Decision-making using longitudinal modelling presence of model uncertainty

Longitudinal modelling: Time to take the next step?

Tobias Mielke

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Objective

„Making the best possible decision at the earliest time-point in the most efficient manner“

Increasing efficiency in drug development:

– Adaptive study designs, Biomarker, Portfolio optimization

What else to increase efficiency?

- Ensure that the collected data is actually utilized (e.g. longitudinal)
- Ensure that we leverage internal/external knowledge in our designs (e.g. MBMA, (Bayesian) modelling)

Implications:

- Do the right decision with higher certainty: Better informed decisions
- Do the same decision earlier: Quicker transition to next phase

Leveraging Longitudinal Data

Longitudinal data are **routinely** collected in our experiments

- e.g. body weight, viral load, blood pressure, heart rate, tumor growth, …

Often, primary analyses focus on AUC analyses, cross-sectional endpoint analyses or changes from baseline. Good practice to reflect on:

- The question of interest:
  - What is the study objective? Any difference between groups or specific difference?
  - What is the pattern of change? Is the pattern different over treatment groups?
    - ↳ Earlier onset iso size of effect as differentiator?
- Possible suboptimal use of information:
  - Do we use all measurements? Which assessments provide valuable information?
  - Do we use optimal analysis techniques? Dichotomization = loss of information!

What is longitudinal modelling?

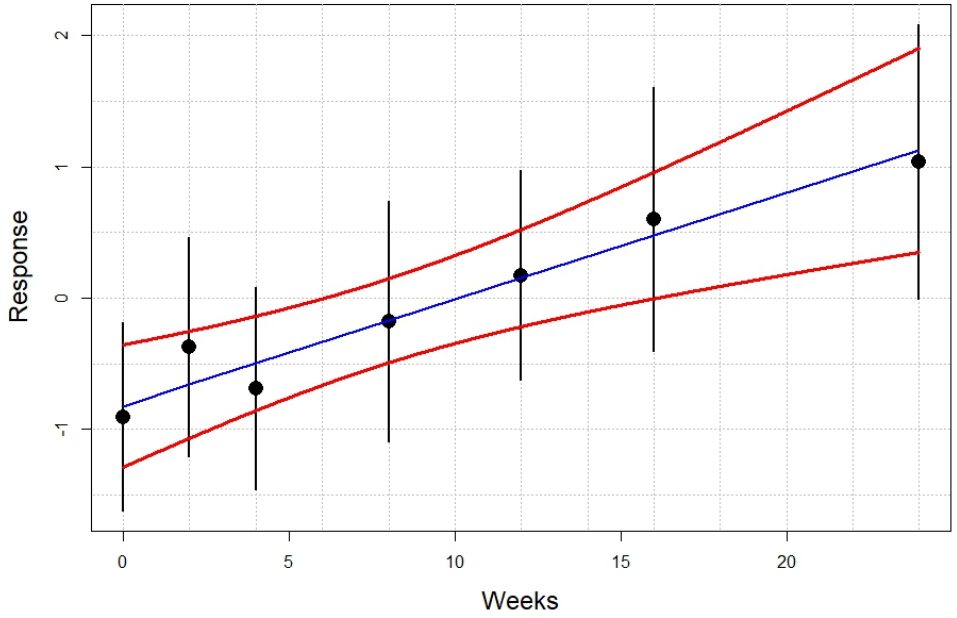
Variety of methods combining multiple assessments into the analysis via model

E.g. looking at binary endpoints:

- Direct modelling of transition probabilities:
  - What is probability that non-responder at week 12 becomes responder at week 24?
- Fitting generalized linear (mixed) models
  - Are dependencies appropriately taken into account?
- Modelling as time-to-event variable if only one “0”-> “1” transition possible
  - Timing of event to provide additional information
- Modelling of an underlying continuous model, driving “response”
  - Fit continuous endpoint and estimate response probabilities from continuous model => more info

Here focus on: Continuous endpoint and mixed effects models.

Why (longitudinal) modelling…



Too nice to be true…

Fitting linear model through 2 points:

- No benefit on these points
- … but for points in between: yes

Fitting linear model through 3 points:

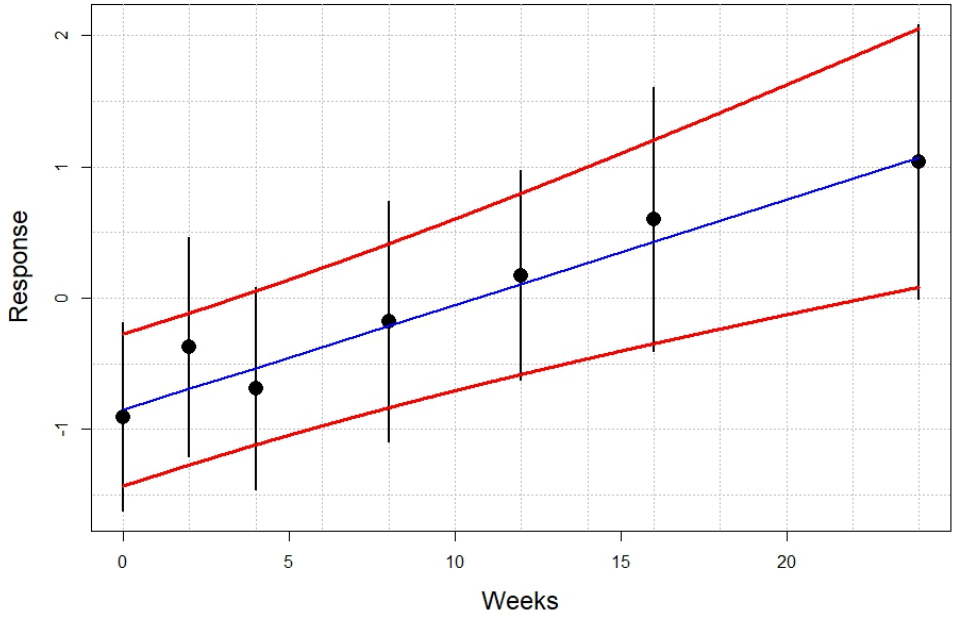
- Also CI on boundaries shrinks

… but best is to use all visits

Visits used in analysis	Efficiency	Required size	Equivalent size
2	100%	50	50
3	89.75%	45	55
4	73.57%	37	68
7	55.08%	27	90

… and what to consider: Correlation of observations

7 measurements on one subject ≠ 7 subjects with one measurement

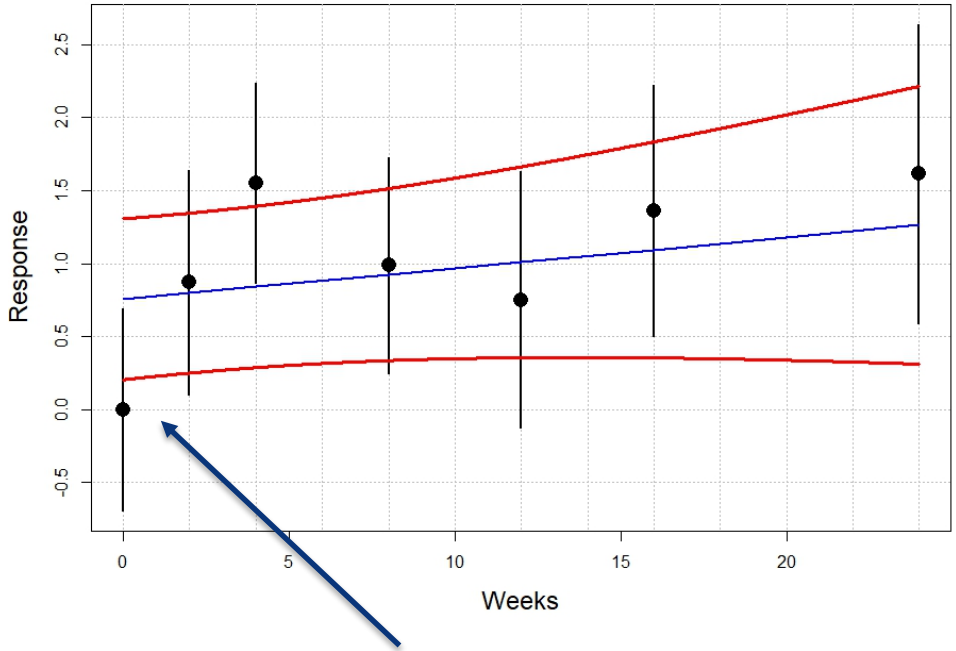


… 6 patients can make 1 month

- Observations from one experimental unit (e.g. patient) are correlated
- Benefit depends on correlation within subject:
  - High correlation => less learnings

Visits used in analysis	Efficiency	Required size	Equivalent size
2	100%	50	50
3	96.37%	48	52
4	92.45%	46	54
7	88.48%	44	56

… and what else to consider: Uncertainty on the model



Problems with wrong shape:

- Biased estimator
- False coverage probability

↳ Misleading results

Does this mean: Don't use modelling?

- No.
- Be cautious and aware of uncertainty.
- Include uncertainty in design evaluation.

Model-based CI doesn't include observed mean => "Too bad to be correct?"

The question of interest: What is the objective?

Apply statistical methodology to support efficient ***decision making***

What is the considered rule for decision making?

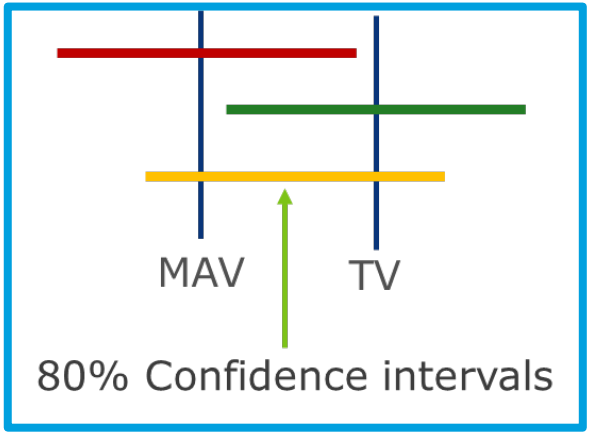
Situation 1

Is there any difference between control and test at any time?

H_0: $\cap_{i=1}^{T} \mu_i^1 \leq \mu_i^0$ at α_0

Situation 2

Do we reach a targeted effect size



Situation 3

What is the conditional or predictive power?

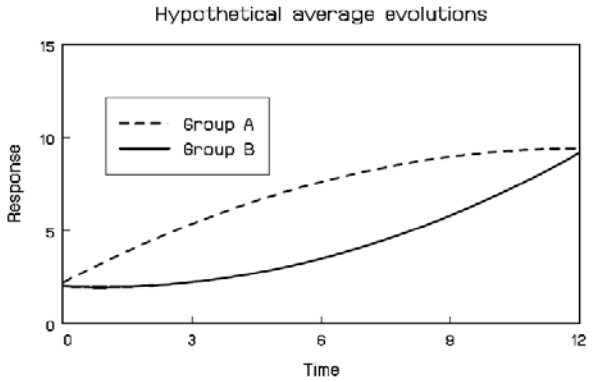
Probability of study/program success given data <x?

Situation 1: Test for any difference => Look at Björn's part

Is there any difference between control and test at any time?

- Summarize multiple assessments into one
- Optimal weighting of assessments:

$$P_\delta\left(\frac{c^T(\bar{Y}^1-\bar{Y}^0)}{\sqrt{2c^T\Sigma c}} > z_{1-\alpha}\right) = 1-\Phi\left(z_{1-\alpha}-\frac{c^T\delta}{\sqrt{2c^T\Sigma c}}\right)$$



Benefit: Higher power as compared to standard cross-sectional

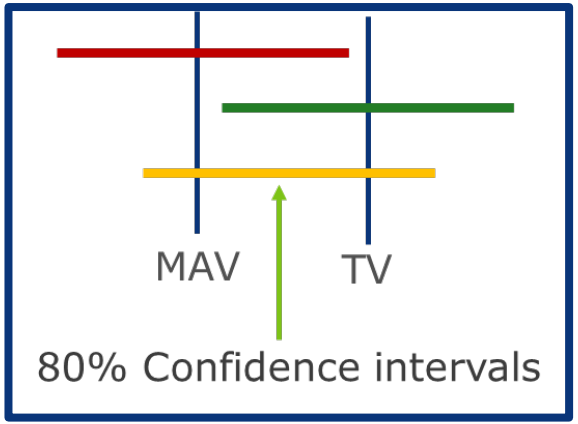
- ***Does not severely depend on distribution of data***

Limitation: No insight on size of effect.

Situation 2: Test for relevant difference $\Delta = \mu_T^1 - \mu_T^0$

Dual "Go – No go" approach (Lalonde 2007)

- Is the effect worth continued investment?
  - **"Stop"** if Δ significantly below some target effect at some α_{TV}
- Was a minimum relevant effect observed?
  - **„Go"** if Δ above max. non-acceptable effect at some level α_{MAV}



Properties/Problems:

1. Model-based approach: info is drawn from model ➔ **Error-control not guaranteed**
2. Non-model based approach: significant amount of data not utilized
3. Definition of Δ: Parameter in a longitudinal model or difference at certain time?

Situation 3: Conditional Power

Final analysis:

Test $H_0: \mu_T^1 \leq \mu_T^0$ *at level* α*=5% (two-sided), using standard methods*

Interim analysis: *What is the probability of a positive study at final analysis?*

$$Conditional\ Power = P_{\delta,\sigma}\left(\sqrt{w_1}z_1 + \sqrt{w_2}Z_2 + \sqrt{w_3}Z_3 > z_{1-\alpha} \middle| Z_1, Z_2^*\right)$$

δ, σ: Assumption on effects: Estimated from all data using parametric model

z_1: Completers at interim analysis: We know their data

Z_2: Some data available at interim analysis: We can predict how they develop

Z_3: No data available at interim analysis: We can estimate their effect

Situations 2 & 3: How does it work?

1. **Assume a parametric model, e.g. (N)LME:**

$$Y_{ij} = \eta(\beta_i, t_j, trt) + \epsilon_{ij}, \beta_i \sim N(\beta, D), \epsilon_{ij} \sim N(0, \sigma^2), i = 1, \ldots, N, j = 1, \ldots, m$$

   - Linear / non-linear?
   - Interpretation? $1 + trt \times time$ vs. $1 + trt + trt \times time$?

2. **Given data, fit model:**
   - Estimates for: $\beta, D, \sigma^2, \beta_i$ together with measures of uncertainty (FIM)

3. **Calculate decision metrics of interest:**
   - Model based confidence interval (simple, using estimates of β, D, σ^2)
   - Conditional power (a bit trickier, using estimates of $\beta, D, \sigma^2, \beta_i$ and actual data)

Impact of model uncertainty: Impacts of wrong model

Test for any difference: No loss of validity

- Loss in power: Likely still better than cross-sectional approach

Test for relevant difference: Loss of validity

- Biased estimator results in wrong coverage probability
- Short confidence intervals are possible, suggesting wrong stop/go decisions

Conditional power: Potentially misleading results

- Biased estimator and model lead to biased estimates of conditional power

Some strategies for mitigating model uncertainty

1. **Use flexible model**
   - MMRM: While validity not harmed, no gain in efficiency
2. **Fit several models to the data and pick the best***
   - Not best model in terms of “maximum effect”: Standard multiplicity problem
   - Best model in terms of best fit: Underestimation of variability => false coverage probability
3. **Fit several models to the data and conduct model averaging*:**
   - Similar to “2”, but characteristics potentially better
4. **Use of longitudinal modelling as back-up only**

Given all these complexities: What is the opportunity space?

What is the potential benefit?

Example: Phase 3 study designs in IPF

Based on publication:

*"Efficacy & Safety of Nintedanib in Idiopathic Pulmonary Fibrosis"**

Primary endpoint:

- Annual rate of decline in FVC (measured in milliliters per year)
- Assessments at weeks: 0, 2, 4, 6, 12, 24, 36, 52 and 56

Secondary endpoint: Absolute change from baseline at week 52

Analysis: RCR model with random intercept and slopes

Sample size: 90% power to detect…

- between group difference of 100ml in annual rate of decline
- … assuming standard deviation of 300ml on change from baseline
- … and 3:2 randomization

Richeldi et al., NEJM 2014, 370:2071-2082

Example: Phase 3 study designs in IPF

*"Efficacy & Safety of Nintedanib in Idiopathic Pulmonary Fibrosis"**

Sample size planned looking only at change from baseline week 52

… no need to make assumptions on within-subject correlation

… more assessments: information will increase => conservative sample size

What we'll look at instead:

1. What would be the power using all assessments?

   … and what is the impact of variance composition on operating characteristics?

2. How would deviations from assumed linear model impact power?

   … and how would model selection/averaging compare

Richeldi et al., NEJM 2014, 370:2071-2082

Example: Phase 3 study designs in IPF

*What would be the power using all assessments? (*0, 2, 4, 6, 12, 24, 36, 52)

Looking at standard LME:

$$Y_{ij} = \beta_{i1} + (\beta_{i2} + trt \times \beta_{i3})t_{ij} + \epsilon_{ij} \text{ with } \beta_i \sim N(\beta, D) \text{ and } \epsilon_i \sim N(0, \Sigma)$$

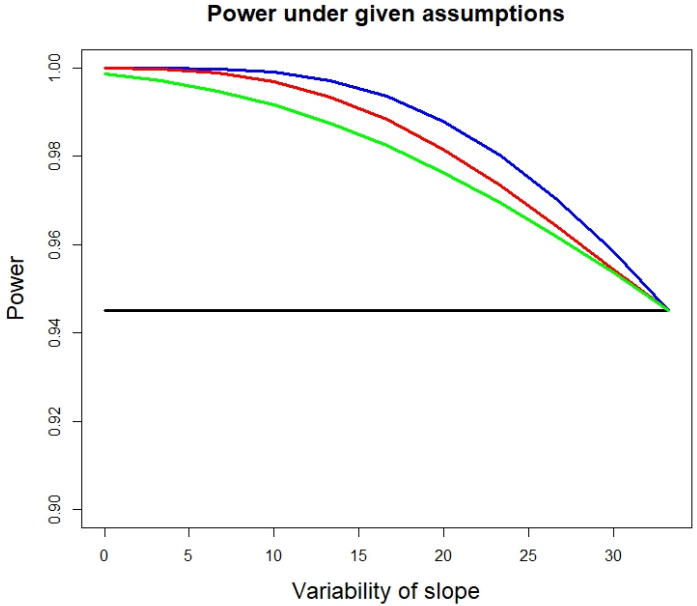
Assumption on change from baseline: $c^T Y_i \sim N(\mu, \sigma^2_{CFB})$ $(c = (-1, 0, \ldots, 0, 1)^T)$

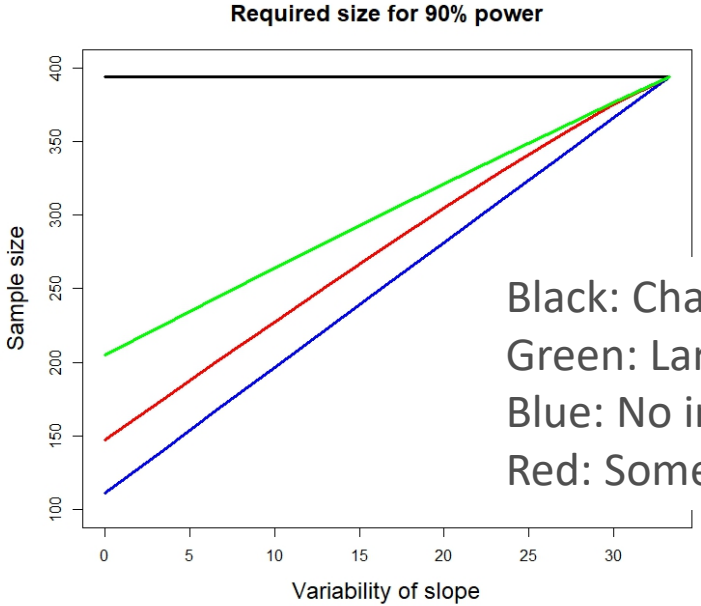
- Assuming no intra-subject correlation: $\Sigma = \sigma^2_\epsilon I_m$
- Assuming no random treatment effects $(Var(\beta_{i3}) = 0)$
- … resulting assumption on variability: $\sigma^2_\epsilon = (\sigma^2_{CFB} - t^2_m Var(\beta_{i2}))/2$

Given Unknown

Example: Phase 3 study designs in IPF

- *What would be the power using all assessments?*





Black: Change from baseline
Green: Large intercept variability
Blue: No intercept variability
Red: Something in between

Result: potentially overly conservative study design

Example: Phase 3 study designs in IPF

2. How would deviations from assumed linear model impact power?

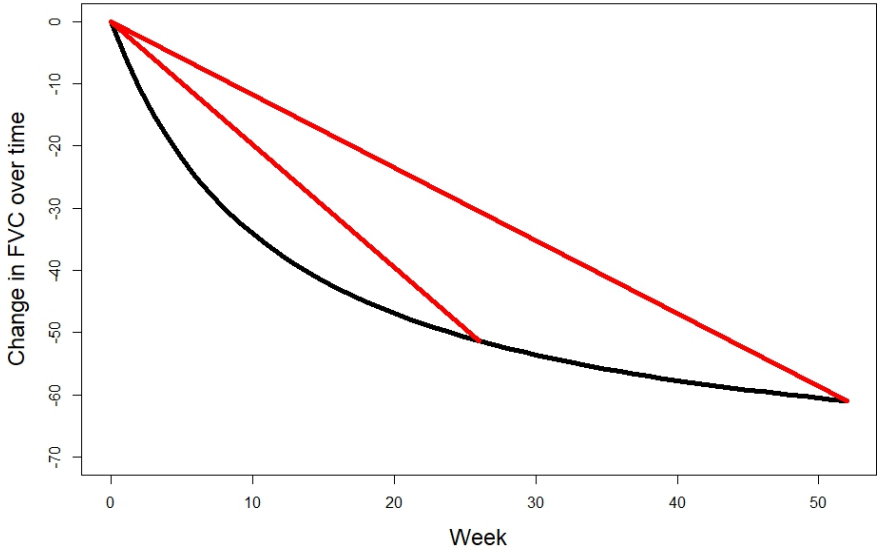
What is the annual rate of decline?

- **Linear progression: simple**
- **Non-linear progression: depends on study duration**

Looking at landmark (e.g. estimate on wk52 change from baseline informed from longitudinal model) may resolve this problem:

$$\mathbf{H_0}: \boldsymbol{\Delta = \mu_{52}^1 - \mu_{52}^0 \leq 0}$$

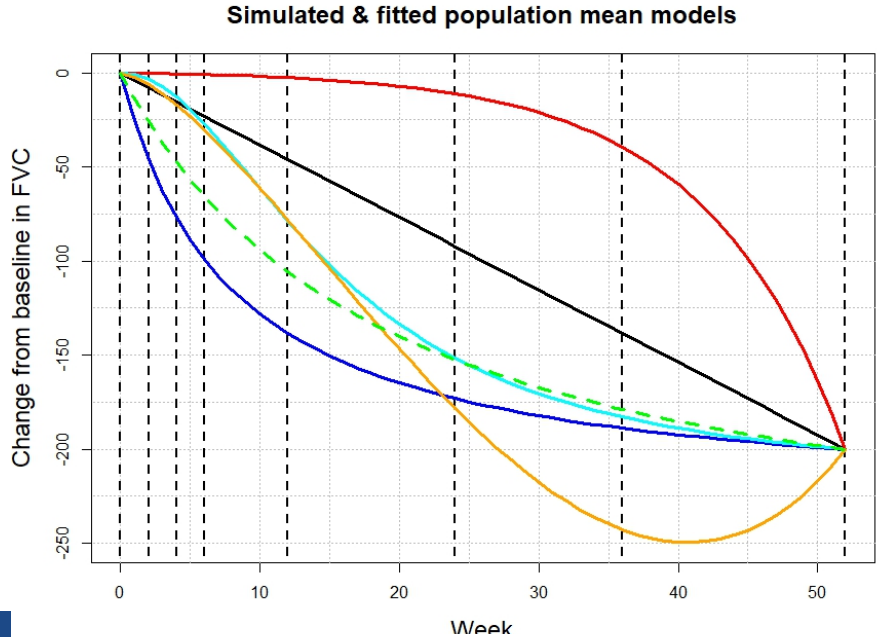
… but will be different endpoint.

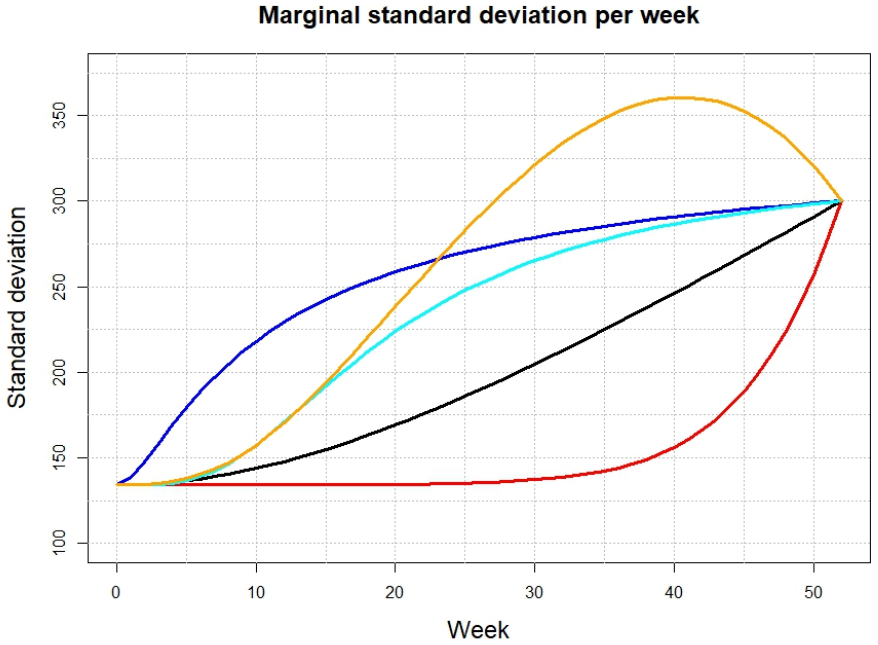


Example: Phase 3 study designs in IPF

2. How would deviations from assumed linear model impact power?

Simulated & fitted scenarios:





Note: linear models were fitted (but non-linear in time)!

Example: Phase 3 study designs in IPF

2. How would deviations from assumed linear model impact power?

Coverage probabilities of 95%-confidence interval:

		Fitted shape						
		Linear	Emax	Exponential	Sigmoidal	Beta model	Selection (1)	Selection (2)
Simulated shape	Linear	95,10%	83,13%	88,57%	91,77%	79,70%	95,10%	91,77%
	Emax	78,03%	94,47%	45,30%	76,10%	58,93%	94,47%	76,10%
	Exponential	88,03%	48,23%	95,13%	59,33%	41,97%	95,13%	88,03%
	Sigmoidal	94,10%	90,80%	73,47%	94,67%	92,17%	94,67%	92,17%
	Beta model	92,40%	93,20%	70,63%	93,60%	94,87%	94,87%	93,60%
	Emax (2)	94,00%	94,47%	53,97%	92,00%	78,30%	93,23%	93,23%

- Selection (1/2): Simulated shape available / not available for selection
- Model averaging: performs like model selection as difference in profiles detected
- Model selection can mitigate some concerns, but not all (need adequate model)

Example: Phase 3 study designs in IPF

2. How would deviations from assumed linear model impact power?

Type 1 error:

		Fitted shape						
		Linear	Emax	Exponential	Sigmoidal	Beta model	Selection (1)	Selection (2)
Simulated shape	Linear	2,40%	2,60%	2,70%	2,60%	2,60%	2,40%	2,60%
	Emax	2,40%	2,30%	1,10%	2,30%	2,60%	2,30%	2,30%
	Exponential	2,20%	2,50%	2,00%	2,60%	3,20%	2,00%	2,20%
	Sigmoidal	2,70%	2,70%	3,20%	2,60%	2,90%	2,60%	2,90%
	Beta model	2,40%	2,00%	1,80%	2,50%	2,50%	2,50%	2,50%
	Emax (2)	2,20%	2,30%	2,70%	2,00%	1,80%	2,50%	2,50%

Type-1 error less of an issue here:

- Model selection not conducted to establish some effect, but to get best fit.
- Within each analysis model, the type-1 error is controlled

Example: Phase 3 study designs in IPF

2. How would deviations from assumed linear model impact power?

Power (assuming effect of 100):

		Fitted shape						
		Linear	Emax	Exponential	Sigmoidal	Beta model	Selection (1)	Selection (2)
Simulated shape	Linear	96,10%	91,70%	89,50%	95,90%	95,80%	96,10%	95,90%
	Emax	83,60%	96,90%	41,60%	87,70%	86,30%	96,90%	87,70%
	Exponential	94,60%	88,60%	96,80%	94,80%	94,00%	96,80%	94,60%
	Sigmoidal	96,60%	91,70%	80,30%	96,80%	96,70%	96,80%	96,70%
	Beta model	96,70%	90,20%	71,30%	97,30%	97,30%	97,30%	97,30%
	Emax (2)	97,70%	98,90%	82,60%	98,10%	97,80%	98,30%	98,30%

Power affected by model:

- Wrong model selected for fitting => impact on power
- True EMax but different model fitted => effect underestimated => lower power

Summary

Longitudinal modelling:

- … may add substantial efficiency, allowing earlier decision making
- … but requires more specifications/assumptions in design stage (e.g. variability)

Wrong model may lead to incorrect conclusions:

- … but approaches like model averaging may yield robust and efficient results
- … worthwhile at least for internal decision making

And a last thought:

- Do we really need the right model?
- Estimator from wrong model may still be informative due to numerics:
  - E.g. estimator of effect at final visit vs. maximum information from final visit in linear model

Summary: Uncertainty on the model ≠ don't try to utilize modelling

References

References:

Van Lancker, K., Vandebosch, A., Vansteelandt, S., de Ridder, F. (2019), "Evaluating futility of a binary clinical endpoint using early read-outs" *Statistics in Medicine*, **38**: 5361-5375

Buatois, S., Ueckert, S., Frey, N., Retout, S., Mentré, F. (2018), "Comparison of Model Averaging and Model Selection in Dose Finding Trials Analyzed by Nonlinear Mixed Effects Models", *The AAPS Journal, 20:56*

Lalonde, R.L., Kowalski, K.G., Hutmacher, M.M., Ewy, W., Nichols, D.J., Milligan, P.A., Corrigan, B.W., Lockwood, P.A., Marshall, S.A., Benincose, L.J., Tensfeldt, T.G., Parivar, K., Amantea, M., Glue, P., Koide, H. And Miller, R., (2007), "Model-based Drug Development". *Clinical Pharmacology & Therapeutics,* **82:** 21-32

Richeldi L., du Bois, R., Raghu, G., Azuma, A., Brown, K., Costabel, U., Cottin, V., Flaherty, K., Hansell, D., Inoue, Y., Kim, D., Kolb, M., Nicholson, A., Noble, P., Selman, M., Taniguchi, J., Brun, M., Le Maulf, F., Girard, M., Stowasser, S., Schlenker-Herceg, R., Disse, B., Collard, H., INPULSIS Trial Investigators (2014), "Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis" *New England Journal of Medicine*, **29**: 2071-2082