odd), that is:

Subjects 1, 3, 5 = 0 drug effect

Subjects 2, 4, 6 = -2 drug effect

…not a good drug model!

Note: the trick here is using IOV to make things look nice, when they are not.

23

Advice

The drug effect depends on the subject ID number (even or odd), that is:

Subjects 1, 3, 5 = 0 drug effect

Subjects 2, 4, 6 = -2 drug effect

…not a good drug model!

Be careful not to over interpret these basic plots.
Use random effects carefully. They can make a bad model look good!
Buy an old stats book!...Residual Analysis. Try DV-IPRED on the y-axis!

Advice Box

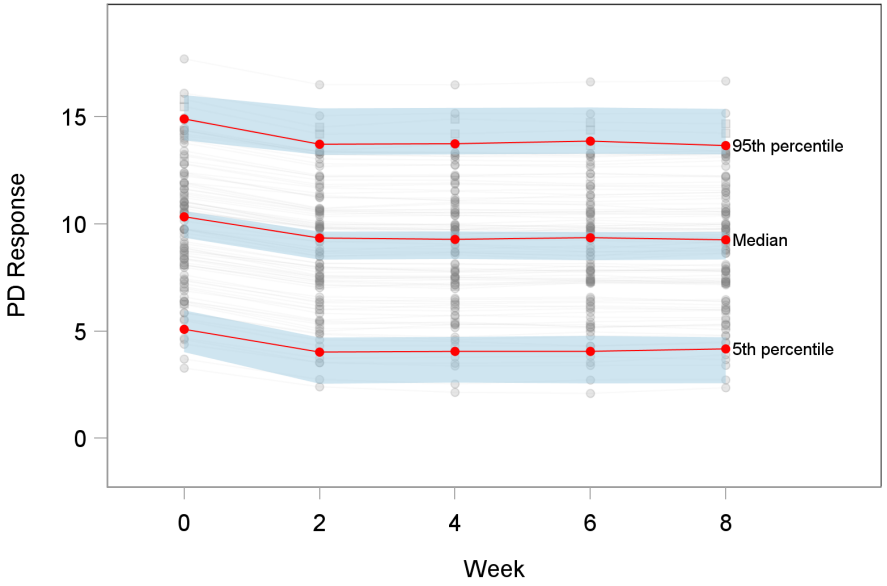
Error 4: VPCs: Using simple VPCs to argue a model is sound

24

Predictive checks (or VPCs in the PM world!) are an excellent way to challenge a model, although most VPCs I see do not challenge the most critical parts of the model (e.g. the drug model) in any meaningful way.

A nice VPC is often just a reflection of baseline IIV…use more challenging VPCs than this!

25



Longitudinal Model – Drug Arm

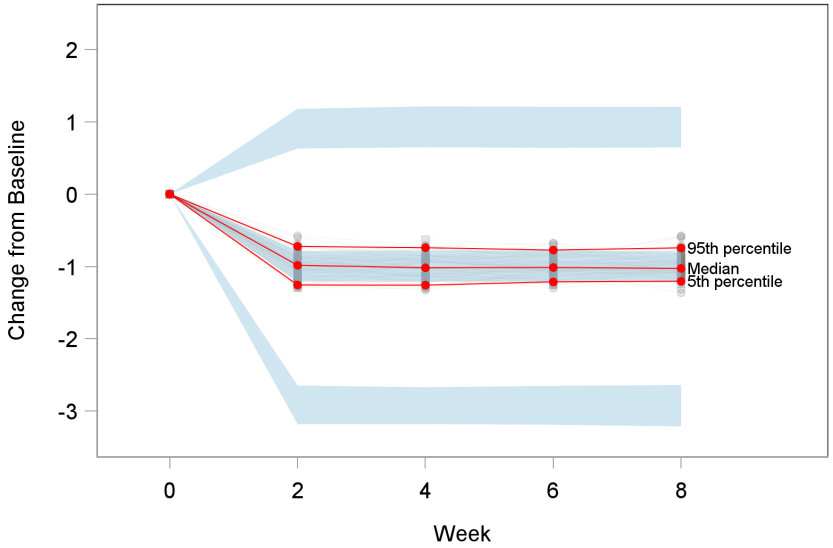
The same (nonsense) drug model as before, but with n=100 subjects and a higher baseline IIV

… the baseline IIV is hiding the model inadequacies.

This type of basic VPC is not very useful for PD data

26

…a better VPC shows the model inadequacies



Using Change from Baseline

The individual changes from baseline are all about -1, unlike the odd model !

27

Try some more interesting VPCs !

It is great if you can show that the model can describe each component of the observed data (note : use these during model development, not after)

Can the model describe Baseline?

As well as the mean, median, 5th/95th percentiles, the CDF, interquartile ranges, SD, min, max etc.

Can the model describe Baseline + Placebo?

Absolute score, changes from baseline, change between time points (e.g. week 26 – week 12), % of patients changing by category (0-25%, 25-50%, >50% improvement or deterioration etc.)

Can the model describe Baseline + Placebo + Drug?

As above, but now also the "delta" between Placebo and Drug, or different dose levels …e.g. difference in medians, means, percentiles, categorical % change etc. For example:

% Drug Patients with > 50% improvement – % Placebo Patients with > 50% improvement (by time point)

28

Advice

It is great if you can show that the model can describe each component of the observed data (note : use these during model development, not after)

Can the model describe Baseline?

As well as the mean, median, 5th/95th percentiles, the CDF, interquartile ranges, SD, min, max etc.

Can the model describe Baseline + Placebo?

Absolute score, changes from baseline, change between time points (e.g. week 26 – week 12), % of patients changing by category (0-25%, 25-50%, >50% improvement or deterioration etc.)

Can the model describe Baseline + Placebo + Drug?

As above, but now also the "delta" between Placebo and Drug, or different dose levels …e.g. difference in medians, means, percentiles, categorical % change etc. For example:

% Drug Patients with > 50% improvement – % Placebo Patients with > 50% improvement (by time

Use VPCs to challenge every part of your model (baseline, placebo and drug effects)
Look at "deltas" of changes from baseline, changes between time points/doses etc.

Advice Box

Error 5: Model uncertainty: Selecting a single “final” model

29

Our (historic) goal to seek a single, final model is not good science. This mind-set needs to change.

We should view model uncertainty in the same way as we see parameter uncertainty (i.e. it is always there, and it is naïve to pretend it does not exist)

Robustness of conclusions (e.g. dose selection) across a range of candidate models is compelling. If results change meaningfully across this model space, we can plan accordingly (i.e. hope for the best, plan for the worst)

30

Advice

Our (historic) goal to seek a single, final model is not good science. This mind-set needs to change.

We should view model uncertainty in the same way as we see parameter uncertainty (i.e. it is always there, and it is naive to pretend it does not exist)

Robustness of conclusions (e.g. dose selection) across a range of candidate models is compelling. If results change meaningfully across this model space, we can plan accordingly (i.e. hope for the best, plan for the worst)

If you cannot clearly distinguish between competing models, don't!
Progress all candidate model to the prediction stage, to see if model choice is critical.

Advice Box

Error 6: Parameter uncertainty: say NO to the NP bootstrap !

31

The non-parametric bootstrap is ubiquitous…we were all amazed at its simplicity, and coded it quickly…but

…who taught you the theory part?

We wish to know the joint distribution of the model parameters, conditioning on the model and data (P(Θ | model, data).

Why are 1000 sets of Maximum Likelihood estimates equivalent to that? (clue, they are not!)

32

Advice

The non-parametric bootstrap is ubiquitous…we were all amazed at its simplicity, and coded it quickly…but

…who taught you the theory part?

We wish to know the joint distribution of the model parameters, conditioning on the model and data (P(Θ | model, data).

Why are 1000 sets of Maximum Likelihood estimates equivalent to that? (clue, they are not!)

Bayesian MCMC is a beautiful way to describe the joint distribution of the parameters.
If you haven't starting using Bayesian analysis, try it ! …you will not regret it !

Advice Box

33

Summary

It is important to (continually) challenge our modelling toolkit.

34

Summary

It is important to (continually) challenge our modelling toolkit.

I hope this was informative, and made you think.

35

Summary

It is important to (continually) challenge our modelling toolkit.

I hope this was informative, and made you think.

Also on YouTube https://youtu.be/E3T2p6Mv0Xc

36

Summary

It is important to (continually) challenge our modelling toolkit.

I hope this was informative, and made you think.

I hope the advice will be useful.

37

Summary

It is important to (continually) challenge our modelling toolkit.

I hope this was informative, and made you think.

I hope the advice will be useful.

Disagree? I am always happy to learn. Call me!

38

Summary

It is important to (continually) challenge our modelling toolkit.

I hope this was informative, and made you think.

I hope the advice will be useful.

Disagree? I am always happy to learn. Call me!

Good Luck!

39

Thank You to…

Janet Wade and Ekaterina Gibiansky for kindly providing comments on an early draft of these slides

40

Thank You !! …Any Questions?

Or feel free to contact me at : al_in_sweden@hotmail.com

41

Backup

…poor decisions do not always lead to poor outcomes, but no way to go through life!

42

To minimise the risk of poor outcomes, we should strive for excellence at every stage of modelling.

Be aware of the our own "anecdotal" viewpoint – "In my experience…nothing bad happened"

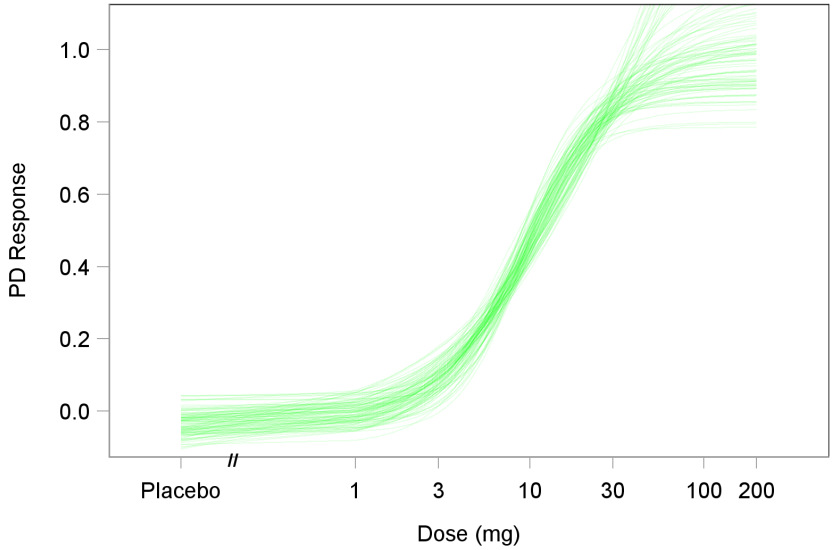
…like this young lady saying:

"But I have never been hit!"

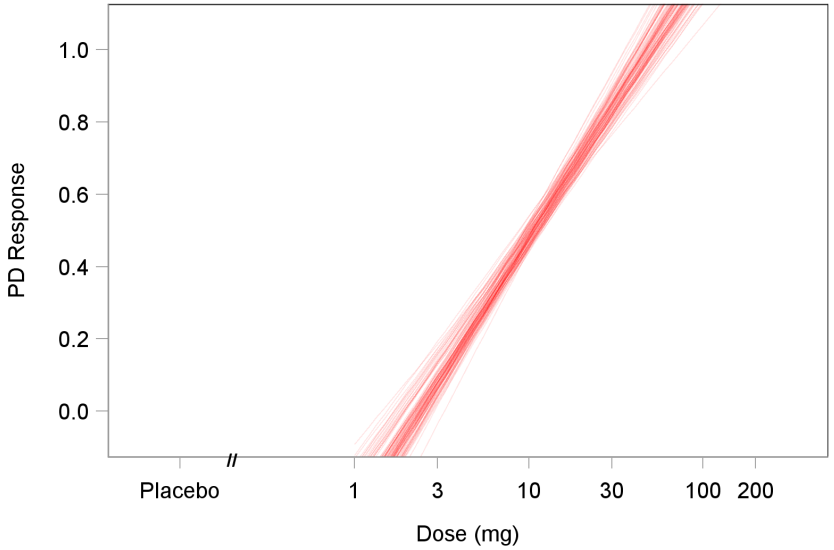
The first 100 (MCMC) D-R relationships for the "poor design" case

43

Sigmoidal Emax fit



Log-linear fit

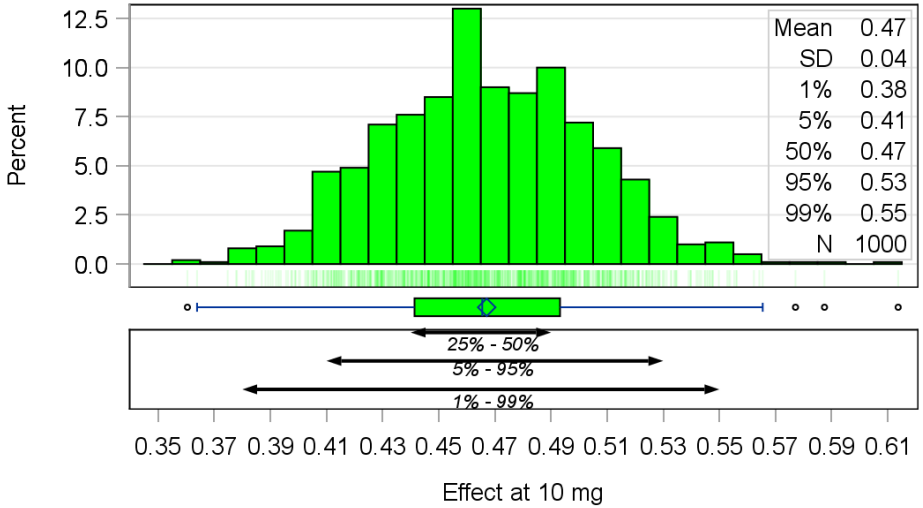


Bayesian fits for both models.

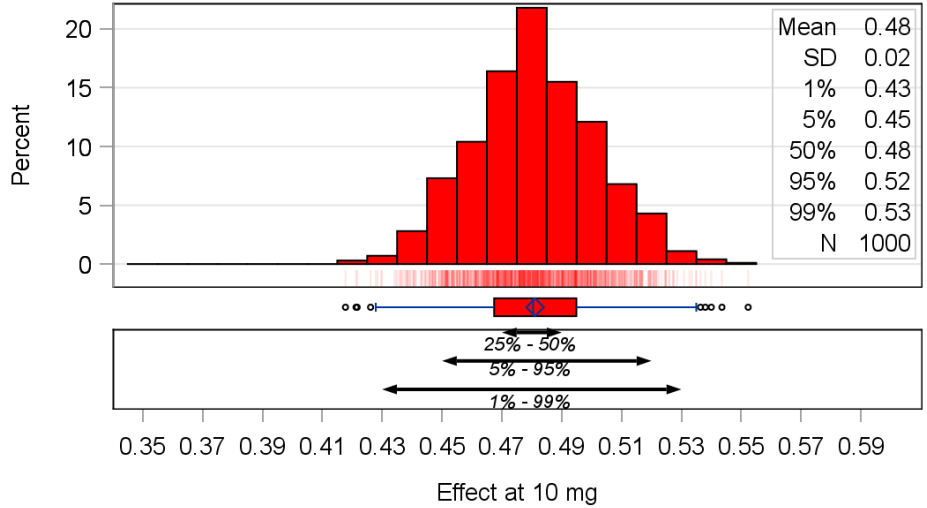
…note the (erroneous) ‘gain’ in precision at the 10 mg dose level

44

Sigmoidal Emax fit



Log-linear fit

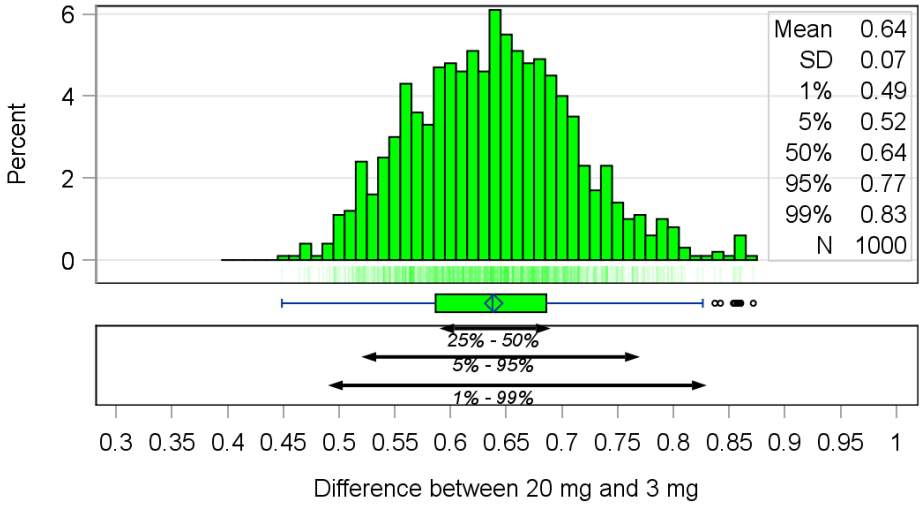


When we condition on dubious models (like the log-linear model), the resulting prediction intervals may be spurious.

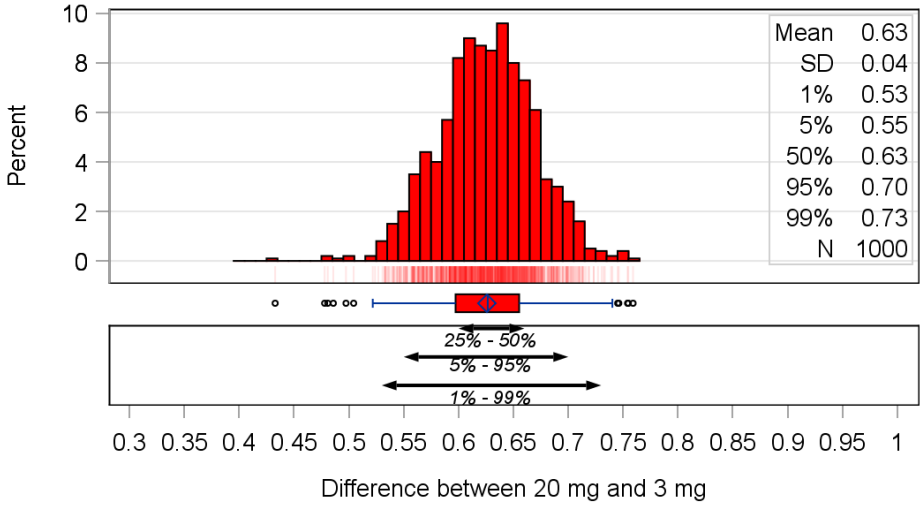
…and the difference between the 20 mg and the 3 mg dose levels

45

Sigmoidal Emax fit



Log-linear fit

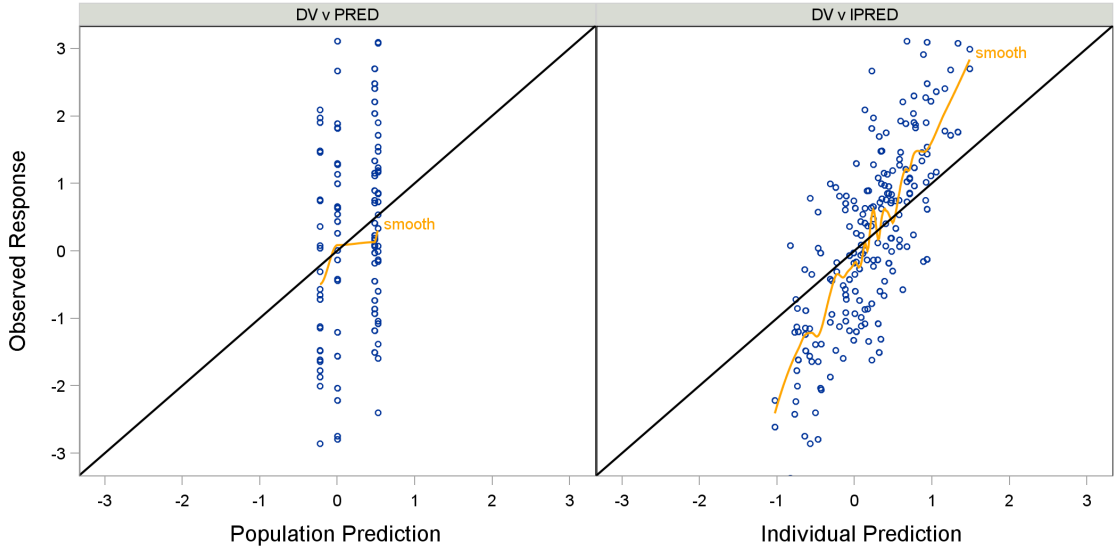


Bayesian fits for both models.

An example where the DV v PRED and DV v IPRED look wrong, but the model is correct!

46

Example 2: A simple linear mixed effect model



This plot looks poor, but the model is exactly right!

The simulated longitudinal model and the "very odd" model

47

Simulated Model (N=6 subjects for Placebo and Drug, Weeks 0, 2, 4, 6, 8)

Response = 10 + Sub_eff + Drug_eff + Error

Sub_eff	~ N (0,1)
Drug_eff	= 0 for Placebo, -1 for Drug (weeks 2, 4, 6, 8)
Error	~ N (0, 0.01)

Fitted Model (estimated parameters were Drug_eff and OCC_{SD})

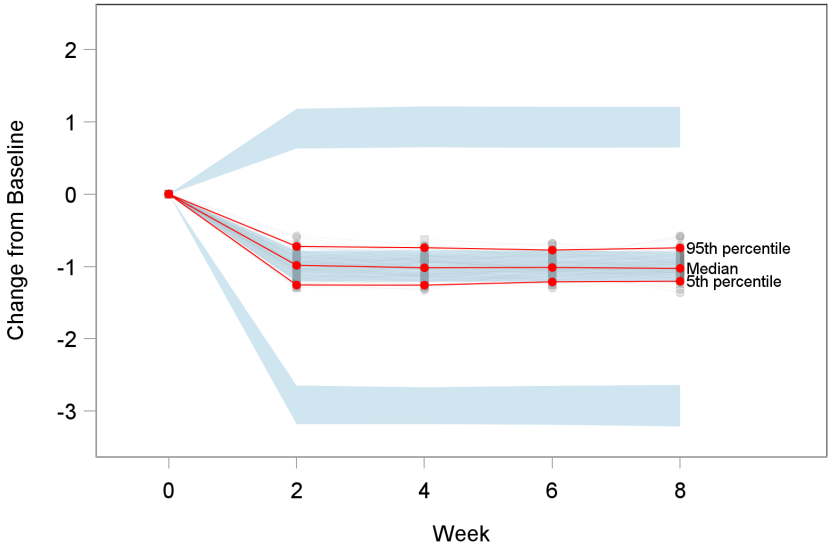
Response = $Base_{observed}$ + Drug_Model + Occasion + Error

Drug_Model	= Drug_eff + 1	for ID (1, 3, 5)
Drug_Model	= Drug_eff - 1	for ID (2, 4, 6)
Occasion	~ N(0, OCC^2_{SD})	for weeks 2, 4, 6 and 8
Error	~ N(0, 0.01) (fixed)	

48

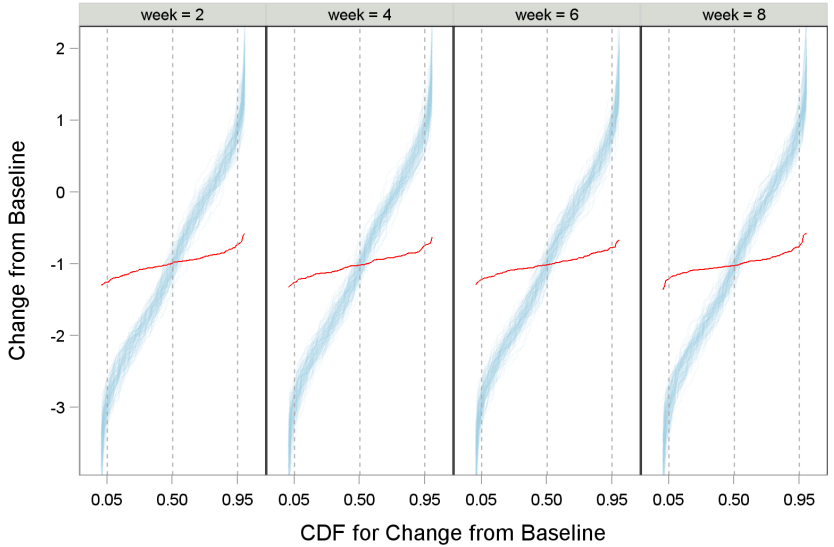
…better VPCs …a CDF example

Using Change from Baseline



The individual changes from baseline are all about -1, unlike the odd model !

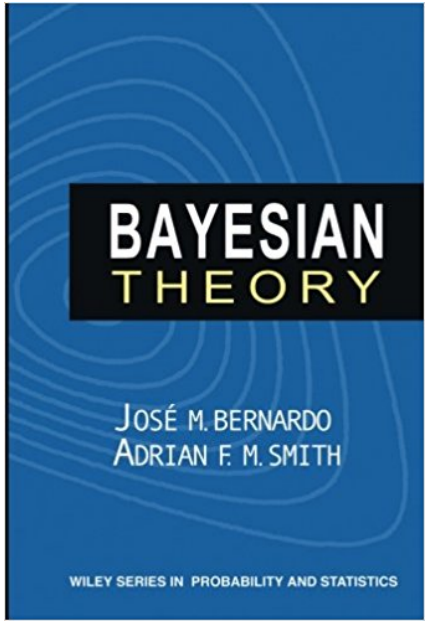
Using the cumulative density function

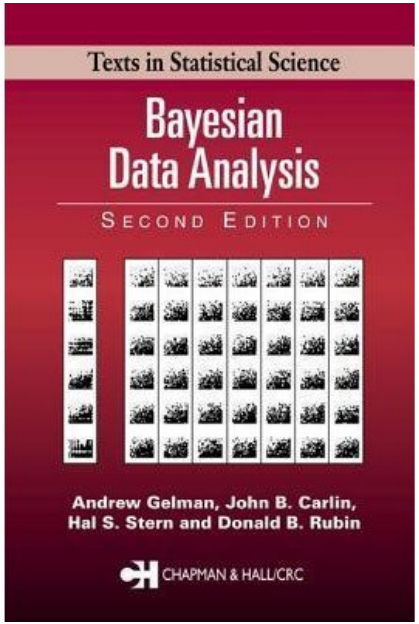


Shows the whole distribution (not just median and 5th / 95th percentiles)

It is interesting to note how the Bayesian community have ignored the NP bootstrap

49





NP Bootstrap – Example 1 – it doesn't have the correct coverage probabilities

50

True Rate	N	Observed	NP Bootstrap Percentiles 2.5%	97.5%	P (Observed)	Cumulative Totals
0.5	100	0	0.0	0.0	<0.000001	
0.5	100	1	0.0	0.3	<0.000001	
		...				1.76%
		...				
0.5	100	38	0.29	0.48	0.004473	
0.5	100	39	0.30	0.49	0.007111	
0.5	100	40	0.31	0.50	0.010844	
		...				
		...				
0.5	100	50	0.40	0.60	0.079589	96.48%
		...				
		...				
0.5	100	60	0.50	0.69	0.010844	
0.5	100	61	0.51	0.70	0.007111	
0.5	100	62	0.52	0.71	0.004473	
		...				1.76%
		...				
0.5	100	99	0.97	1.00	<0.000001	
0.5	100	100	1.00	1.00	<0.000001	

N=100 p(true) = 0.5

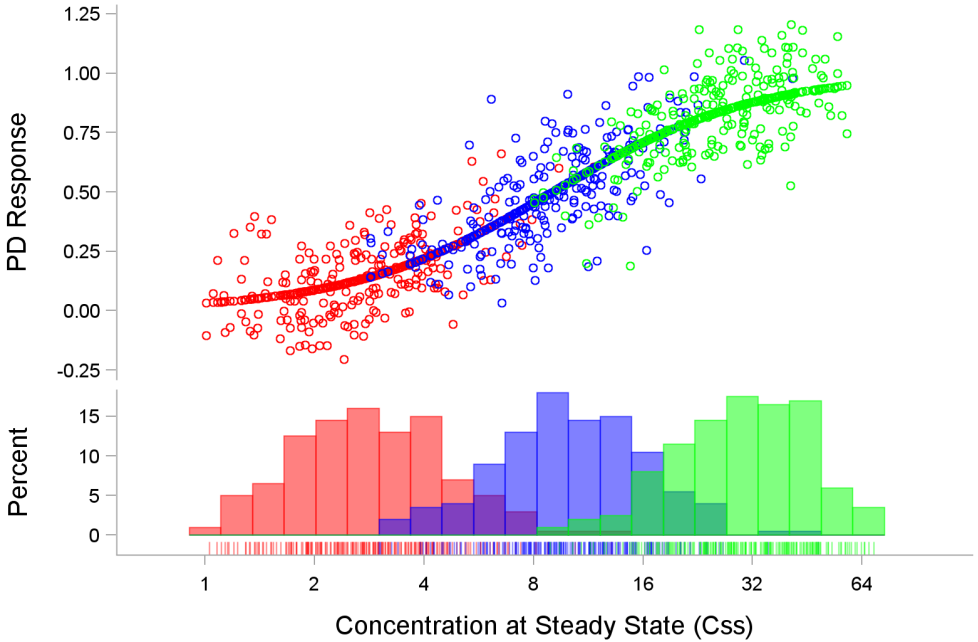
101 possible (observed) study outcomes.

If we observe between 40 to 60 events, the 95% NP bootstrap contains the true mean (0.5 (green area). Otherwise it will not (red area). Each red area occurs 1.76% instead of 2.5% of the time.

If repeat for different P, N or different intervals (1%, 5%, 10%, 20% etc. instead of 2.5%), there is no consistent pattern (sometimes conservative, sometimes anti-conservative)

NP Bootstrap – Example 2 – how to implement it correctly?

51



What stratification is correct?

Three dose levels (% female) are: 3mg (45%), 10mg (40%), 30mg (35%)

Exposure-Response with covariates (sex on EC50)

With a Dose-Response, we would stratify the sampling by Dose. But with exposure? Do we need to maintain the ‘same’ distribution of Css, or the same %Female/Male?

No ‘correct way’ – all choices seem ad-hoc.

52

…and finally, some quotes from Churchill

“To improve is to change, so to be perfect is to change often.”

“You have enemies? Good. That means you’ve stood up for something, sometime in your life.”