

# Calibrated borrowing: getting the questions and answers right

Christian Röver

UNIVERSITÄTSMEDIZIN : UMG  
GÖTTINGEN

Department of Medical Statistics,  
University Medical Center Göttingen,  
Göttingen, Germany

GMDS 2025, Jena  
September 9, 2025

“Finding the question  
is often more important  
than finding the answer.”

(John W. Tukey)

- statistical analyses yield **probabilistic statements**  
(credible / confidence intervals, . . .)
- to be meaningful, probabilities need to be **calibrated**
- sometimes given **by construction**  
(e.g.: assumptions are met)
- sometimes need to **verify** in sensitivity analyses (simulations, . . .)  
(e.g.: asymptotics are used, assumptions are violated)

- statistical analyses yield **probabilistic statements**  
(credible / confidence intervals, . . .)
- to be meaningful, probabilities need to be **calibrated**
- sometimes given **by construction**  
(e.g.: assumptions are met)
- sometimes need to **verify** in sensitivity analyses (simulations, . . .)  
(e.g.: asymptotics are used, assumptions are violated)
  
- **proper setup** is crucial  
to yield **sensible** (coverage, type-I error, power, . . .) statements

# Bayesian vs. frequentist calibration

“marginal” vs. “conditional”

- Bayesian and frequentist methods come with differing implications / requirements regarding calibration:

## Bayesian

intervals provide coverage *on average* over the prior distribution:  
a “*marginal*” property

(marginalisation over  $p(\theta, y)$ )

## frequentist

intervals provide coverage *for any point* (“*true parameter*”) in parameter space:  
a “(*uniform*) *conditional*” property

(marginalisation over  $p(y|\theta)$ )

# Bayesian calibration

## Simple (one-stage) example

### Setup:

- prior:

$$\theta \sim \text{Normal}(\mu_0 = 0, \sigma_0^2 = 1)$$

- likelihood:

$$\bar{y}|\theta \sim \text{Normal}(\theta, s_y^2 = \frac{1}{m})$$

# Bayesian calibration

## Simple (one-stage) example

### Setup:

- prior:

$$\theta \sim \text{Normal}(\mu_0 = 0, \sigma_0^2 = 1)$$

- likelihood:

$$\bar{y}|\theta \sim \text{Normal}(\theta, s_y^2 = \frac{1}{m})$$

- posterior:

$$\theta|\bar{y} \sim \text{Normal}\left(\frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{1}{s_y^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{1}{s_y^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{s_y^2}}\right)$$

# Bayesian calibration

## Simple (one-stage) example

### Example

- a sample of size  $m = 9$  yields average  $\bar{y} = 1$
- posterior:

$$\theta | \bar{y} = 1 \sim \text{Normal}(0.90, 0.32^2)$$

# Bayesian calibration

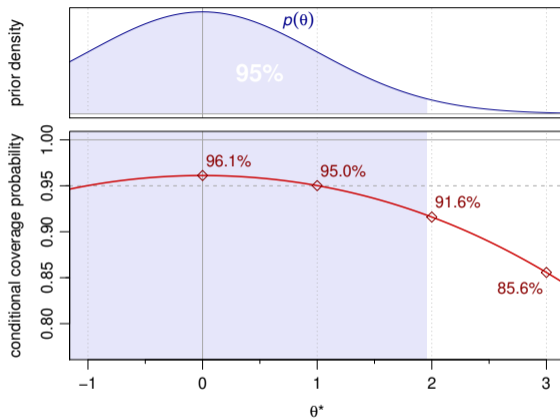
## Point-wise (frequentist) coverage

- checking **frequentist calibration**:  
**vary (fixed) true parameter** ( $\theta^*$ ), marginalize over  $p(\bar{y} | \theta^*)$

true $\theta^*$	CI coverage
0.0	96.1%
1.0	95.0%
2.0	91.6%
3.0	85.6%

# Bayesian calibration

## Point-wise and mean coverage



- “point-wise” coverage varies with  $\theta^*$
- **mean coverage** is (by construction) **exactly** at nominal 95% for  $\theta^* \sim \text{Normal}(\mu_0, \sigma_0^2)$

# Bayesian calibration

Mean coverage for “analysis” and “design” priors

- analysis is based on  $\text{Normal}(\mu_0, \sigma_0^2)$  prior (“**analysis prior**”)
- may compute **mean coverage** based on differing “**design prior**”  $\text{Normal}(\mu_*, \sigma_*^2)$  (i.e., assumptions are violated!)

design prior		
$\mu_*$	$\sigma_*$	coverage
<b>0.0</b>	<b>1.0</b>	<b>95.0%</b>
0.0	0.5	
0.0	2.0	
1.0	1.0	

- 95% intervals are **calibrated** for matching priors ( $\mu_* = \mu_0, \sigma_* = \sigma_0$ )

# Bayesian calibration

Mean coverage for “analysis” and “design” priors

- analysis is based on  $\text{Normal}(\mu_0, \sigma_0^2)$  prior (“**analysis prior**”)
- may compute **mean coverage** based on differing “**design prior**”  $\text{Normal}(\mu_*, \sigma_*^2)$  (i.e., assumptions are violated!)

design prior		
$\mu_*$	$\sigma_*$	coverage
<b>0.0</b>	<b>1.0</b>	<b>95.0%</b>
0.0	0.5	95.8%
0.0	2.0	91.4%
1.0	1.0	93.8%

- 95% intervals are **calibrated** for matching priors ( $\mu_* = \mu_0, \sigma_* = \sigma_0$ )
- **not calibrated** otherwise  
(some settings may be “optimistic” / “pessimistic” / “conservative”)

# Bayesian calibration with historical data

## Two-stage example

### 2-stage example: setup

- prior:

$$\theta \sim \text{Normal}(\mu_0 = 0, \sigma_0^2 = 1)$$

- **first** sample (size  $n$ ):

$$\bar{x}|\theta \sim \text{Normal}(\theta, s_x^2 = \frac{1}{n})$$

- **second** sample (size  $m$ ):

$$\bar{y}|\theta \sim \text{Normal}(\theta, s_y^2 = \frac{1}{m})$$

- (idea: “historical data”  $\bar{x}$ , “current data”  $\bar{y}$ )

# Bayesian calibration with historical data

## Two-stage example

### 2-stage example: analysis

- posterior may be expressed as:

$$p(\theta | \bar{x}, \bar{y}) \propto p(\bar{y} | \theta) p(\bar{x} | \theta) p(\theta)$$

# Bayesian calibration with historical data

## Two-stage example

### 2-stage example: analysis

- posterior may be expressed as:

$$\begin{aligned} p(\theta | \bar{x}, \bar{y}) &\propto p(\bar{y} | \theta) p(\bar{x} | \theta) p(\theta) \\ &\propto \left( p(\bar{y} | \theta) p(\bar{x} | \theta) \right) p(\theta) \propto \underbrace{p(\bar{x}, \bar{y} | \theta)}_{\text{joint likelihood}} \underbrace{p(\theta)}_{\text{prior}} \end{aligned}$$

# Bayesian calibration with historical data

## Two-stage example

### 2-stage example: analysis

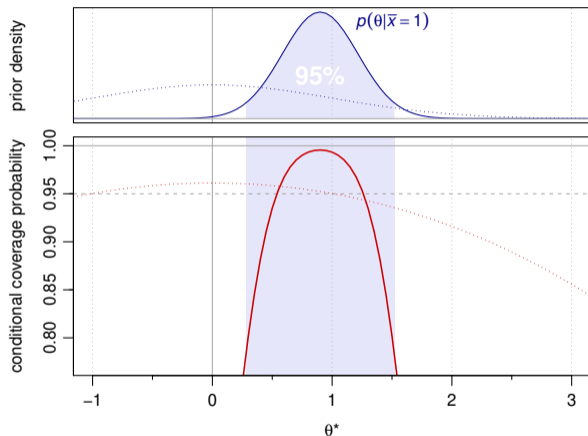
- posterior may be expressed as:

$$\begin{aligned} p(\theta | \bar{x}, \bar{y}) &\propto p(\bar{y} | \theta) p(\bar{x} | \theta) p(\theta) \\ &\propto \left( p(\bar{y} | \theta) p(\bar{x} | \theta) \right) p(\theta) \propto \underbrace{p(\bar{x}, \bar{y} | \theta)}_{\text{joint likelihood}} \underbrace{p(\theta)}_{\text{prior}} \\ &\propto p(\bar{y} | \theta) \left( p(\bar{x} | \theta) p(\theta) \right) \propto \underbrace{p(\bar{y} | \theta)}_{\bar{y} \text{ likelihood}} \underbrace{p(\theta | \bar{x})}_{\text{historical prior}} \end{aligned}$$

- historical-data** ( $\bar{x}$ ) posterior as **(informative) prior** for new data ( $\bar{y}$ )
- “technically” little has changed: calibration unaffected  
(may also think of data as larger sample of size  $n + m$ )

# Bayesian calibration with historical data

## Two-stage example

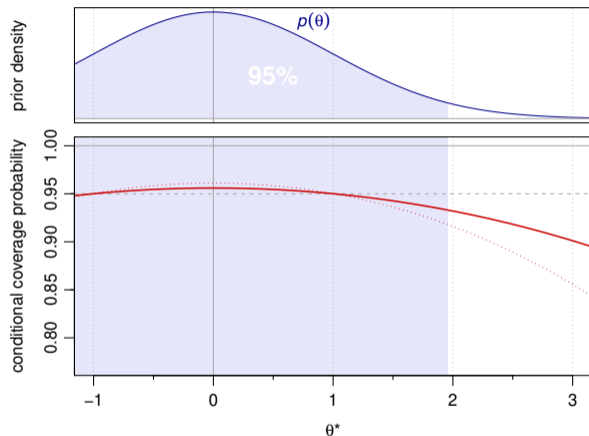


(dotted lines from previous figure)

- may again investigate “conditional” coverage probabilities:
  - fixing  $\bar{x}$ , marginalizing over  $p(\bar{y} | \theta^*, \bar{x})$   
**(Is this concerning? Does this suggest worse properties?)**

# Bayesian calibration with historical data

## Two-stage example

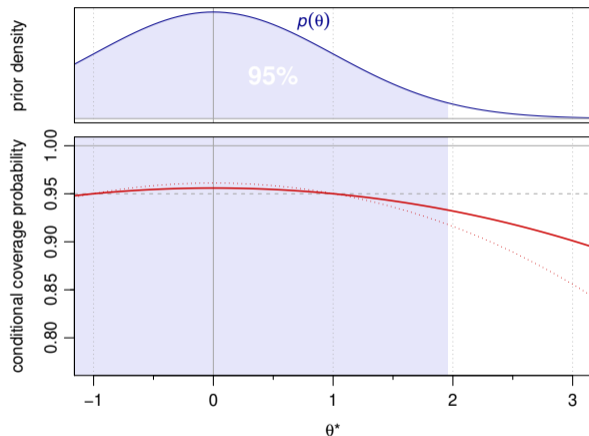


(dotted lines from previous figure)

- may again investigate “conditional” coverage probabilities:
  - fixing  $\bar{x}$ , marginalizing over  $p(\bar{y} | \theta^*, \bar{x})$   
**(Is this concerning? Does this suggest worse properties?)**
  - or marginalizing over  $p(\bar{x}, \bar{y} | \theta^*)$   
**(Does this suggest improved properties?)**

# Bayesian calibration with historical data

## Two-stage example

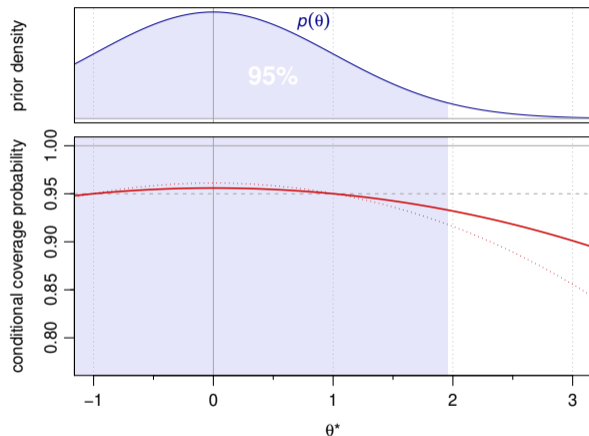


(dotted lines from previous figure)

- may again investigate “conditional” coverage probabilities:
  - fixing  $\bar{x}$ , marginalizing over  $p(\bar{y} | \theta^*, \bar{x})$   
**(Is this concerning? Does this suggest worse properties?)**
  - or marginalizing over  $p(\bar{x}, \bar{y} | \theta^*)$   
**(Does this suggest improved properties?)**
- implications unclear

# Bayesian calibration with historical data

## Two-stage example

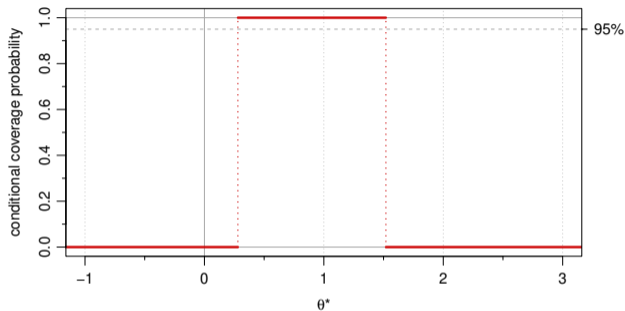


(dotted lines from previous figure)

- may again investigate “conditional” coverage probabilities:
  - fixing  $\bar{x}$ , marginalizing over  $p(\bar{y} | \theta^*, \bar{x})$   
**(Is this concerning? Does this suggest worse properties?)**
  - or marginalizing over  $p(\bar{x}, \bar{y} | \theta^*)$   
**(Does this suggest improved properties?)**
- implications unclear
- **mean coverage** again guaranteed in both cases  
(w.r.t. both  $p(\bar{y} | \bar{x})$  or  $p(\bar{x}, \bar{y})$ )

# Bayesian calibration with historical data

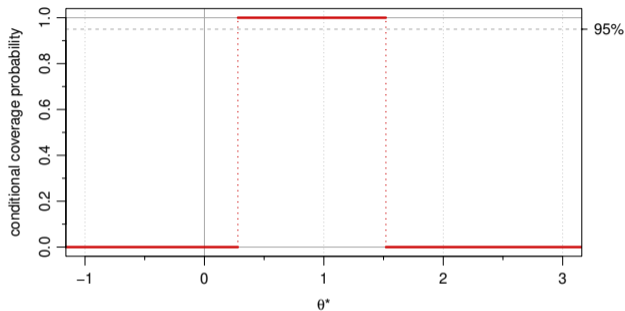
Operating characteristics: conditioning and marginalizing



- more pronounced example:  
consider coverage probability  
based on 1st-stage data only:  
95% interval based on  $p(\theta|\bar{x} = 1)$   
(conditioning on  $\theta^*$ )

# Bayesian calibration with historical data

Operating characteristics: conditioning and marginalizing

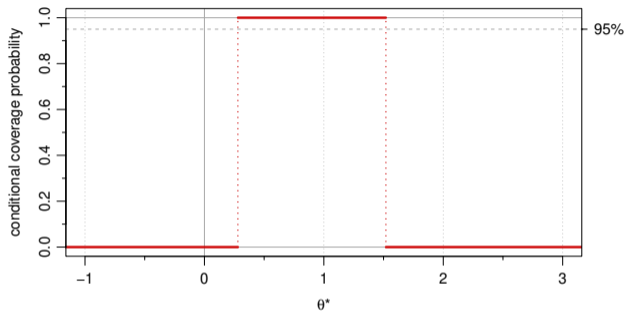


- more pronounced example:  
consider coverage probability  
based on 1st-stage data only:  
95% interval based on  $p(\theta|\bar{x} = 1)$   
(conditioning on  $\theta^*$ )

This is terrible!

# Bayesian calibration with historical data

Operating characteristics: conditioning and marginalizing



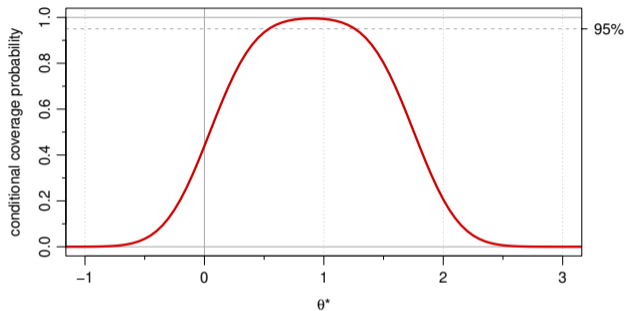
- more pronounced example:  
consider coverage probability  
based on 1st-stage data only:  
95% interval based on  $p(\theta|\bar{x} = 1)$   
(conditioning on  $\theta^*$ )

This is terrible!

Problem is, parameter  $\theta^*$  and data  $\bar{x}$  are varied independently.  
This is no concern as long as the model is correct  
(which needs to be viewed *as a whole*).

# Bayesian calibration with historical data

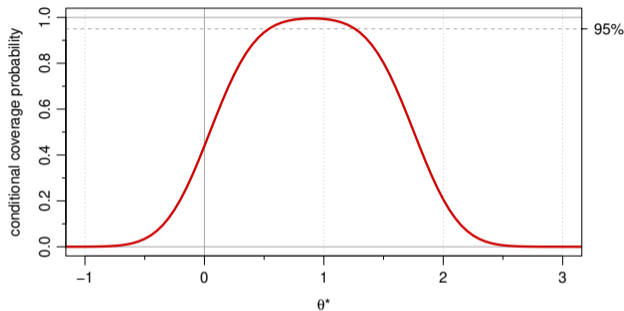
Operating characteristics: conditioning and marginalizing



- back to previous example:  
consider coverage probability  
based on 1st- and 2nd-stage data:  
interval based on  $p(\theta|\bar{x} = 1, \bar{y})$ ,  
(conditioning on  $\theta^*$  and  $\bar{x} = 1$   
marginalizing over  $p(\bar{y}|\theta^*, \bar{x} = 1)$ )

# Bayesian calibration with historical data

Operating characteristics: conditioning and marginalizing

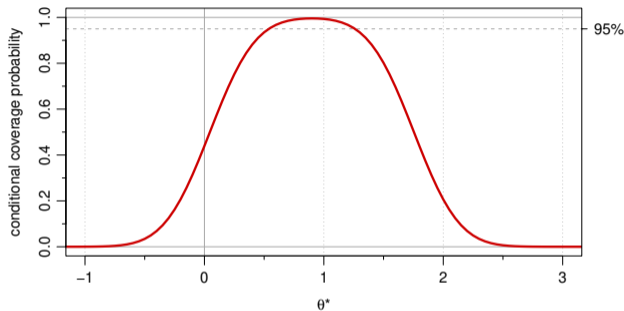


- back to previous example:  
consider coverage probability  
based on 1st- and 2nd-stage data:  
interval based on  $p(\theta|\bar{x} = 1, \bar{y})$ ,  
(conditioning on  $\theta^*$  and  $\bar{x} = 1$   
marginalizing over  $p(\bar{y}|\theta^*, \bar{x} = 1)$ )

This is terrible!

# Bayesian calibration with historical data

Operating characteristics: conditioning and marginalizing



- back to previous example:  
consider coverage probability  
based on 1st- and 2nd-stage data:  
interval based on  $p(\theta|\bar{x} = 1, \bar{y})$ ,  
(conditioning on  $\theta^*$  and  $\bar{x} = 1$   
marginalizing over  $p(\bar{y}|\theta^*, \bar{x} = 1)$ )

This is terrible!

... parameter  $\theta^*$  and (part of) data ( $\bar{x}$ ) are varied independently.  
(...)  
You need to make sure the *whole* model is plausible.

# Bayesian calibration with historical data

## Different marginalization schemes

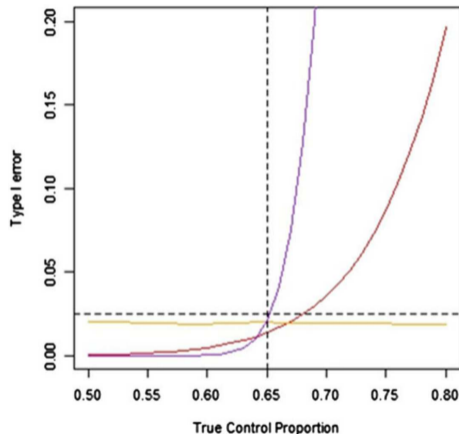
- several **calibration “perspectives”**,  
marginalisation over different distributions

		conditioning on:
1.)	$p(\theta, \bar{x}, \bar{y})$	“(prior) marginal”
2.)	$p(\bar{x}, \bar{y}   \theta)$	“conditional on $\theta^*$ ”      parameter $\theta$
3.)	$p(\bar{y}, \theta   \bar{x})$	“( $\bar{x}$ -posterior) marginal”      data $\bar{x}$
4.)	$p(\bar{y}   \theta, \bar{x})$	“conditional on $\bar{x}$ and $\theta^*$ ”      data $\bar{x}$ and parameter $\theta$

- problematic: **conditioning on parameters and data**  
4th option commonly considered (e.g., previous slides 12–14)
- suggestion:
  - **prefer 1st** (or 3rd)
  - investigate **design prior variations** to gauge operating characteristics

# “What if” questions

## Conditioning & calibration

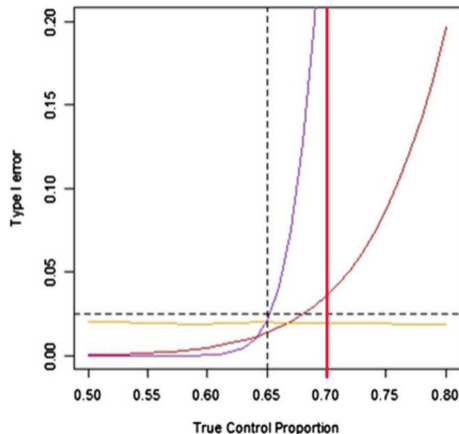


- note: **conditioning** may be considered posing a “**what if**” question
- consider type-I error investigation for fixed historical rate of 0.65

K. Viele *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, **13**(1):41–54, 2014.

# “What if” questions

## Conditioning & calibration



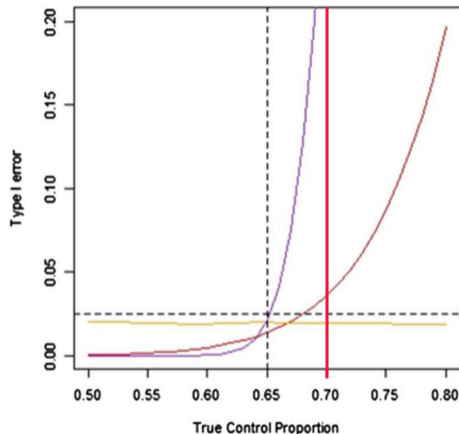
K. Viele *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, **13**(1):41–54, 2014.

- note: **conditioning** may be considered posing a “**what if**” question
- consider type-I error investigation for fixed historical rate of 0.65
- e.g., evaluation at proportion 0.70 may be interpreted as

Q: “*What’s the long-run error if the true value was 0.70, while each time the historical data indicated a rate of 0.65?*”

# “What if” questions

## Conditioning & calibration



K. Viele *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, **13**(1):41–54, 2014.

- note: **conditioning** may be considered posing a “**what if**” question
- consider type-I error investigation for fixed historical rate of 0.65
- e.g., evaluation at proportion 0.70 may be interpreted as

Q: “*What’s the long-run error if the true value was 0.70, while each time the historical data indicated a rate of 0.65?*”

A: “*In that case, your model would be wrong anyway.*”

# Calibrated borrowing

## Summary

- when applying Bayesian methods
  - “**point-wise**” coverages (errors, power, . . .) are not instructive
  - avoid **paradoxical conditioning** schemes (parameters and downstream data)

# Calibrated borrowing

## Summary

- when applying Bayesian methods
  - “**point-wise**” coverages (errors, power, ...) are not instructive
  - avoid **paradoxical conditioning** schemes (parameters and downstream data)
- note: **conditioning** corresponds to posing a “**what if**” **question**

# Calibrated borrowing

## Summary

- when applying Bayesian methods
  - “**point-wise**” coverages (errors, power, ...) are not instructive
  - avoid **paradoxical conditioning** schemes (parameters and downstream data)
- note: **conditioning** corresponds to posing a “**what if**” question
- focus on **assumption variations** for **design priors** instead (e.g, mixture weights, vague priors, heterogeneity, ...)
- specification of relevant **design priors** may immediately **feed back** into **analysis prior** specification

# Calibrated borrowing

## Summary

- when applying Bayesian methods
  - “**point-wise**” coverages (errors, power, ...) are not instructive
  - avoid **paradoxical conditioning** schemes (parameters and downstream data)
- note: **conditioning** corresponds to posing a “**what if**” question
- focus on **assumption variations** for **design priors** instead (e.g, mixture weights, vague priors, heterogeneity, ...)
- specification of relevant **design priors** may immediately **feed back** into **analysis prior** specification
  
- **ask sensible questions to get sensible answers**



(radio tower photo by Losch, adapted under CC-BY-SA-3.0)

Three parallel workshops, 20 participants each:

Workshop 1:

**Meta-analysis**

Speaker:

- **Wolfgang Viechtbauer**  
(Maastricht University)
- **Christian Röver**  
(UMG Göttingen)
- **Sebastian Weber**  
(Novartis Basel)

Workshop 2:

**Causal inference**

Speaker:

- **Vanessa Didelez**  
(BIPS Bremen)
- **Arthur Allignol**  
(Daichi Sankyo)
- **Oliver Kuß**  
(DDZ Düsseldorf)
- **Alexandra Strobel**  
(UMH Halle)

Workshop 3:

**Time-to-event analysis**

Speaker:

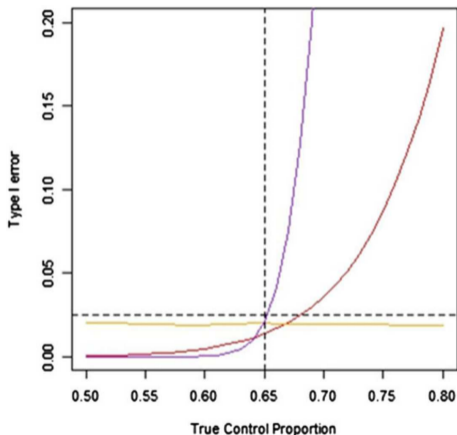
- **Hannes Buchner**  
(Staburo München)
- **Xiaofei Liu**  
(MHM Hannover)
- **Ann-Kathrin Ozga**  
(UKE Hamburg)



+++ additional slides +++

# Examples

Viele *et al.* (2014)



- “Comparison of (...) type I error (...) for separate (orange), pooled (red), and single arm trial (purple) designs. Generally, there is a ‘sweet spot’ near 0.65 where borrowing simultaneously achieves lower MSE, lower type I error, and higher power compared to the separate analysis. Below the sweet spot, we see diminished power with borrowing, and above the sweet spot, we see inflated type I error. Assessing the relative likelihood of these regions is important to assessing the costs and benefits of borrowing.”
- (here: historical rate = 0.65)
- K. Viele *et al.* [Use of historical control data for assessing treatment effects in clinical trials](#). *Pharmaceutical Statistics*, **13**(1):41–54, 2014. (Fig. 2)

# Examples

Schmidli *et al.* (2014)

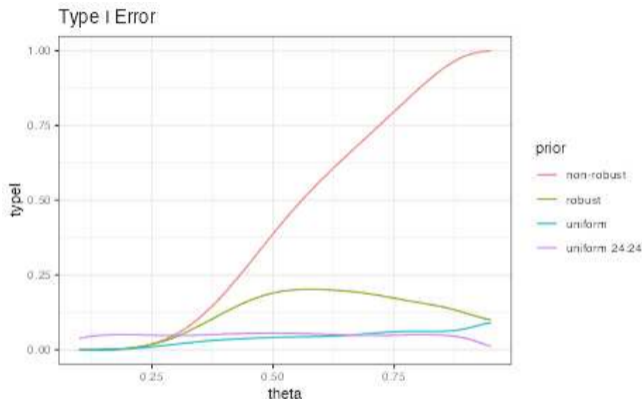
Type I error and power (%) under different control rates  $\psi_*$  and treatment effects  $\delta$  for four priors: Beta(4, 16) (Beta),  $0.5 \times \text{Beta}(4, 16) + 0.5 \times \text{Beta}(1, 1)$  (Mix50),  $0.9 \times \text{Beta}(4, 16) + 0.1 \times \text{Beta}(1, 1)$  (Mix90), Beta(1, 1) (Unif). Also shown is the expected sample size in the control group for the two mixture priors.

Control rate ( $\psi_*$ )	Treatment effect ( $\delta=0$ )				Treatment effect ( $\delta=0.3$ )				Expected sample size control	
	Mix50	Mix90	Beta	Unif	Mix50	Mix90	Beta	Unif	Mix50	Mix90
0.1	0.6	0.1	0.0	1.8	92.0	81.4	81.6	89.7	27.6	20.0
0.2	2.5	1.5	1.6	2.3	88.4	85.7	87.8	82.1	25.5	20.3
0.3	3.9	5.5	6.1	2.4	83.0	88.4	93.4	79.5	28.5	21.2
0.4	4.2	10.4	13.7	2.6	76.7	86.8	97.9	79.5	33.5	23.2
0.5	3.4	12.3	26.0	2.8	77.5	85.4	99.6	81.9	37.4	26.9
0.6	3.0	9.5	44.4	2.6	86.4	89.7	100.0	89.8	38.9	31.8

- “It should be noted that the operating characteristics depend upon the direction of bias induced by incorporating the historical information.”
- H. Schmidli *et al.* [Robust meta-analytic-predictive priors in clinical trials with historical control information](#). *Biometrics*, **70**(4):1023–1032, 2014.

# Examples

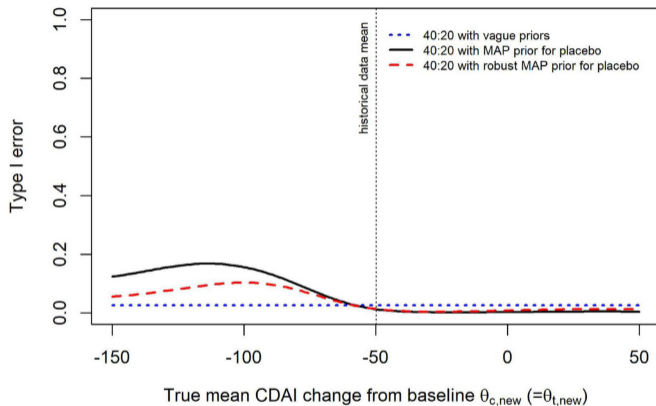
Weber (2025)



- “Note that observing response rates  $> 50\%$  is highly implausible based on the MAP analysis (...)”
- (here: MAP estimate = 0.26 [0.11, 0.46])
- S. Weber. [Getting started with RBest \(binary\)](https://cran.r-project.org/package=RBest). (*RBest package vignette*), <https://cran.r-project.org/package=RBest>, 2025.

# Examples

Best *et al.* (2025)



- “( . . . ) classical Type I error for Bayesian designs with three different analysis priors for the control arm.”
- N. Best *et al.* [Beyond the classical type I error: Bayesian metrics for Bayesian designs using informative priors](#). *Statistics in Biopharmaceutical Research*, **17**(2):183–196, 2025.