Upcycling development data to save experimental resources

Jens Lamerz, Tony Pourmohamad

Nonclinical Statistics Conference 2024
Wiesbaden, Germany

September 25, 2024

Idea:

Upcycling development data to save experimental resources.

Background

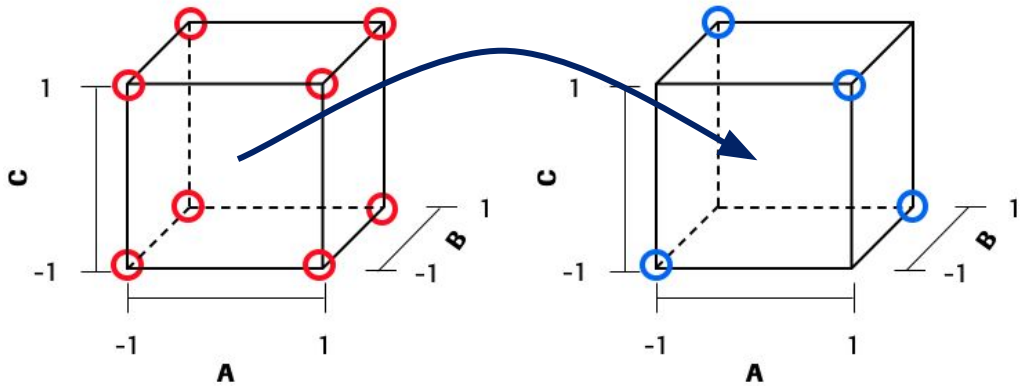
- The Technical Team develops a tablet with dose strength **x**.
- The Clinical Team says ‘thank you’, but now requests a new dose strength **y**.

- Then the Technical Team sets up a new product with same API, dose proportional, ‘same’ process except different tablet weight and diameter, for new dose strength **y**.

- Can we upcycle existing development data to save a relevant number of batches in a new process development project?

Background

- If your first dose strength shows some factors are inactive, you can set up smaller DoE's, and if on top of that two way interactions are inactive, you can save a lot of runs by switching from a resolution V DoE (left panel) to a resolution III DoE (right).



Established Formulation & Process
Dose Strength x mg
Resolution V DoE
e.g. Full Factorial, n= 8 + 3 = 11 runs

Verification
New Dose Strength y mg
Resolution III DoE
e.g. Fractional Factorial, n = 4 + 3 = 7 runs

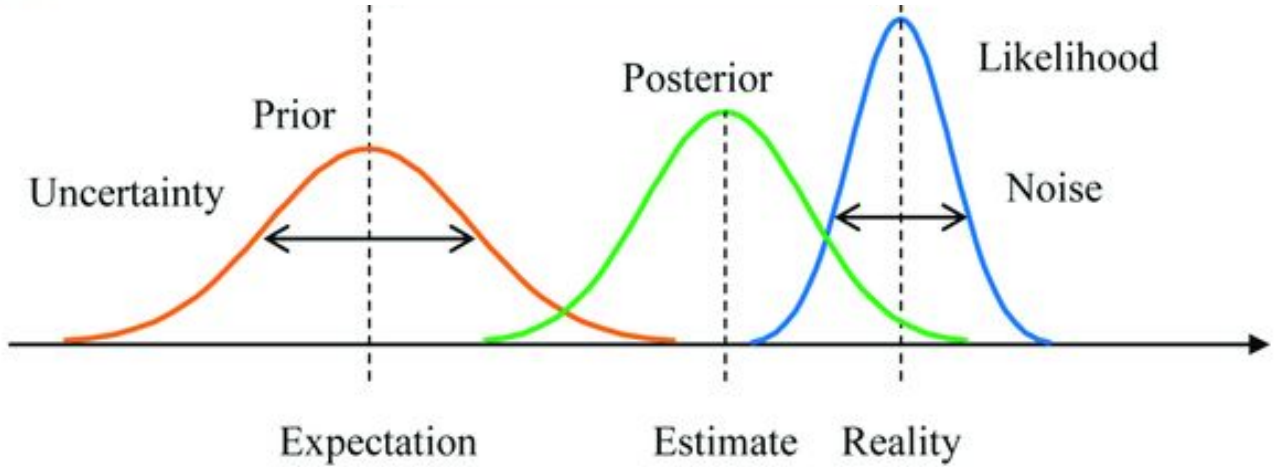
Background

- The reduction of runs comes from downgrading resolution V (e.g. full factorial) to resolution III (e.g fractional factorial).
- How to compensate that the new dose strength is based on much fewer new runs?
- Can we benefit assuming the process for old and new processes are consistent?
- **Note: Upcycling data requires a robust process and analytics, and consistency between old and new process.**

Quick Bayes Recap

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int_\Theta f(x|\theta)p(\theta)d\theta}$$
$$\propto f(x|\theta)p(\theta)$$

- The posterior distribution is proportional to the product of the prior distribution and the likelihood function
- The prior and likelihood both contribute to the final shape and central tendency (location) of the posterior



Our current strategy for specifying the prior

- For example, the posterior associated with the **h**istorical data is the following:

$$p(\beta, \sigma^2 | X_h, y_h) \propto N(y_h | X_h\beta, \sigma^2 I) \times \frac{1}{\sigma^2}$$

- We make use of the following factorization to sample from the posterior distributions

$$p(\beta, \sigma^2 | X_h, y_h) = p(\beta | \sigma^2, X_h, y_h) p(\sigma^2 | X_h, y_h)$$

Normal Distribution

Inverse-Gamma Distribution

- Can use the posterior of the **h**istorical data as the prior for the **c**urrent data, i.e.,

$$p(\beta, \sigma^2 | X_c, y_c) \propto N(y_c | X_c\beta, \sigma^2 I) \times p(\beta | \sigma^2, X_h, y_h) p(\sigma^2 | X_h, y_h)$$

Posterior Distribution of Current Data

Likelihood of Current Data

Prior distribution which is the posterior distribution of the historical data

Why use this prior elicitation strategy?

- We are able to leverage historical data in a noninformative way (i.e., reduce subjectivity of the prior)
- Under a noninformative prior, the Bayesian inferential results for the historical data match the frequentist inferential results
- It's a simple strategy
  - All calculations are straightforward
  - Closed form distributions for all quantities of interests
  - Sampling from the posterior and posterior predictive distributions is very simple

Available Data Sets

Sameinib and Diffimod are fictitious drug names

Sameinib 4 mg*
Full Factorial 8 + 3 CP =

Half Fractional 4+2 CP

+

Half Fractional 4+1 CP

Sameinib 7 mg
Full Factorial 8 + 3 CP =

Half Fractional 4+1 CP

+

Half Fractional 4+2 CP

Diffimod 200 mg**

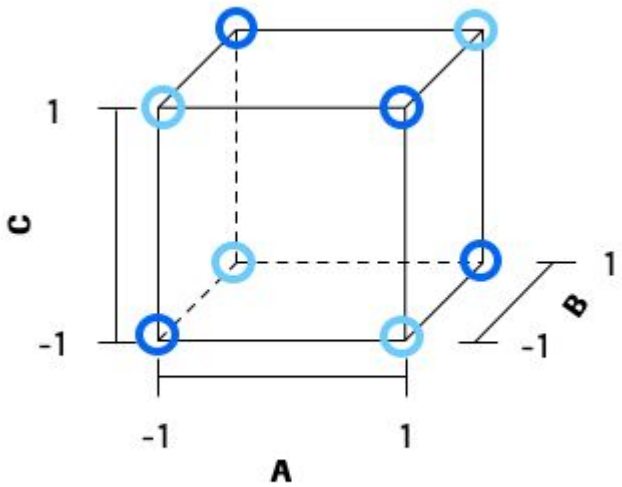
Examples:

* best case DoE, super formulation

** not so good role model formulation

7 mg Sameinib Full Factorial is split into *two* ½ Fract. Factorial subsets

1st subset:
used for model



2st subset:
withheld, used for verification

Sameinib 7 mg
Full Factorial 8 + 3 CP =

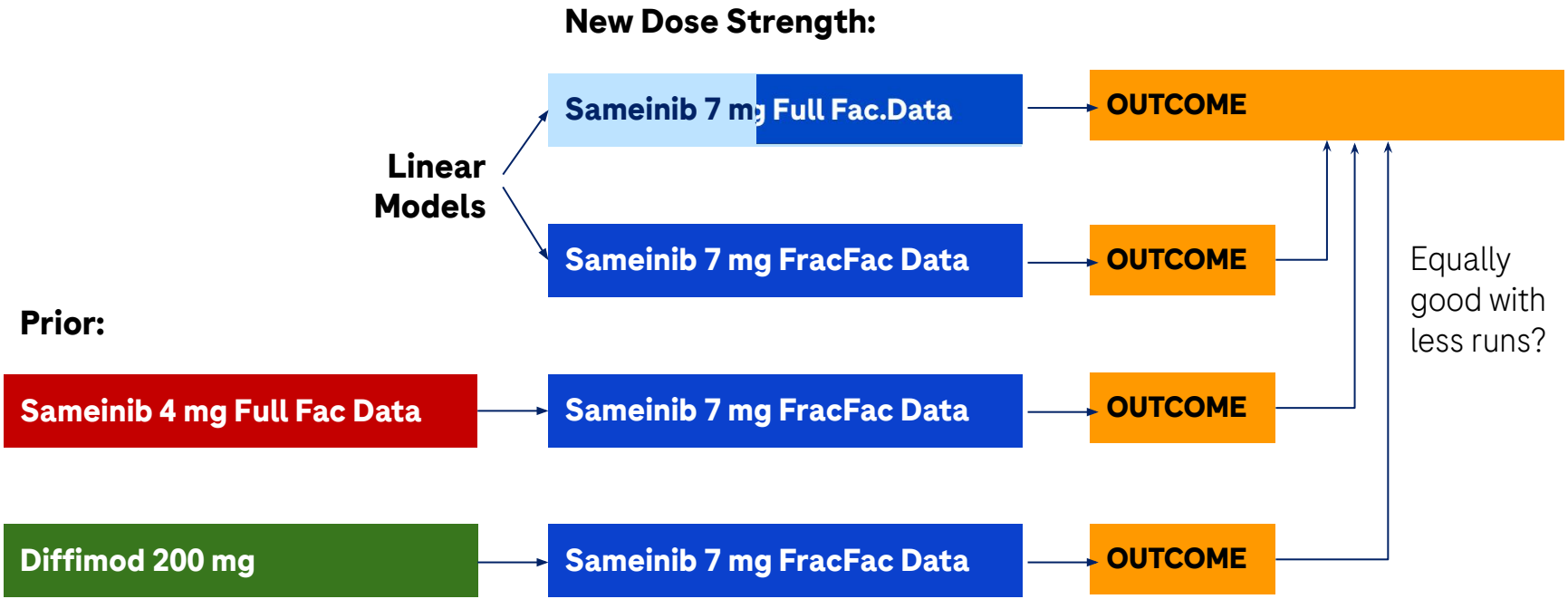
Half Fractional 4+1 CP

+

Half Fractional 4+2 CP

‘Quality’ of Sameinib 7 mg Model

Model Responses: Tablet Core Hardness & Dissolution 15 min



The Statistical Model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Factors: one MA, two PP's

x_1 = `API PSD d50 [um]`,

x_2 = `Comp. Press. [MPa]`,

x_3 = `Dwell Time [ms]`,

Responses: two IPCs, two CQA

y_1 = `Hardness [N]`,

y_2 = `SF [%]`,

y_3 = `Disint. Mean [sec]`,

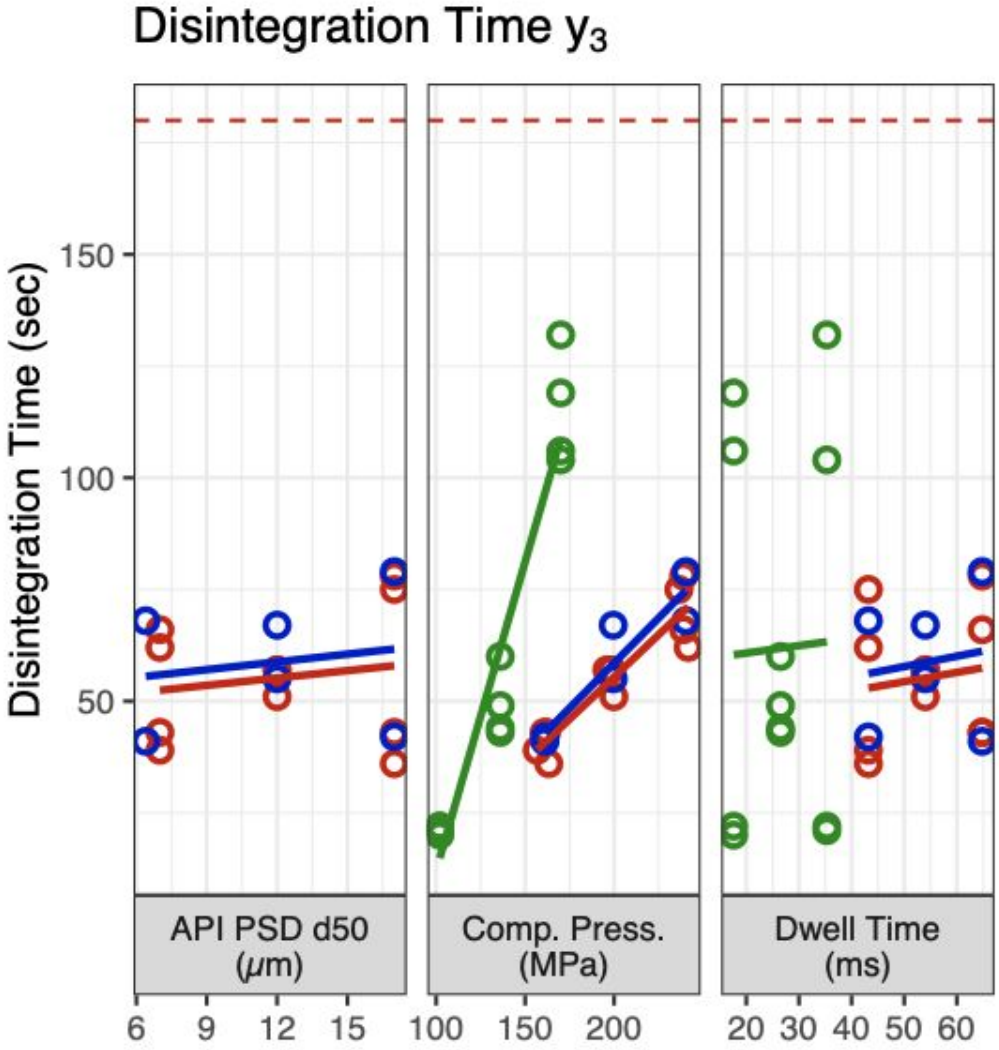
y_4 = `Diss. 15 min Mean [%]` ← see Appendix

Notes

The model for Diffimod model is slightly different as it does not include x_1

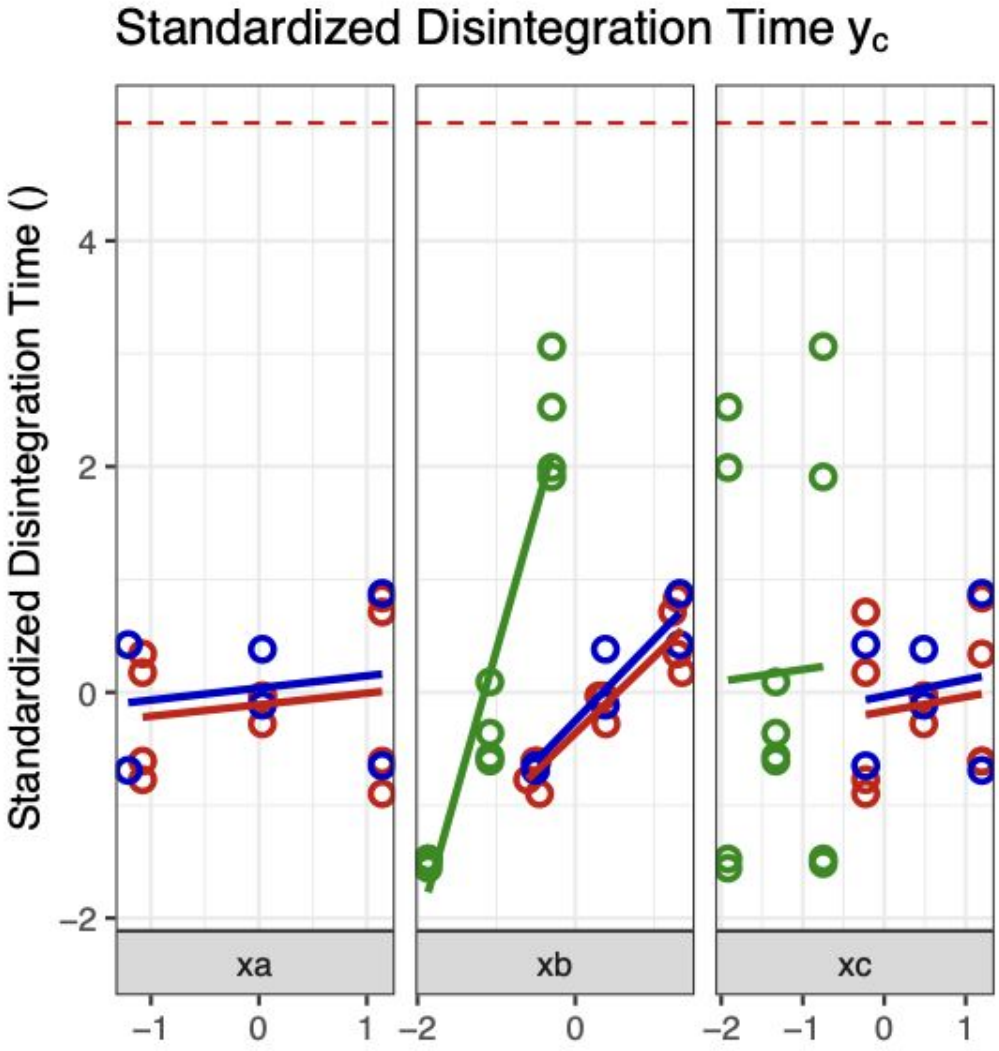
Arabic Numbers in subscript refer to data as is, e.g. y_1 = 180 sec

lowercase latin letters in subscript refer to mean centered and std. data, e. y_a= 5.0



Source

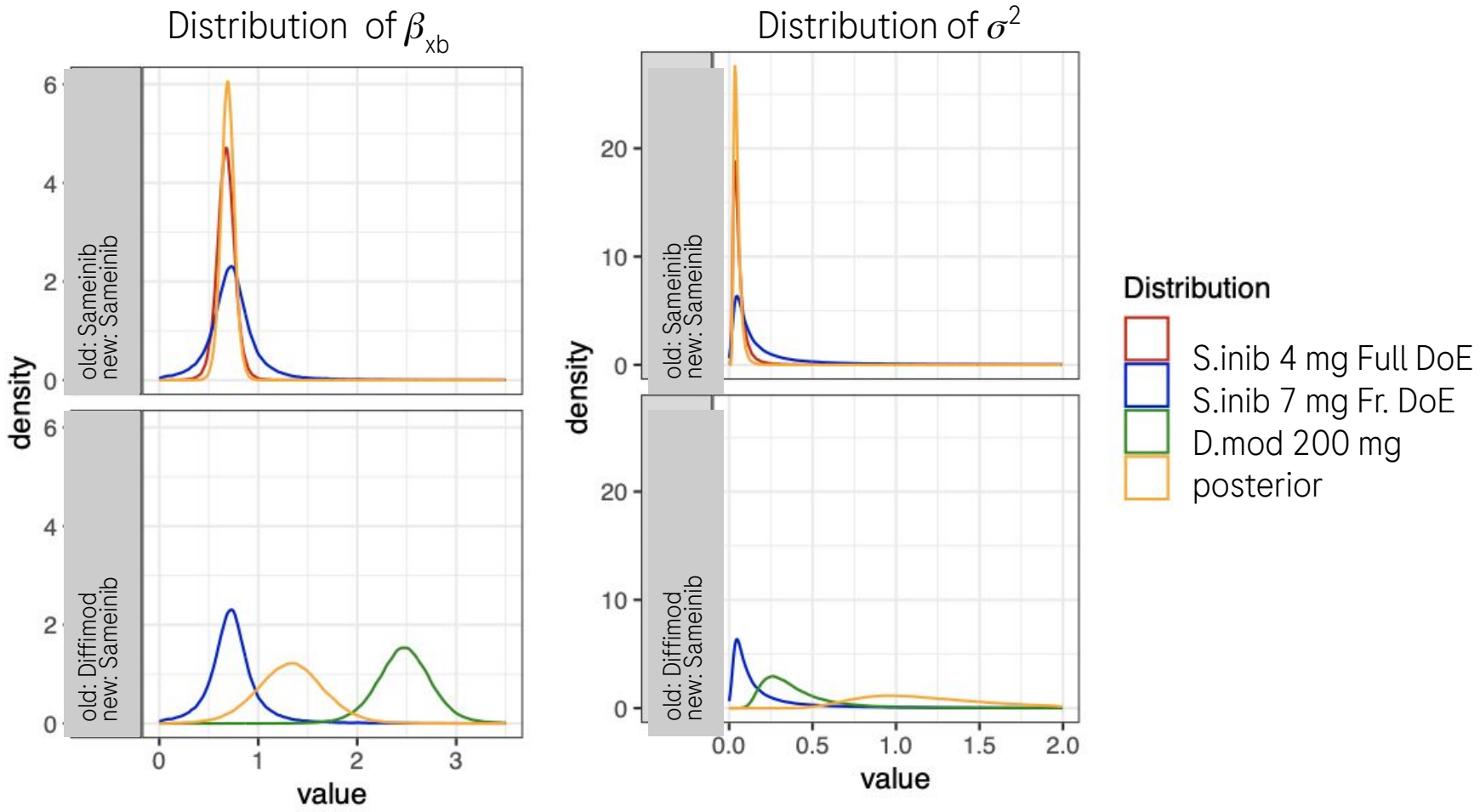
7 mg Sameinib ½ Frac. DoE
4 mg Sameinib Full DoE
200 mg Diffimod



Source

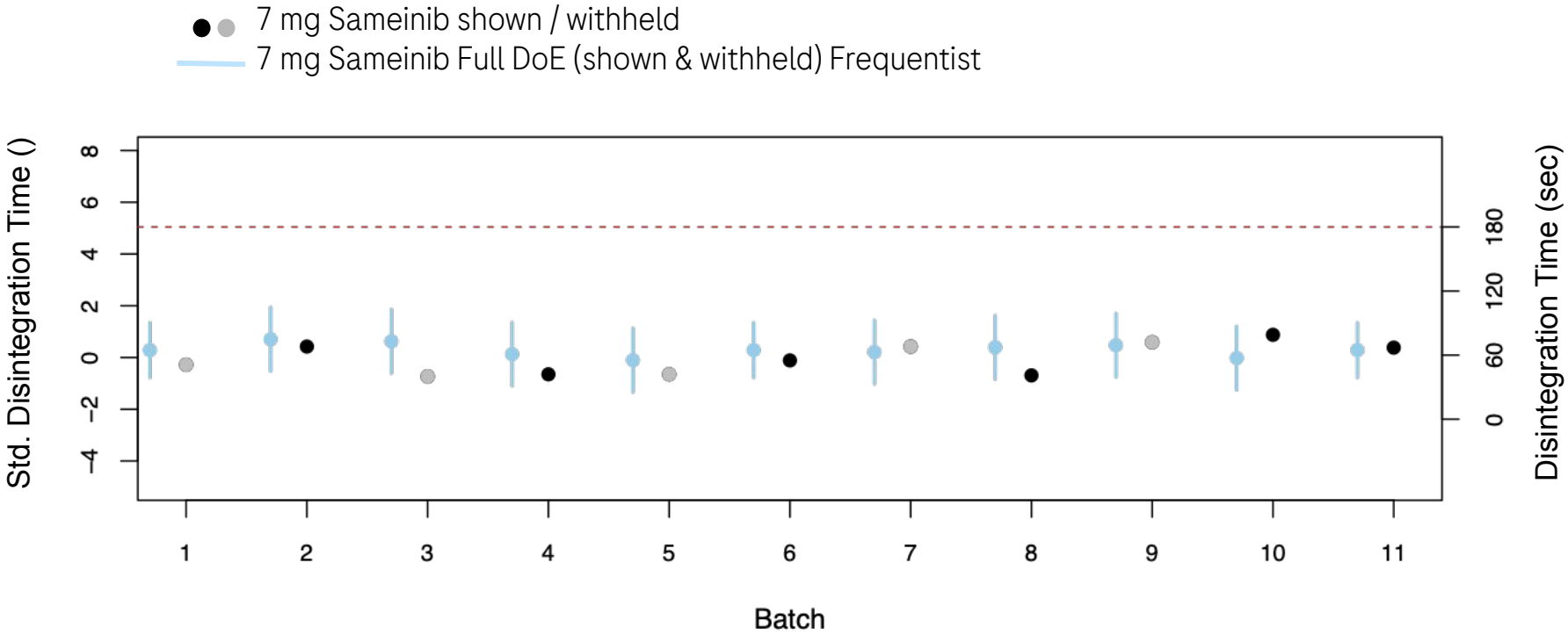
7 mg Sameinib ½ Frac. DoE
4 mg Sameinib Full DoE
200 mg Diffimod

y_c - std. Disintegration Time



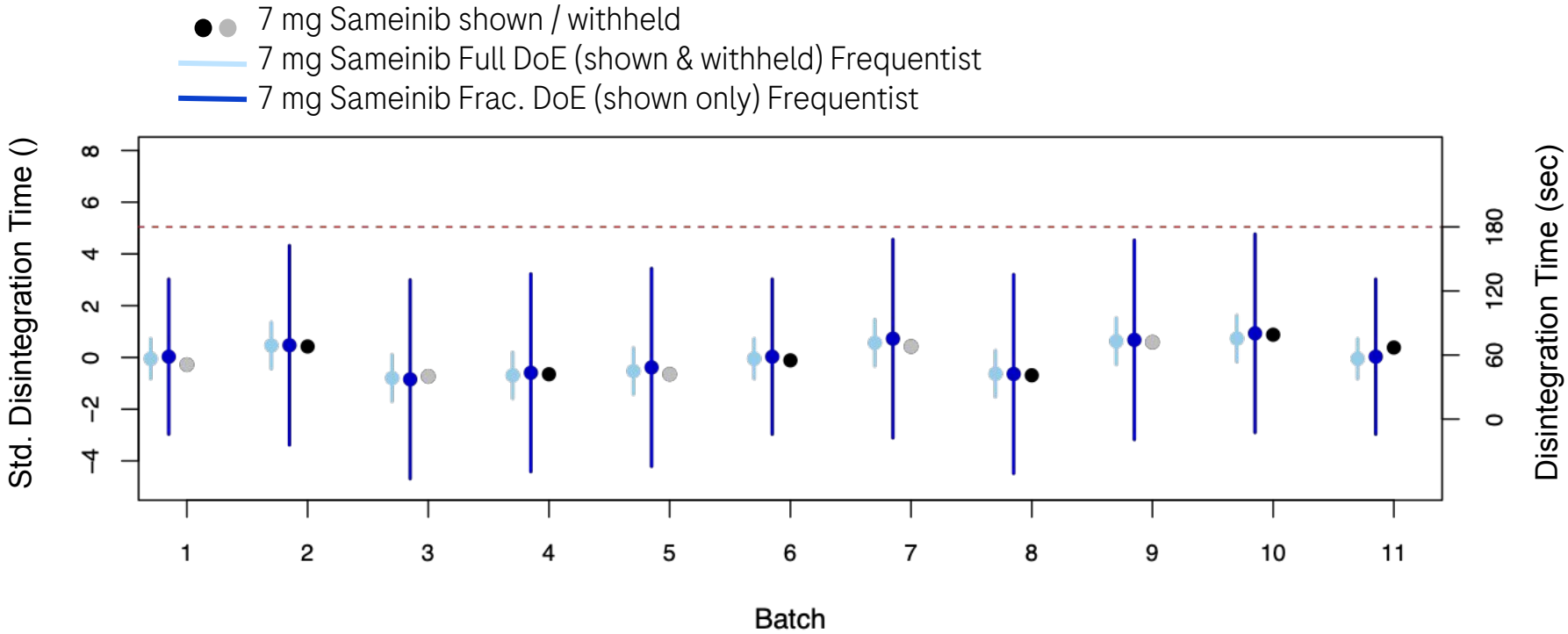
7 mg Full Factorial Data, Frequentist Model

This would be the model if we reinvent the wheel every time we have a new dose strength with n=11 runs



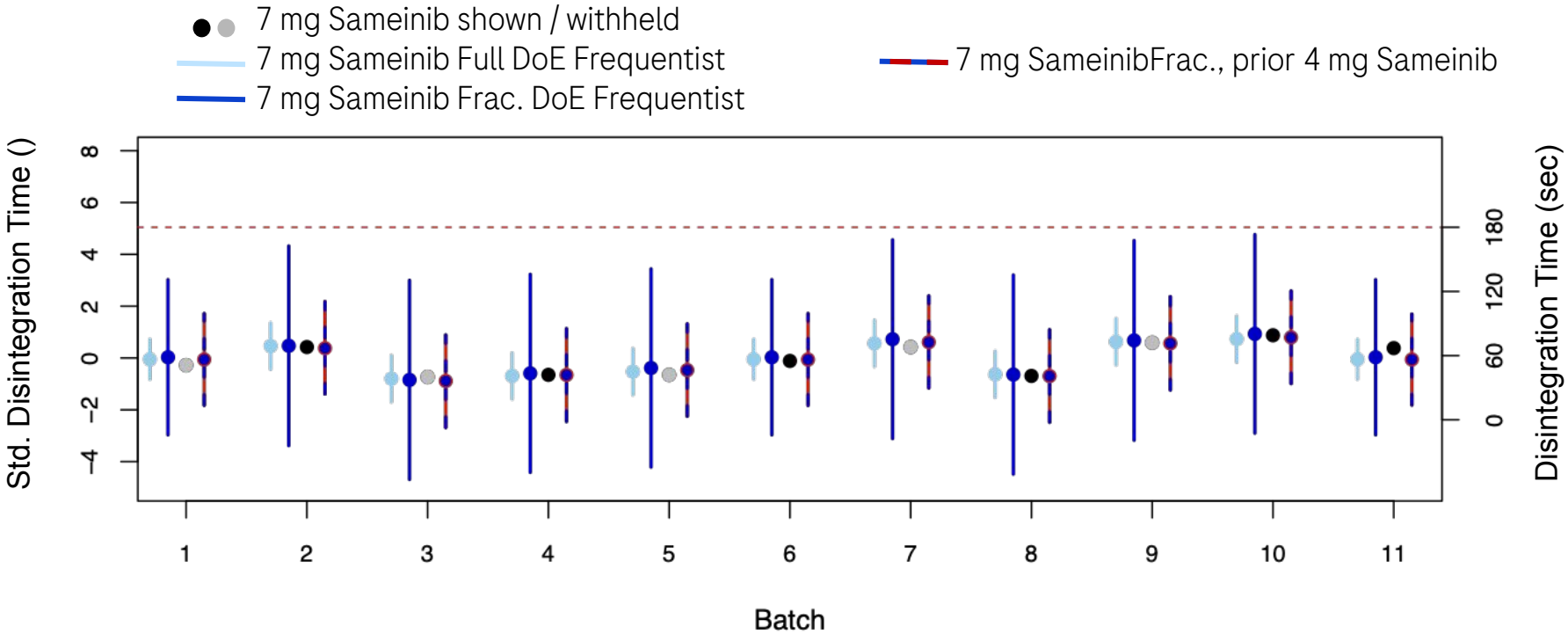
7 mg Fractional Factorial Data, Frequentist Model

The reduced # of runs saves resources, but leads to higher uncertainty = wider pred. intervals.



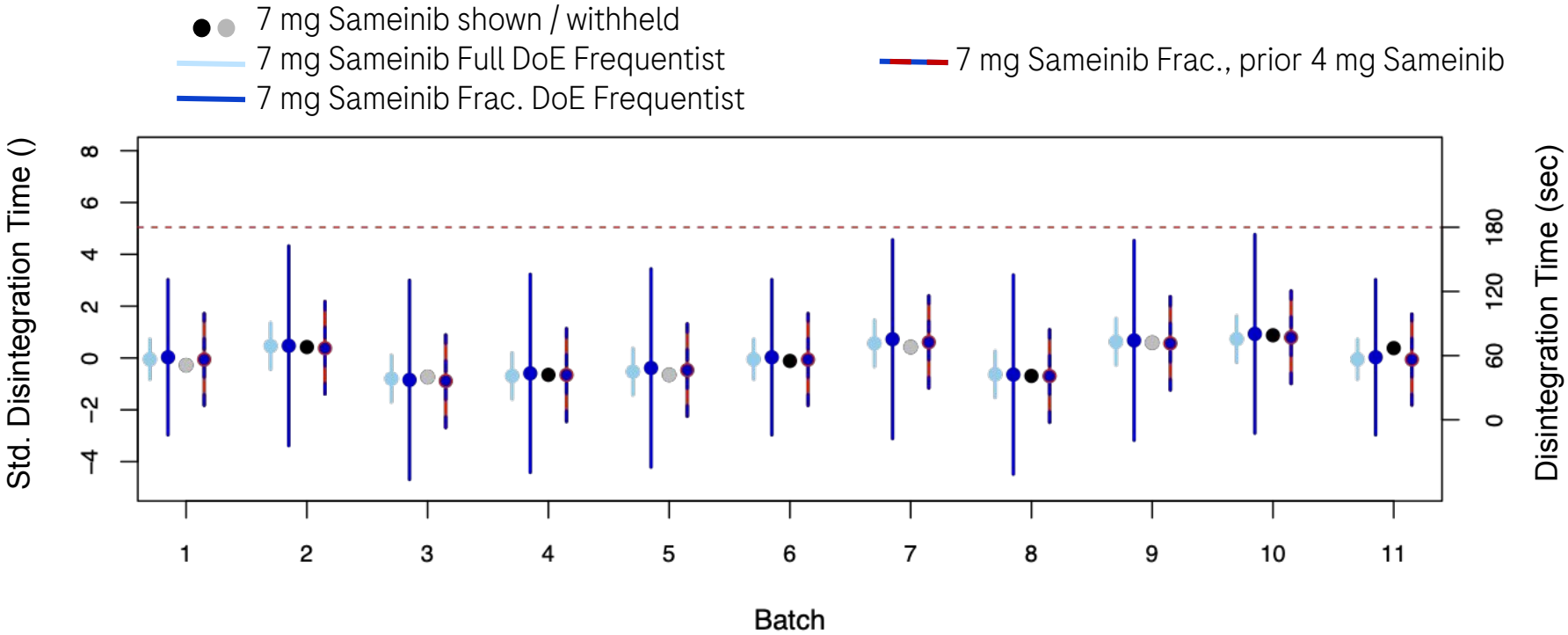
Bayes Model with 4 mg data for prior and 7 mg fractional data

Much smaller prediction intervals for 7 mg we leverage the 4 mg data



Bayes Model with 4 mg data for prior and 7 mg fractional data

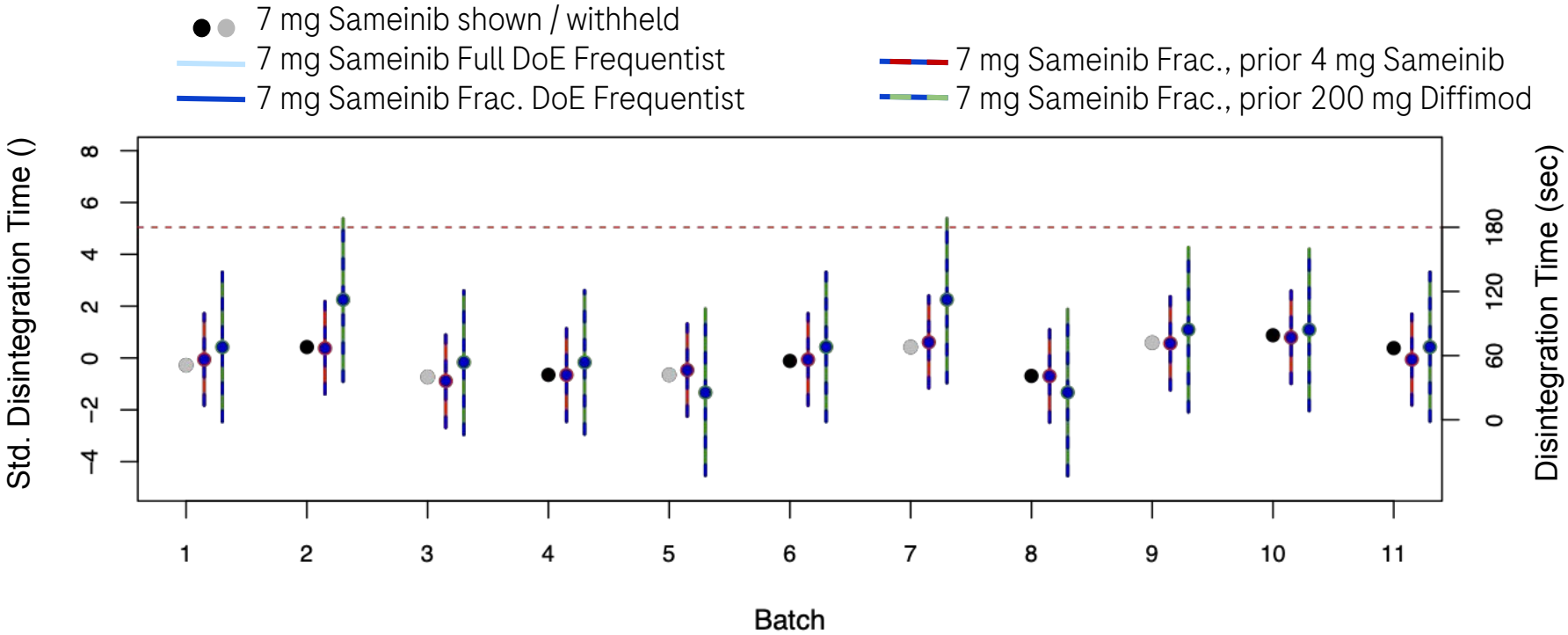
Bayes tradeoff: reduction of variance, increase of bias.



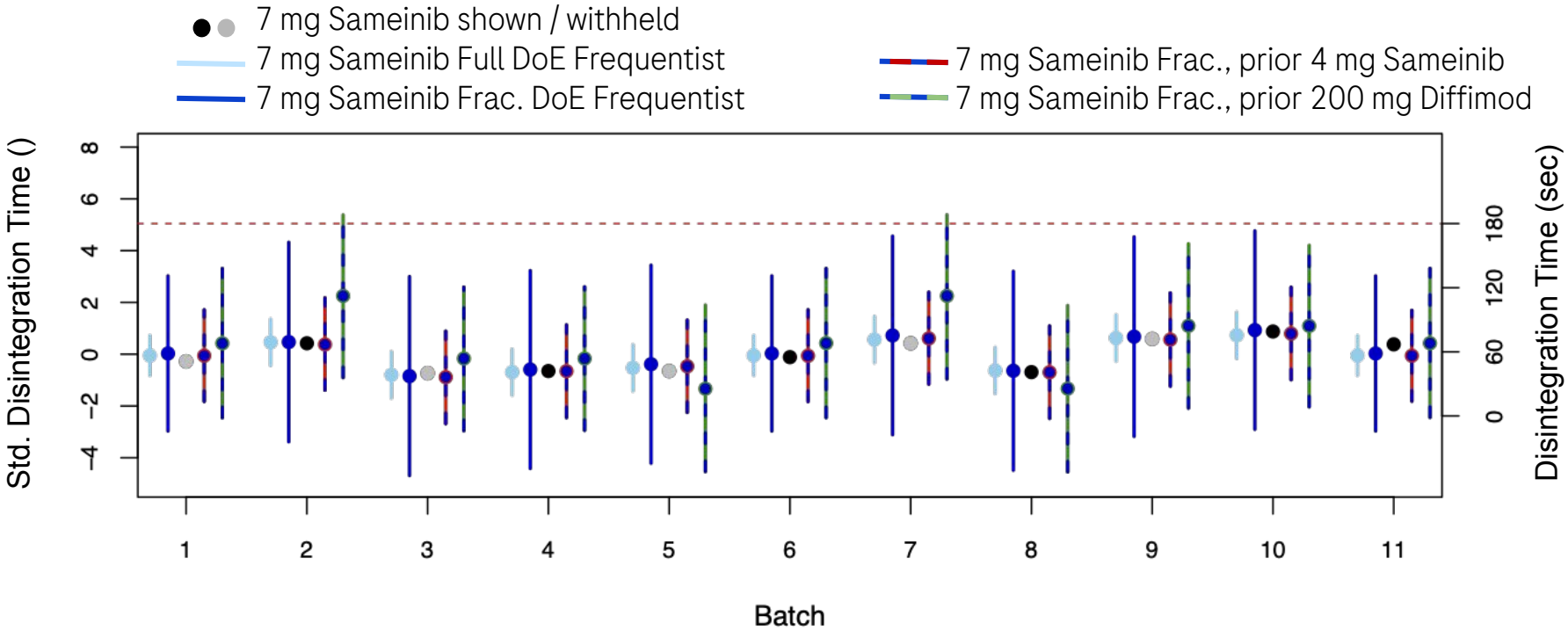
Bayes Model with 200 mg data for prior and 7 mg fractional data

Priors informed by inconsistent process leads to wider uncertainty = prediction intervals.

In-built STOP sign



Disintegration - Overview of Prediction Intervals



Summary

- This use case: Due to elimination of inactive factors and 2FI, a reduced DoE with 6 runs instead of 11 runs renders feasible. This corresponds to 45% less material and analytical efforts.
- A conjugate prior informed on existing development data (4 mg Sameinib) can be used for better predictions in a reduced new dose strength (7 mg Sameinib). Under favorable conditions, the 11 historic runs compensate for 5 runs saved.
- Upgrades and storage of data in FAIR format, cross-sectional discussions to explain the concept, benefits and risks may reduce the overall savings in FTE.
- Upcycling data requires a robust process and analytics, and consistency between old and new process.

Acknowledgement

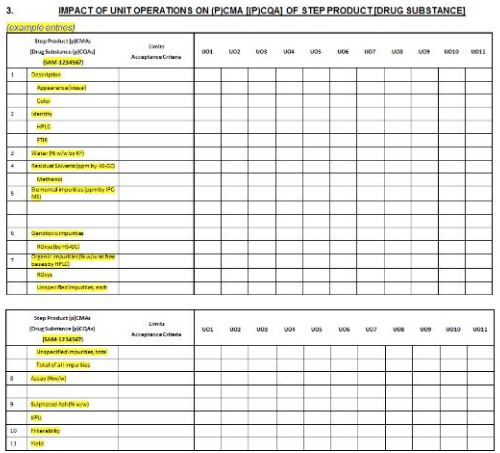
Christian Schmid
Theo Koulis
Lisa Bernstein
Oskar Kalb
Ludmilla Vincenzi

Appendix

Background

- Today, we use our scientific experience for a lean development by conducting appropriate quality risk assessments, efficient DoE and other tools.

pCPP Assessment



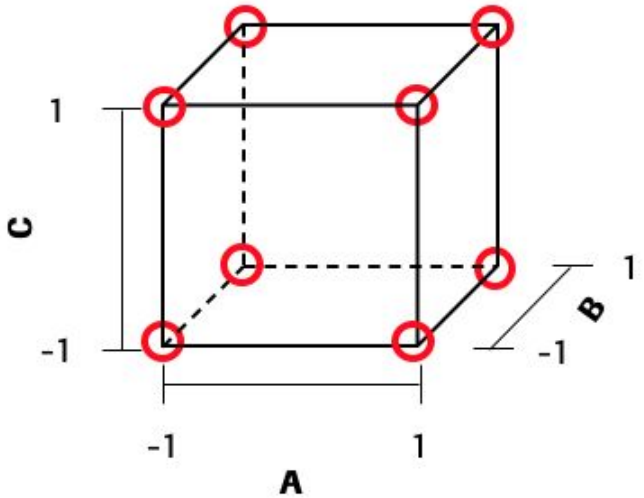
3. IMPACT OF UNIT OPERATIONS ON (P)CMA [(P)CQA] OF STEP PRODUCT [DRUG SUBSTANCE]

(example entries)

	Step Product [pCMAs (Drug Substance [pCQAs] (SAM-1234567)]	Limits Acceptance Criteria	U01	U02	U03	U04	U05	U06	U07	U08	U09	U010	U011
1	Description												
	Appearance (visual)												
	Color												
2	Identity												
	HPLC												
	FTIR												
3	Water (% w/w by KF)												
4	Residual Solvents (ppm by HS-GC)												
	Methanol												
5	Elemental impurities (ppm by ICP-MS)												
6	Genotoxic impurities												
	RDX (by HS-GC)												
7	Trifluoroacetamide (% w/w by free base by HPLC)												
	RDX												
	Unspecified impurities, each												

	Step Product [pCMAs (Drug Substance [pCQAs] (SAM-1234567)]	Limits Acceptance Criteria	U01	U02	U03	U04	U05	U06	U07	U08	U09	U010	U011
	Unspecified impurities, total												
	Total of all impurities												
8	Assay (% w/w)												
9	Sulphated Ash (% w/w)												
	APU												
10	Filterability												
11	Yield												

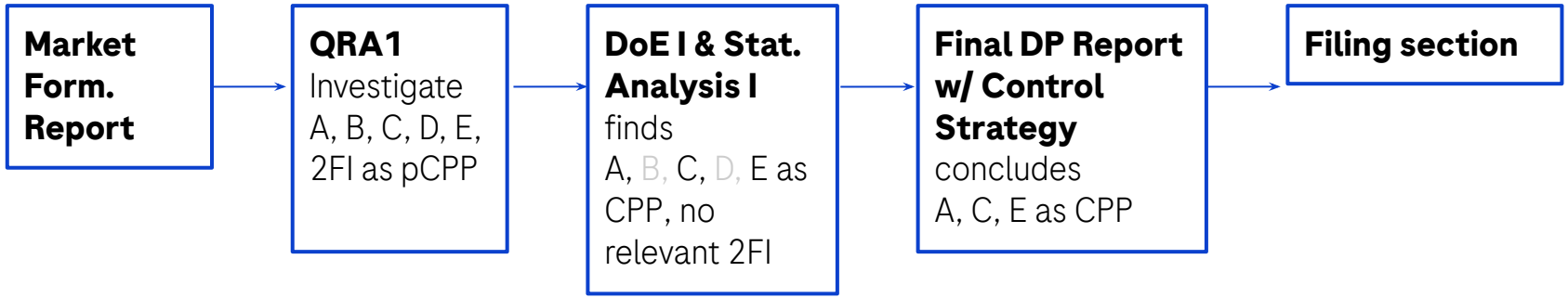
Process Parameter (PP)	Potential Criticality	Pot. Scale Sensitive (y/n)	Set Point/Range	Testing Range	Remark
PP6 Amt. of Reagent (Equiv.)	6c, 6d, 7	n	1.1	1.0 – 1.2	Important for imp
PP7 Temperature (°C)	6c, 6d, 7	n	20	15 – 25	Important for imp
PP8 Amt. of RSM (Equiv.)	3a, 3b	n	1	0.95 – 1.05	Interaction betwe
PP9 Reaction Temperature (°C)	3a, 3b	n	35	25 – 45	and PP9 expected
PP10 Feed time of RSM (min.)		y	30	25 – 120	Feed time can be considerably diffe between lab scale production. There wide testing range chosen to reflect feed times require various scales.



Resolution V DoE

From Market Formulation to Filing Document

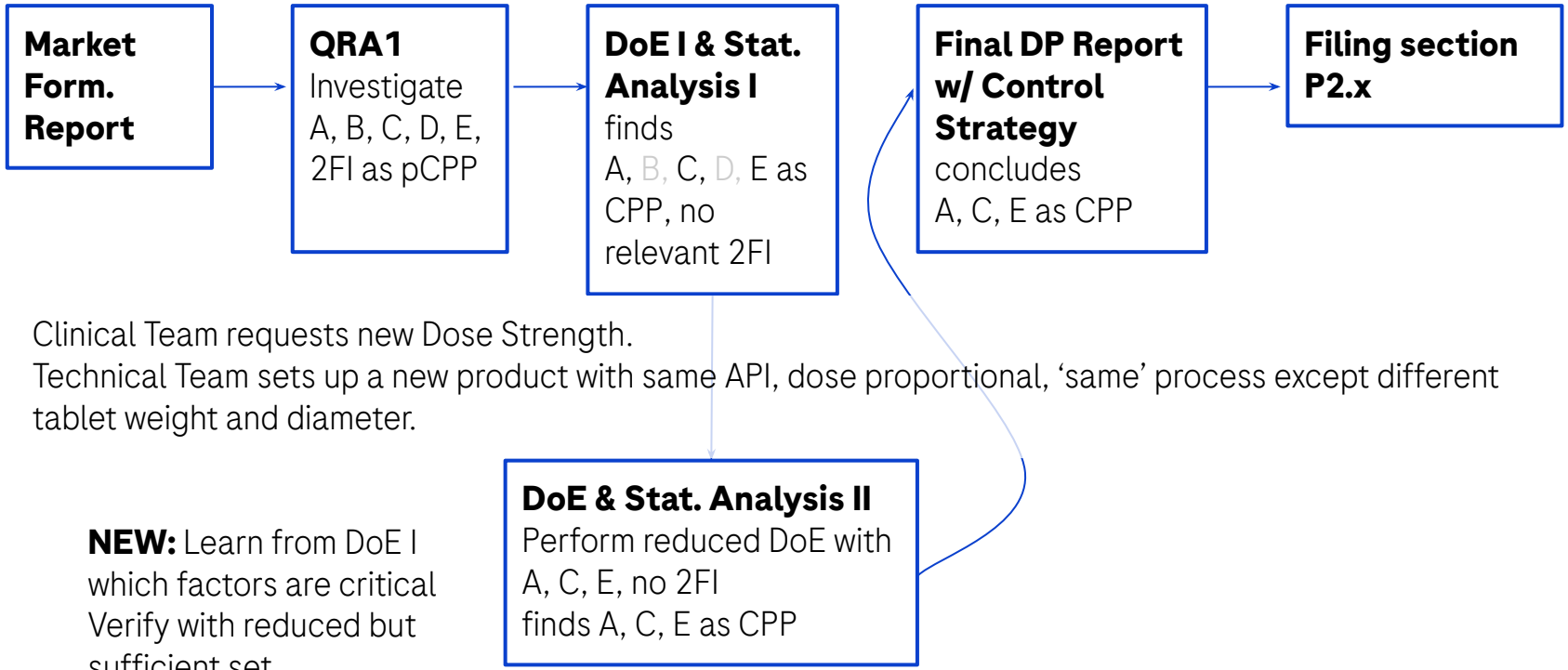
A very simplified overview with some selected milestones - OLD model



2FI: Two Factor Interaction

From Market Formulation to Filing Document

A very simplified overview with some selected milestones - NEW model



Intro to Bayes

A natural choice to incorporating “prior” information into a statistical analysis is to use Bayesian statistics

Assume a probability model for data, i.e., $f(x \mid \theta)$ a.k.a the likelihood function

- The distribution of X depends on the parameter θ
  - To the Frequentist, it is a unknown but fixed constant
  - To the Bayesian, since we don’t know the value of the parameter, it’s a random variable
- Bayesians model their uncertainty about θ with a probability distribution, e.g., $p(\theta)$
  - $p(\theta)$ is called the *prior* distribution
  - Prior because it represents the statistician’s (or scientist’s) beliefs about θ *before* seeing the data
- The distribution of θ after seeing the data is called the *posterior* distribution
  - The posterior distribution is the conditional distribution of the parameter given the data, i.e, $p(\theta \mid x)$
- Prior distribution $p(\theta)$ is based on the best available information
  - But yours might be different than mine. It’s subjective
- Subjectivity is the most frequent objection to Bayesian methods
  - The prior distribution influences the conclusions
  - Two scientists may arrive at different conclusions from the same data, based on the same statistical analysis
- The influence of the prior goes to zero as the sample size increases

Our current strategy for specifying the prior

- Assume that we are using the standard linear model formulation where the observation errors are independent and have the equal variance, i.e.,

$$y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I)$$

where $\mathbf{I}$ is a $\mathbf{n}$ x $\mathbf{n}$ identity matrix and $\boldsymbol{\beta}$ a vector of linear model coefficients

- In general, we can write the posterior distribution as follows:

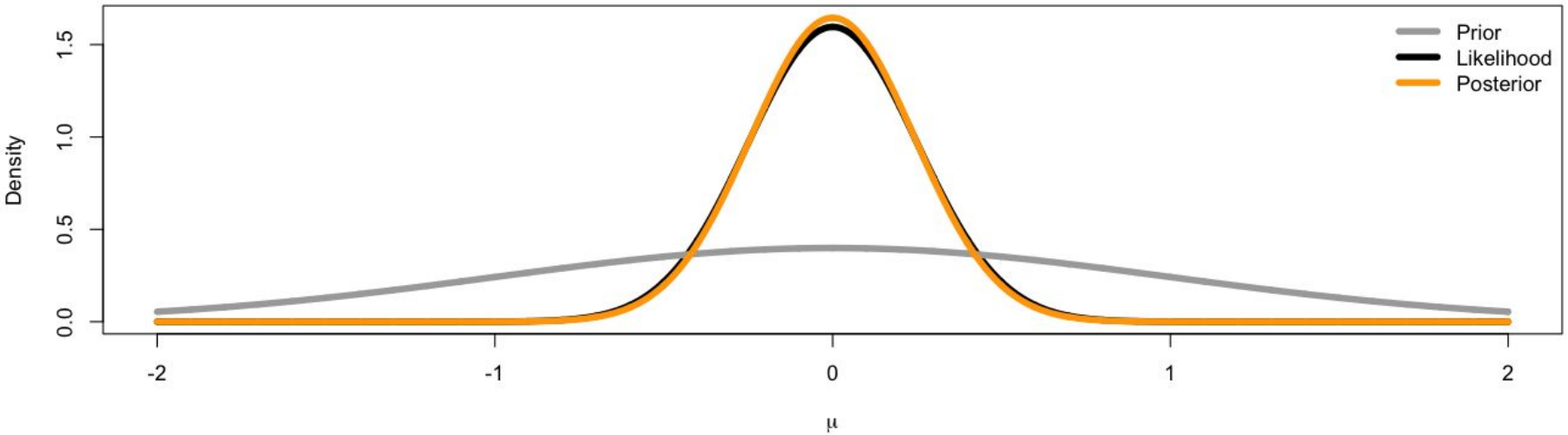
$$p(\beta, \sigma^2|X, y) \propto f(y|\beta, \sigma^2, X)p(\beta, \sigma^2|X)$$

Posterior — Likelihood — Prior

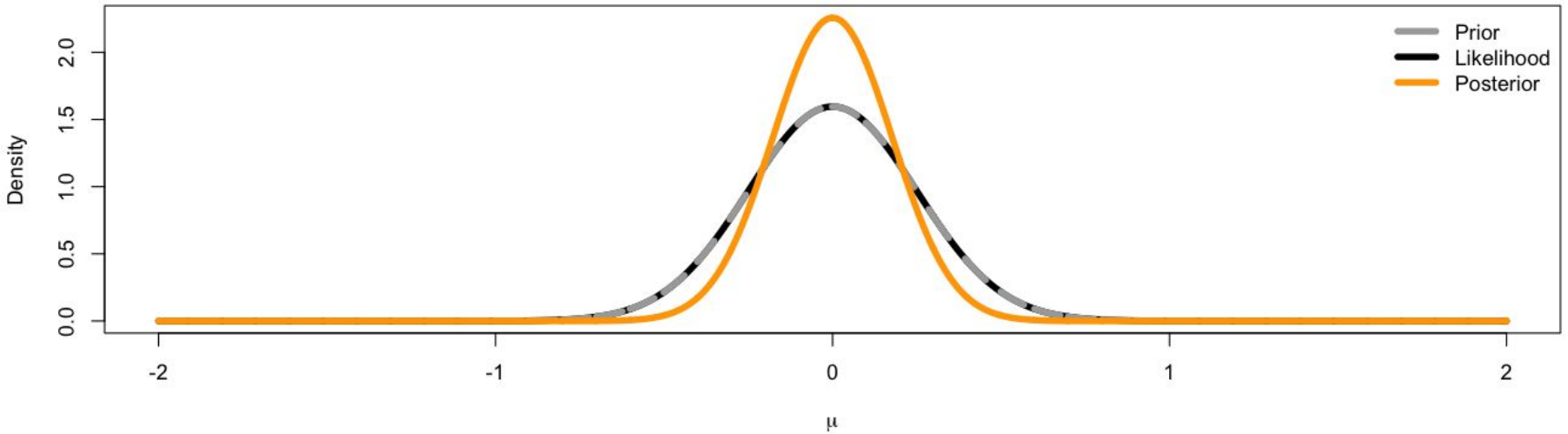
- Based on the "**historical data**" we will assume a noninformative prior of the following form:

$$p(\beta, \sigma^2|X) \propto \frac{1}{\sigma^2}$$

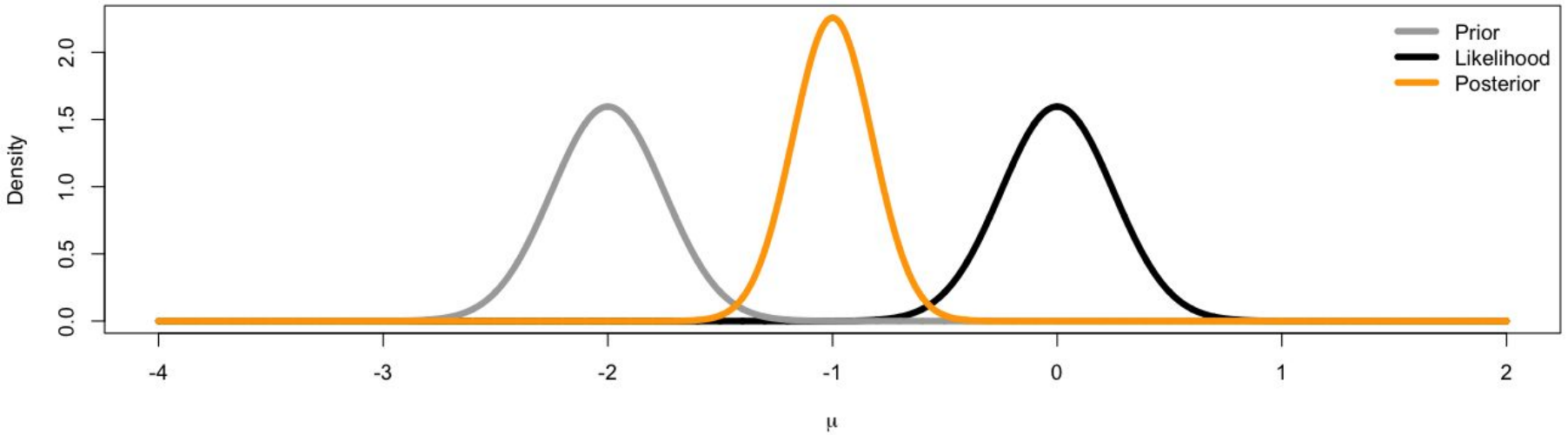
Diffuse Prior



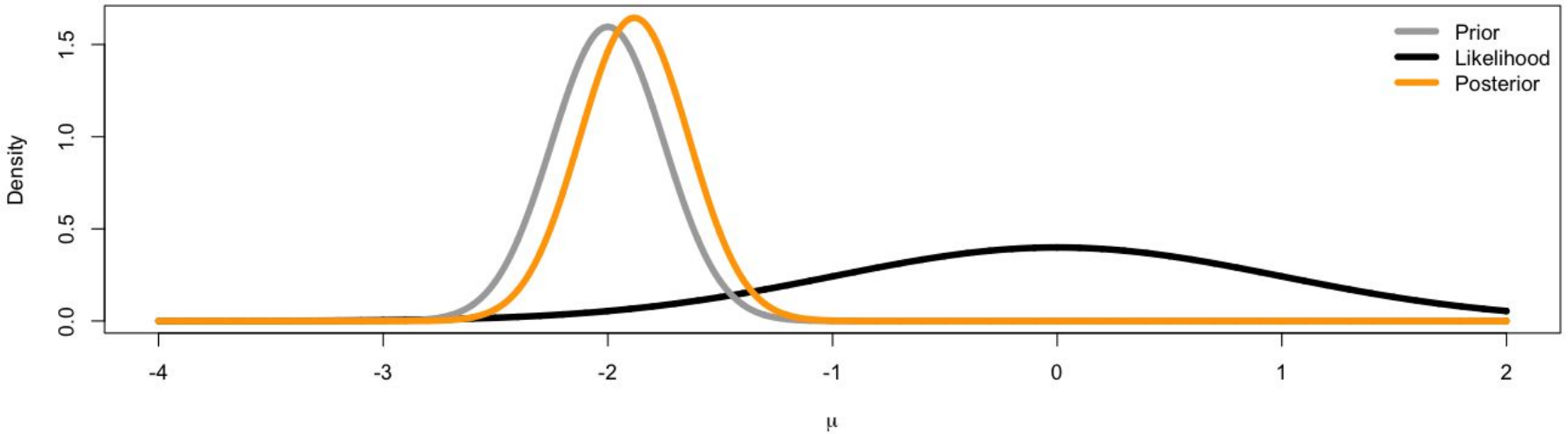
Informative Prior matches the likelihood

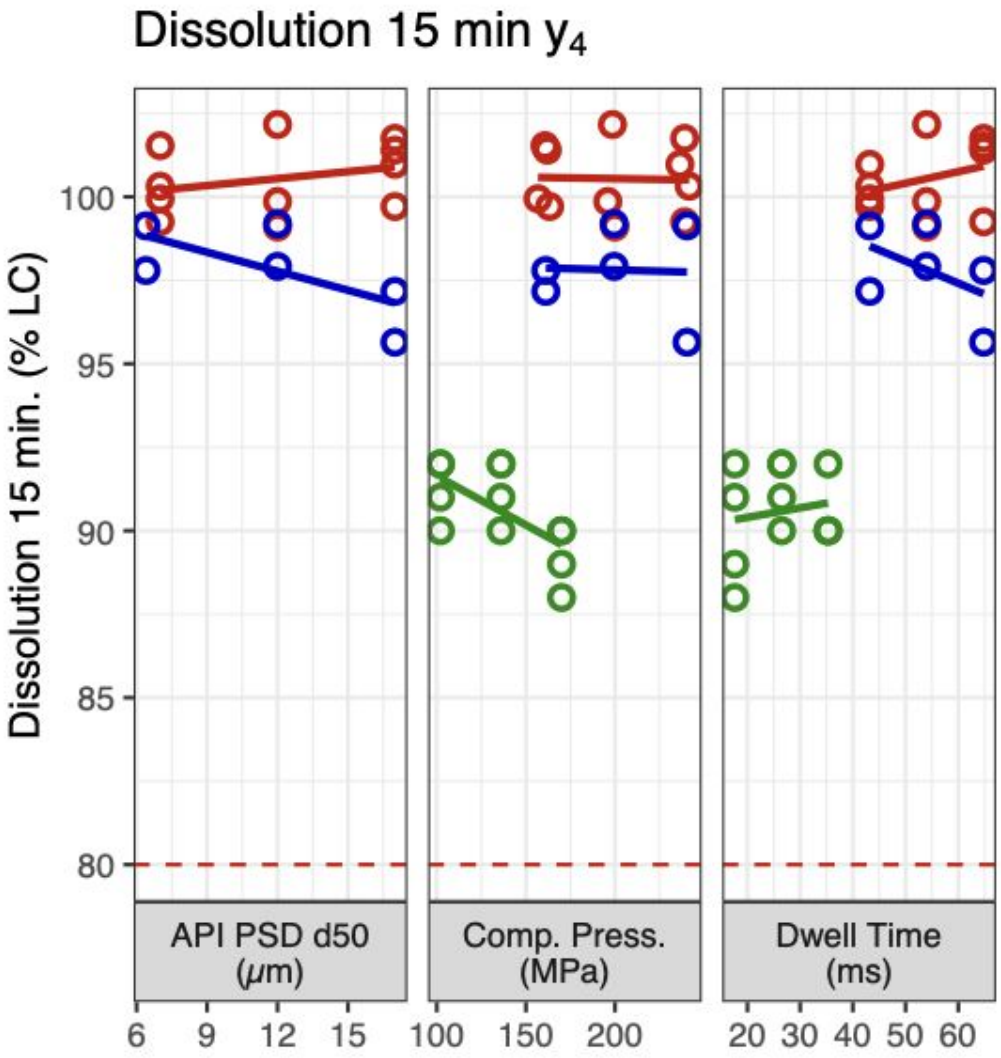


Informative prior doesn't match the likelihood



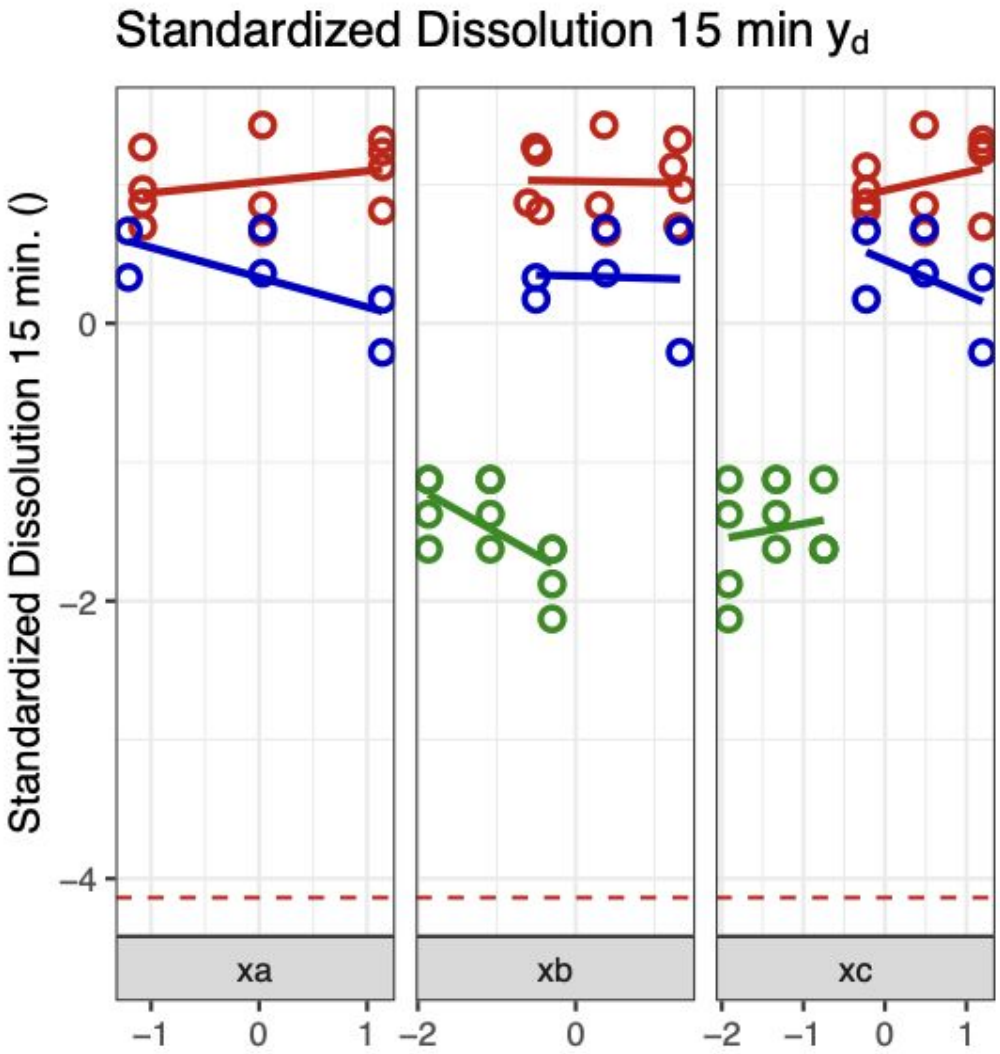
Informative prior and diffuse likelihood





Source

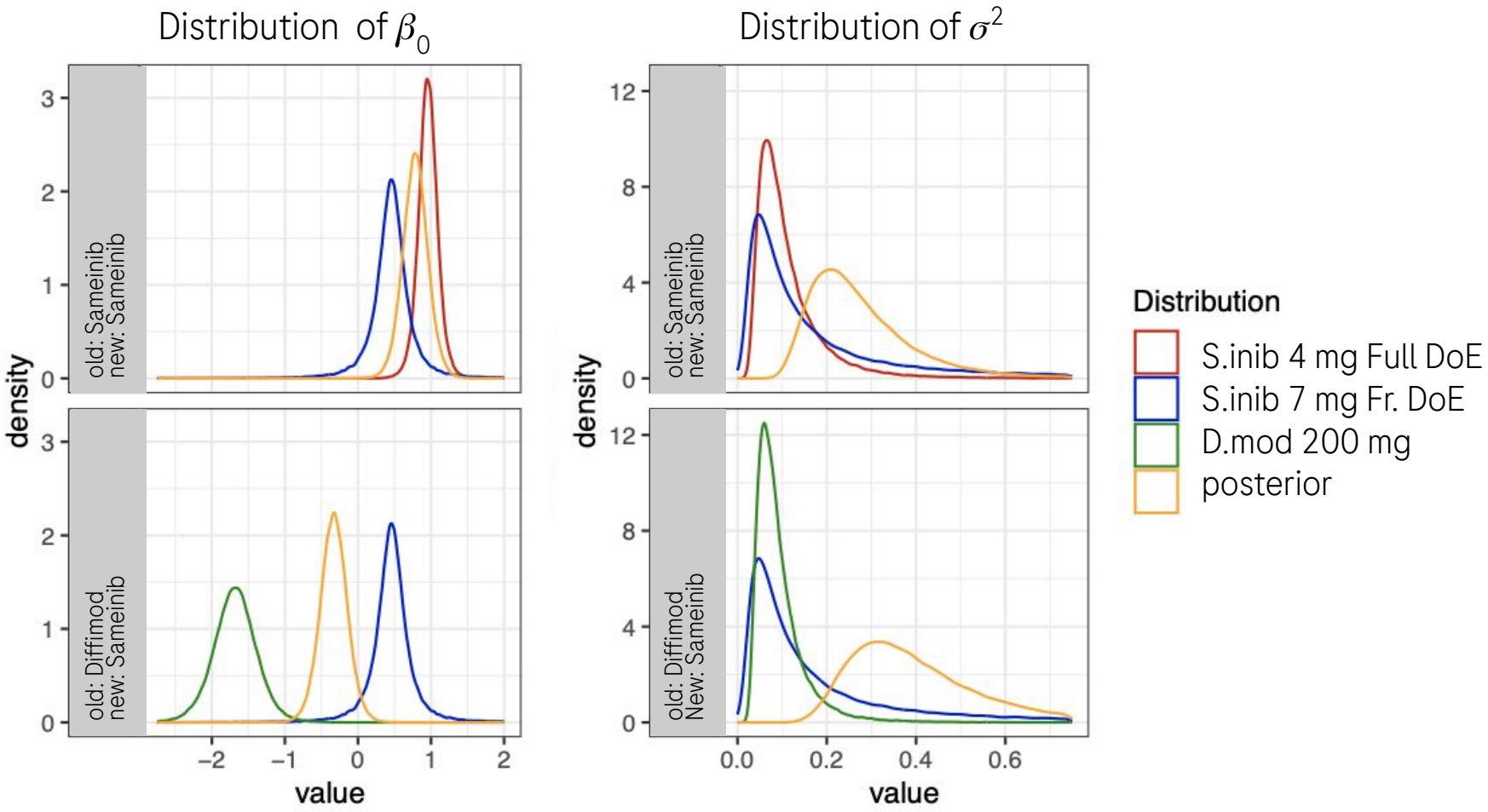
7 mg Sameinib ½ Frac. DoE
4 mg Sameinib Full DoE
200 mg Diffimod



Source

7 mg Sameinib ½ Frac. DoE
4 mg Sameinib Full DoE
200 mg Diffimod

y_d - Dissolution 15 min.



Dissolution - Overview of Prediction Intervals

Both Bayesian predictions have smaller pred. intervals -

What is the meaning if a response like Dissolution 15 min is selected to be in certain range?