Communicating the value of Bayesian approaches in clinical trials: Is it just a prior issue?

Nicky Best

Head of Statistics and Data Science Innovation Hub, GSK

Acknowledgements: PSI Historical Data SIG, Matt Psioda (GSK), Dan Bratton (GSK)

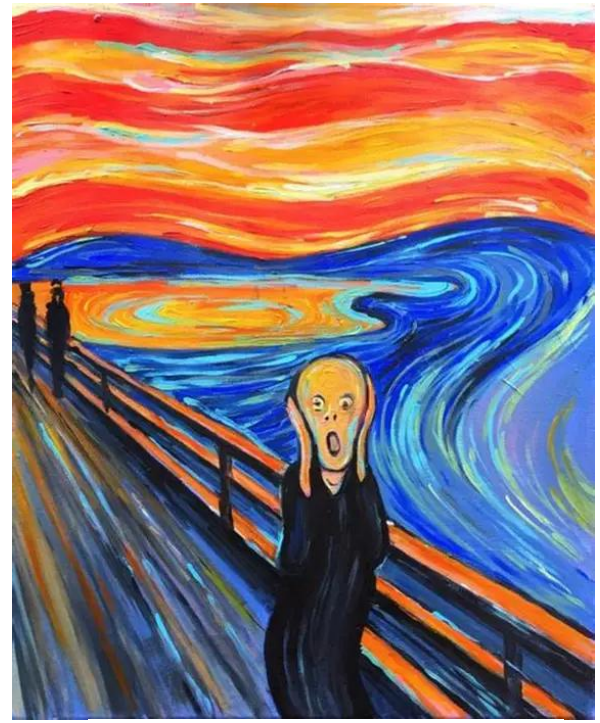
Views expressed are my own and do not necessarily reflect those of GSK

Communicating the value of Bayesian approaches in clinical trials

Katrina and Florian: Puzzle pieces

Communicating the value of Bayesian approaches in clinical trials

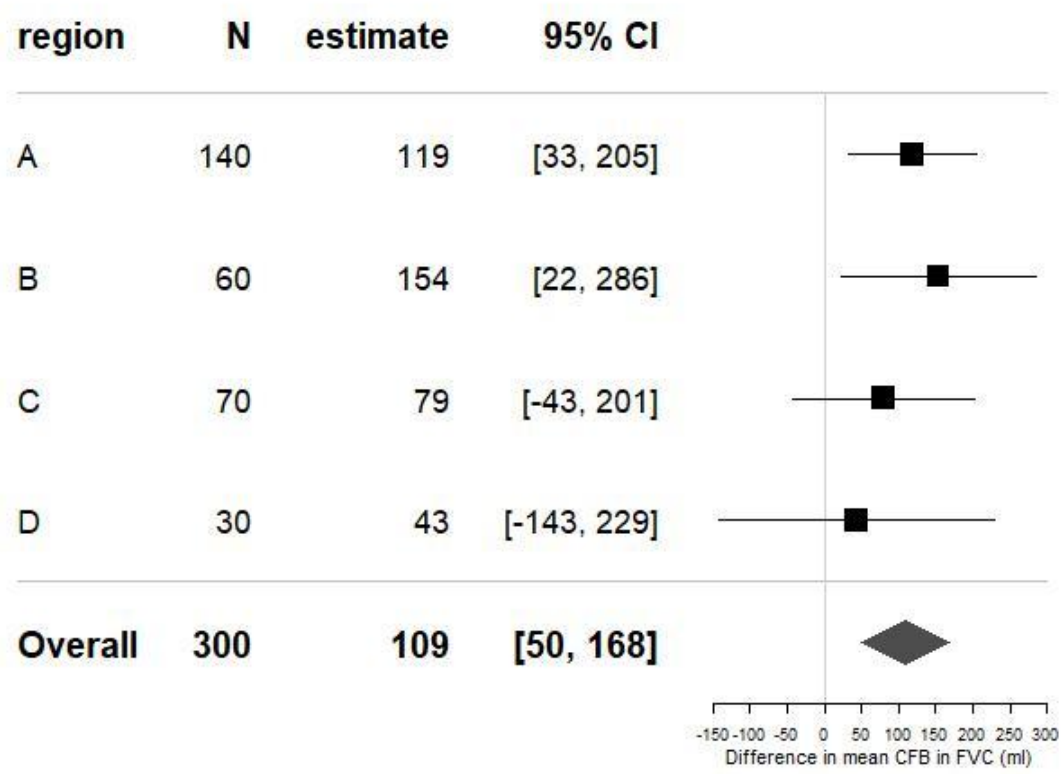
Katrina and Florian: Puzzle pieces



Example 1: Bayesian shrinkage estimation for subgroup effects

Example 1: Bayesian shrinkage estimation for subgroup effects

- Chronic respiratory disease
- Ph3 trial
- Active v control (N = 150 per arm)
- Primary endpoint: CFB in FVC (ml)
- MCID = 100ml
- SD = 260ml
- Subgroups: Regions (4)



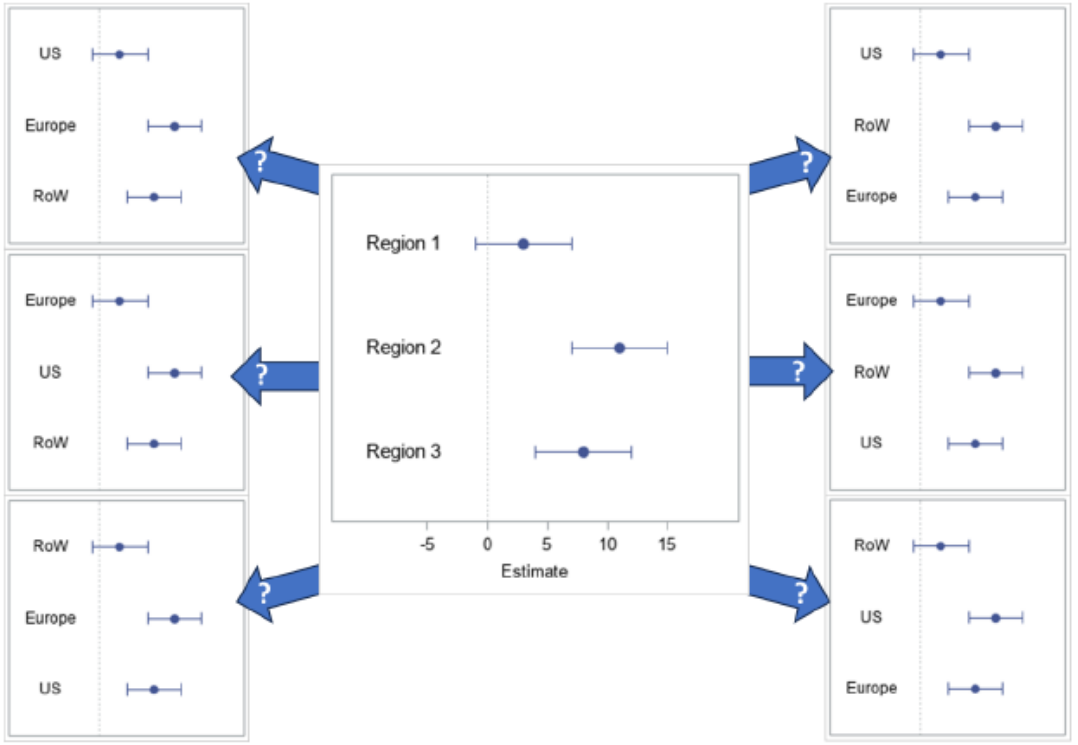
Rationale for shrinkage analysis

- Realistic belief : subgroups might differ slightly but **generally similar** in how they respond to the treatment
  - “Does knowing the effect in subgroup A tell you anything about what to expect in subgroup B?”
  - “Suppose I ask you to predict the treatment effect in subgroup B. If I tell you the effect in subgroup A, does this influence your prediction?”

Rationale for shrinkage analysis

- Realistic belief : subgroups might differ slightly but **generally similar** in how they respond to the treatment
  - ➢ “Does knowing the effect in subgroup A tell you anything about what to expect in subgroup B?”
  - ➢ “Suppose I ask you to predict the treatment effect in subgroup B. If I tell you the effect in subgroup A, does this influence your prediction?”

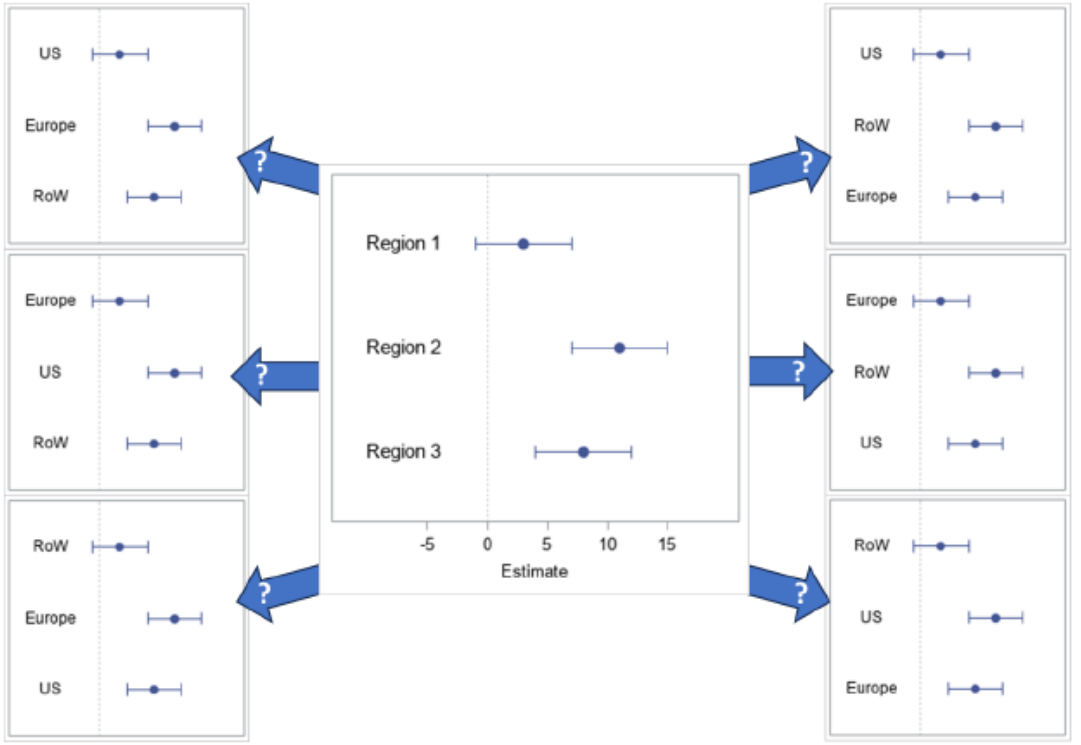
Exchangeability assumption



Rationale for shrinkage analysis

- Realistic belief : subgroups might differ slightly but **generally similar** in how they respond to the treatment
  - ➢ "Does knowing the effect in subgroup A tell you anything about what to expect in subgroup B?"
  - ➢ "Suppose I ask you to predict the treatment effect in subgroup B. If I tell you the effect in subgroup A, does this influence your prediction?"
- Assuming **exchangeability** often more reasonable than independence
  - Exchangeability ≠ identical effects
  - Non-exchangeability → structure that can be modeled

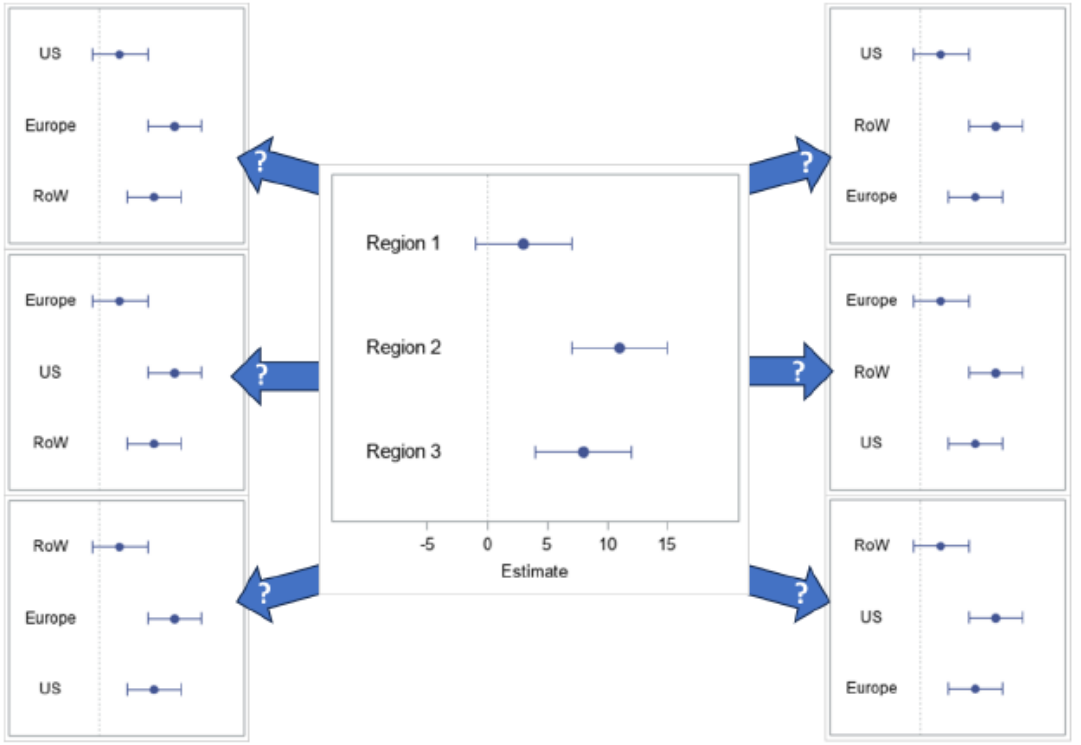
Exchangeability assumption



Rationale for shrinkage analysis

- Realistic belief : subgroups might differ slightly but **generally similar** in how they respond to the treatment
  - ➢ "Does knowing the effect in subgroup A tell you anything about what to expect in subgroup B?"
  - ➢ "Suppose I ask you to predict the treatment effect in subgroup B. If I tell you the effect in subgroup A, does this influence your prediction?"
- Assuming **exchangeability** often more reasonable than independence
  - Exchangeability ≠ identical effects
  - Non-exchangeability → structure that can be modeled
- Statistical rationale: shrinkage gives **lower MSE** than independent estimates

Exchangeability assumption



Shrinkage estimation for subgroup effects: Bayesian statistical model

$\theta_j \sim N(\mu_j, \sigma_j^2)$

θ_j , σ_j = estimated mean & SE of treatment effect in subgroup j

$\mu_j \sim N(\mu, \tau^2)$

μ_j = true treatment effect in subgroup j

$\mu \sim p(\mu)$

μ = overall treatment effect

$\tau \sim p(\tau)$

τ = between-subgroup standard deviation (heterogeneity)

Shrinkage estimation for subgroup effects: Bayesian statistical model

$\theta_j \sim N(\mu_j, \sigma_j^2)$ θ_j , σ_j = estimated mean & SE of treatment effect in subgroup j

$\mu_j \sim N(\mu, \tau^2)$ μ_j = true treatment effect in subgroup j

$\mu \sim p(\mu)$ μ = overall treatment effect

$\tau \sim p(\tau)$ τ = between-subgroup standard deviation (heterogeneity)

Prior on τ - Start by considering fixed values

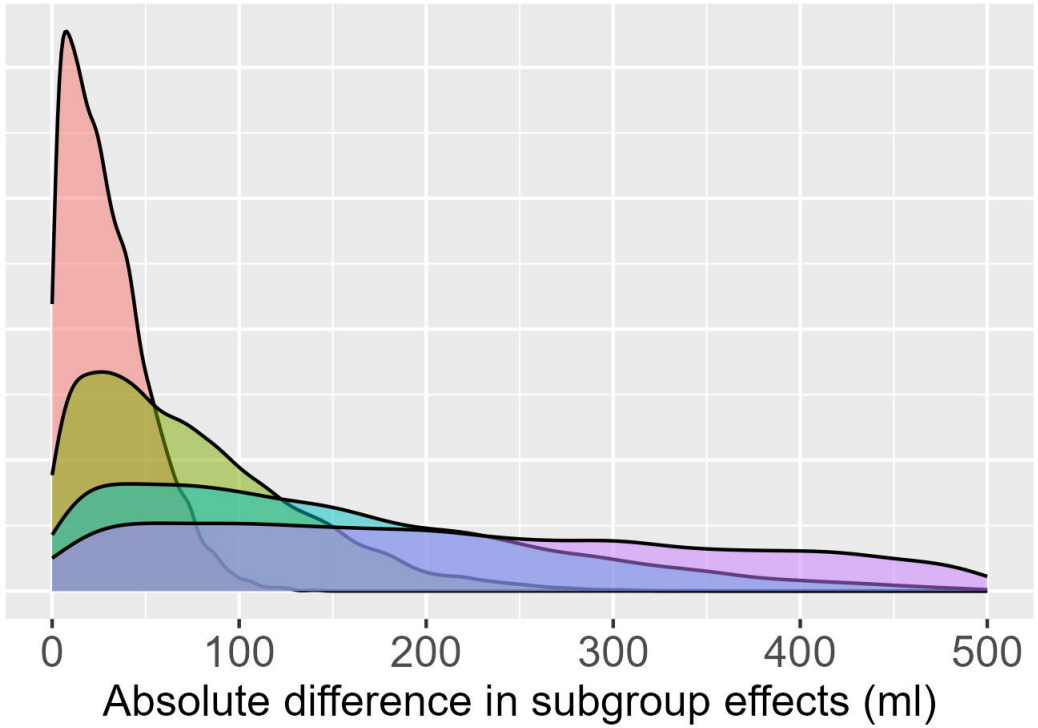
Prior on τ - Start by considering fixed values

Distribution of absolute difference $|\mu_j - \mu_k|$ between treatment effects in 2 randomly selected subgroups for different fixed values of τ

Assumed within-group sampling SD of endpoint = 260 ml; MCID = 100 ml

Prior on τ - Start by considering fixed values

Distribution of absolute difference $|\mu_j - \mu_k|$ between treatment effects in 2 randomly selected subgroups for different fixed values of τ



τ	Quantiles of distribution of $\|\mu_j - \mu_k\|$		
	2.5%	50%	97.5%
26 ml (0.1 SD)	1 ml	25 ml	82 ml
65 ml (0.25 SD)	3 ml	62 ml	206 ml
130 ml (0.5 SD)	6 ml	124 ml	412 ml
260 ml (1 SD)	12 ml	248 ml	824 ml

Assumed within-group sampling SD of endpoint = 260 ml; MCID = 100 ml

Prior on τ – choose a distribution to describe plausible range of values for τ

Half-Normal(ϕ) is often a reasonable choice of prior for the between-subgroup heterogeneity

(Rover et al (2021), Wang et al (2024), Spiegelhalter et al (2004))

Prior on τ

Half-Normal(ϕ) is often a reasonable choice of prior for the between-subgroup heterogeneity
(Rover et al (2021), Wang et al (2024), Spiegelhalter et al (2004))

Choosing value of ϕ:

- $\tau \sim HN(\phi)$ has median 0.67ϕ and 95% interval (0.03ϕ – 2.24ϕ)

Prior scale parameter, ϕ	Quantiles of between subgroup heterogeneity, τ		
	2.5% (0.03ϕ)	50% (0.67ϕ)	97.5% (2.24ϕ)
65	2	44	146
130	4	87	291
260	8	174	582

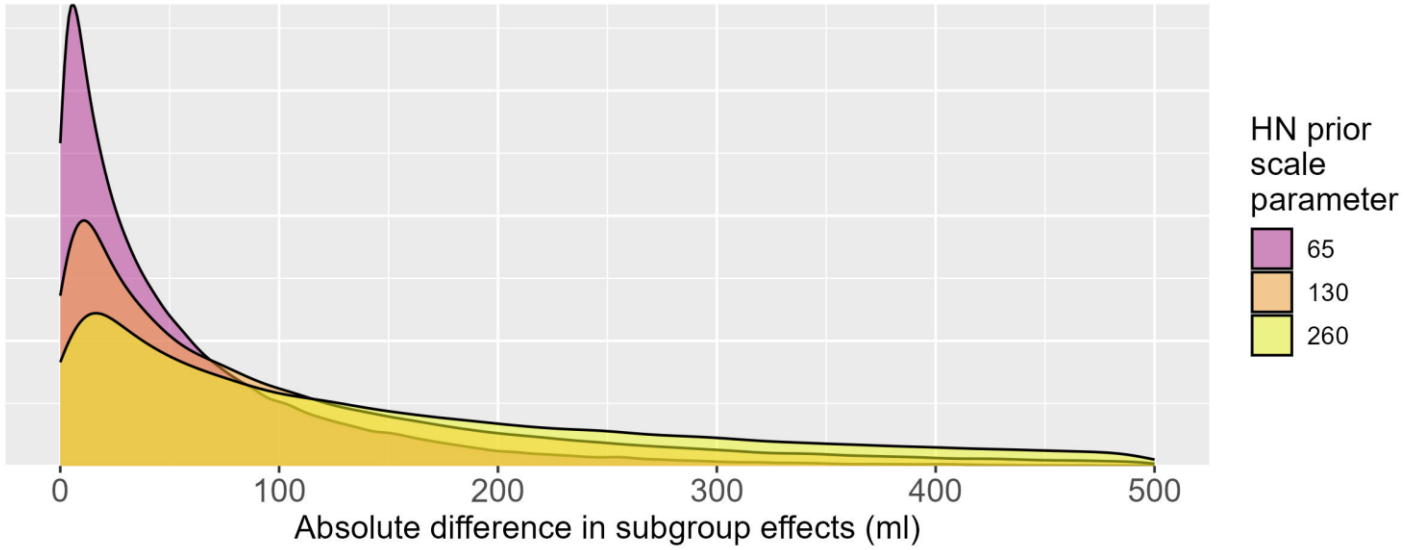
Prior on τ

Half-Normal(ϕ) is often a reasonable choice of prior for the between-subgroup heterogeneity
(Rover et al (2021), Wang et al (2024), Spiegelhalter et al (2004))

Choosing value of ϕ:

- $\tau \sim HN(\phi)$ has median 0.67ϕ and 95% interval (0.03ϕ – 2.24ϕ)
- Look at **induced prior** on $|\mu_j - \mu_k|$

Induced prior on difference in subgroup effects for different choices of scale parameter ϕ for Half Normal(ϕ) prior on τ



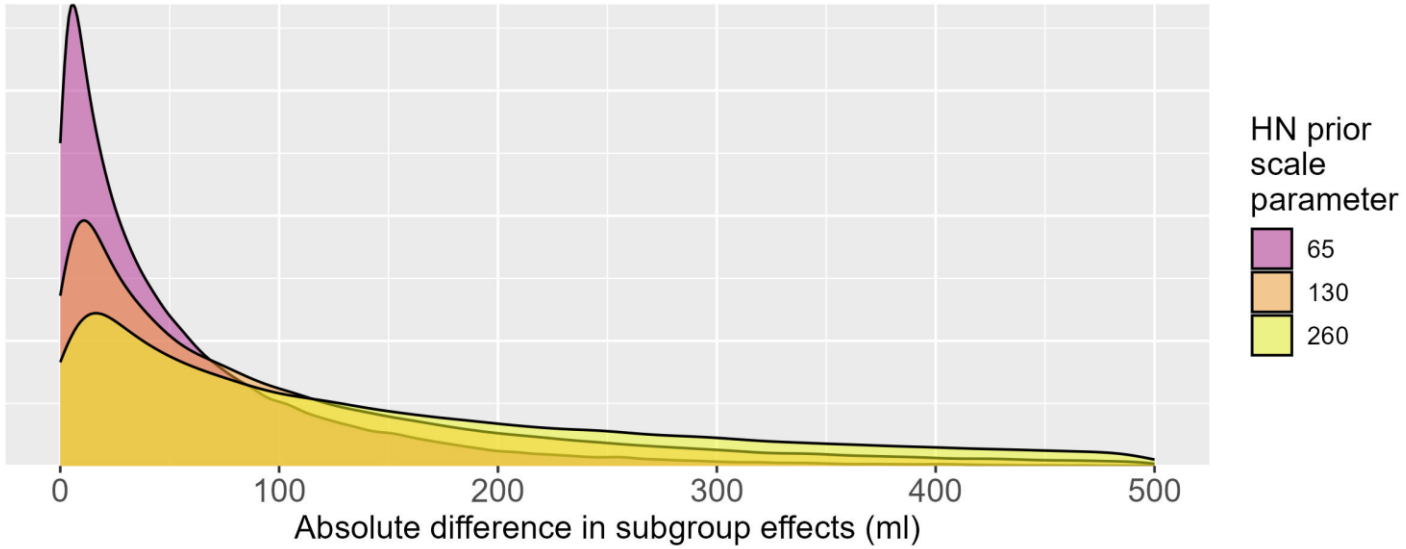
Prior on τ

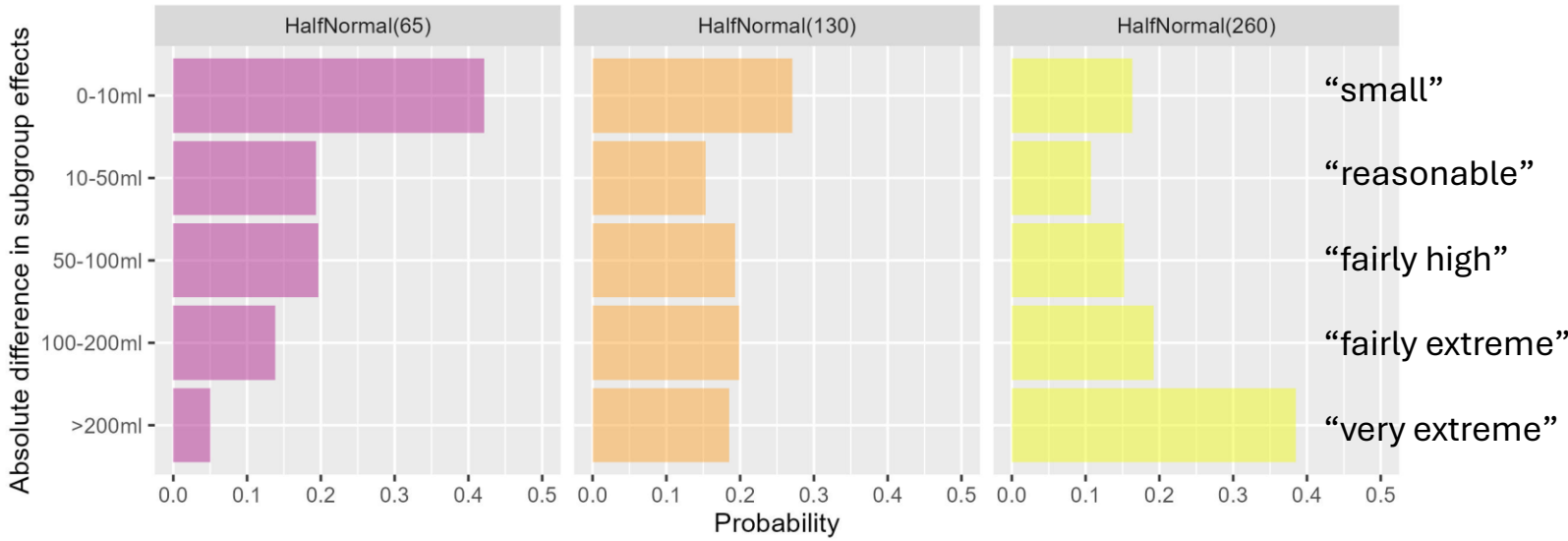
Half-Normal(ϕ) is often a reasonable choice of prior for the between-subgroup heterogeneity
(Rover et al (2021), Wang et al (2024), Spiegelhalter et al (2004))

Choosing value of ϕ:

- $\tau \sim HN(\phi)$ has median 0.67ϕ and 95% interval (0.03ϕ – 2.24ϕ)
- Look at **induced prior** on $|\mu_j - \mu_k|$

Induced prior on difference in subgroup effects for different choices of scale parameter ϕ for Half Normal(ϕ) prior on τ





Prior on τ

Half-Normal(ϕ) is often a reasonable choice of prior for the between-subgroup heterogeneity
(Rover et al (2021), Wang et al (2024), Spiegelhalter et al (2004))

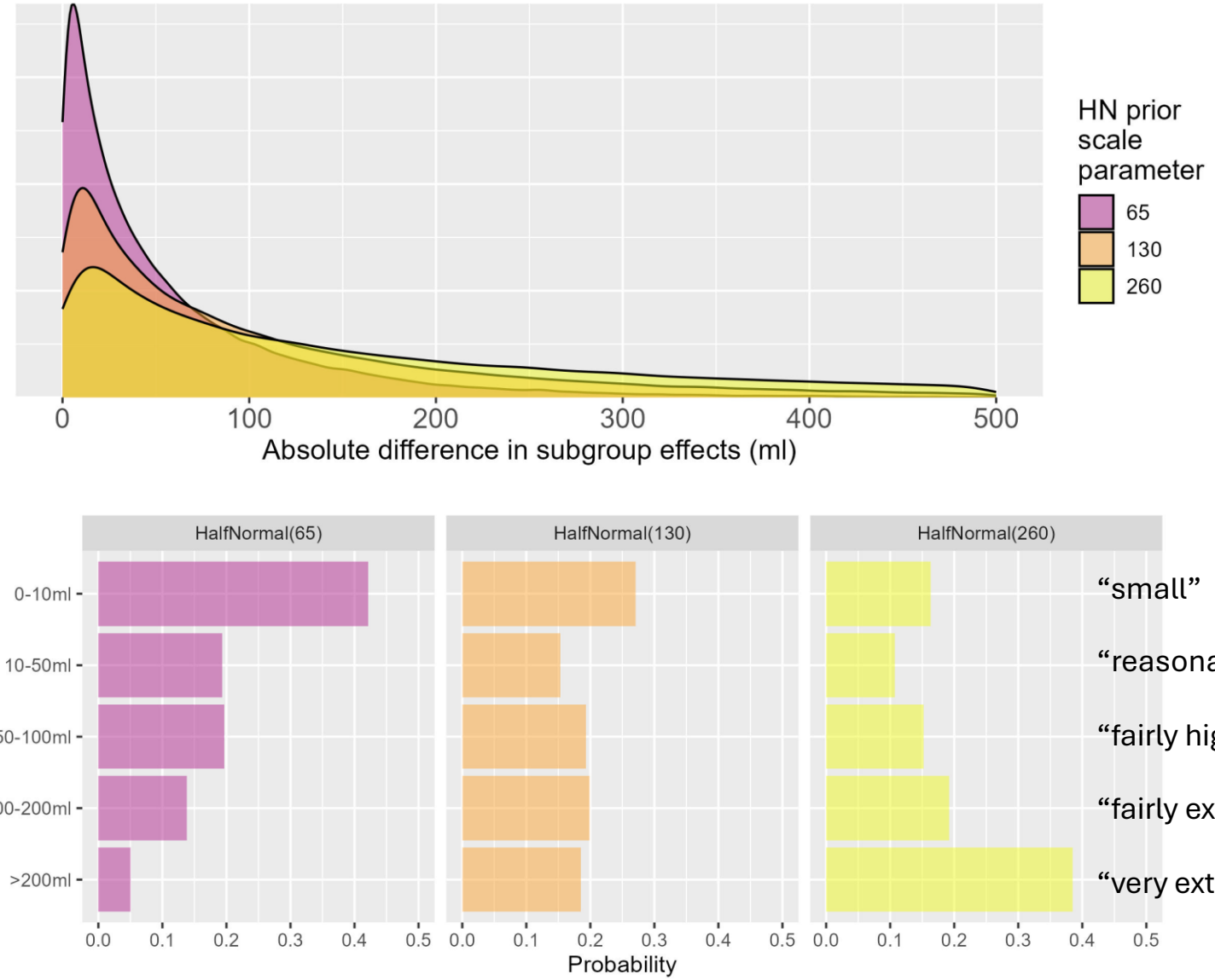
Choosing value of ϕ:

- $\tau \sim HN(\phi)$ has median 0.67ϕ and 95% interval (0.03ϕ – 2.24ϕ)
- Look at **induced prior** on $|\mu_j - \mu_k|$
- Elicit probability p s.t.

$$\mathbf{Pr}(|\mu_j - \mu_k| < \delta) = p$$

e.g. $\Pr(|\mu_j - \mu_k| < 100\ ml) = 0.5$
$\Rightarrow \phi = 194$

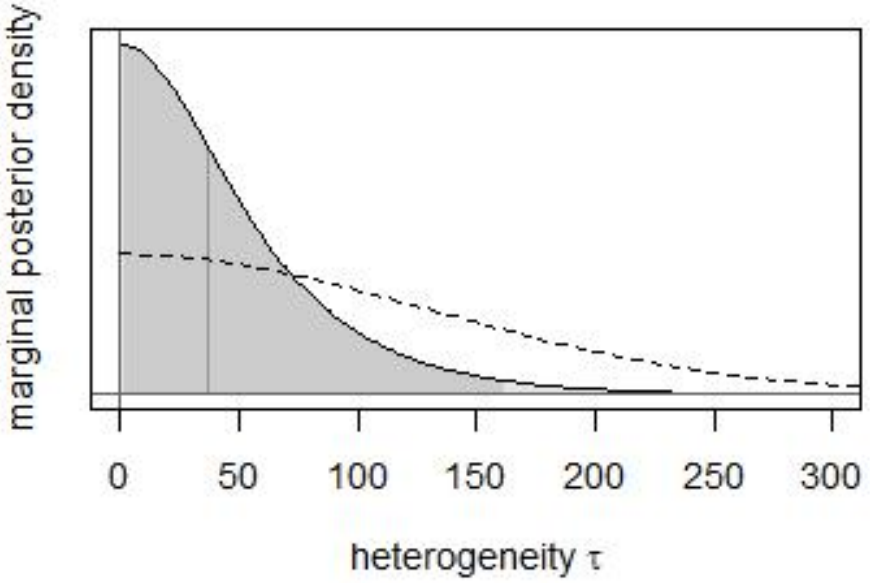
Induced prior on difference in subgroup effects for different choices of scale parameter ϕ for Half Normal(ϕ) prior on τ



Reporting

Primary analysis using $\tau \sim Half\ Normal(130)$ prior

Posterior of τ



prior – – – –

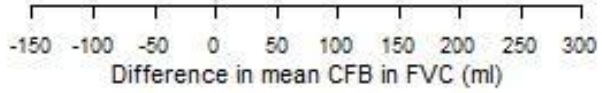
posterior median

posterior 95% CrI

quoted estimate shrinkage estimate

region	N	shrinkage estimate	95% CI
A	140	112	[42, 184]
B	60	121	[35, 222]
C	70	98	[5, 182]
D	30	94	[-31, 193]
Overall	**300**	**109**	**[50, 168]**

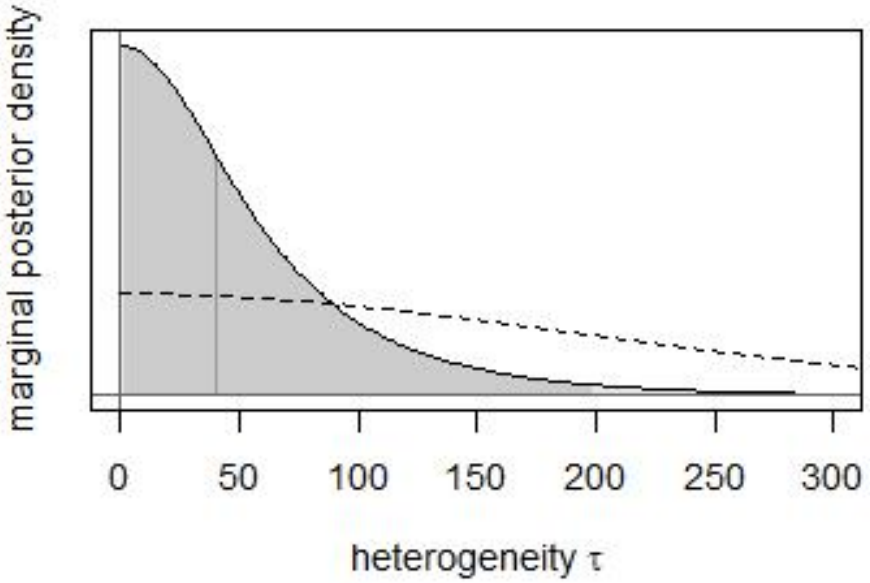
Heterogeneity (tau): 36.6 [1.6, 161.7]



Reporting

Sensitivity analysis using $\tau \sim Half\ Normal(194)$ prior

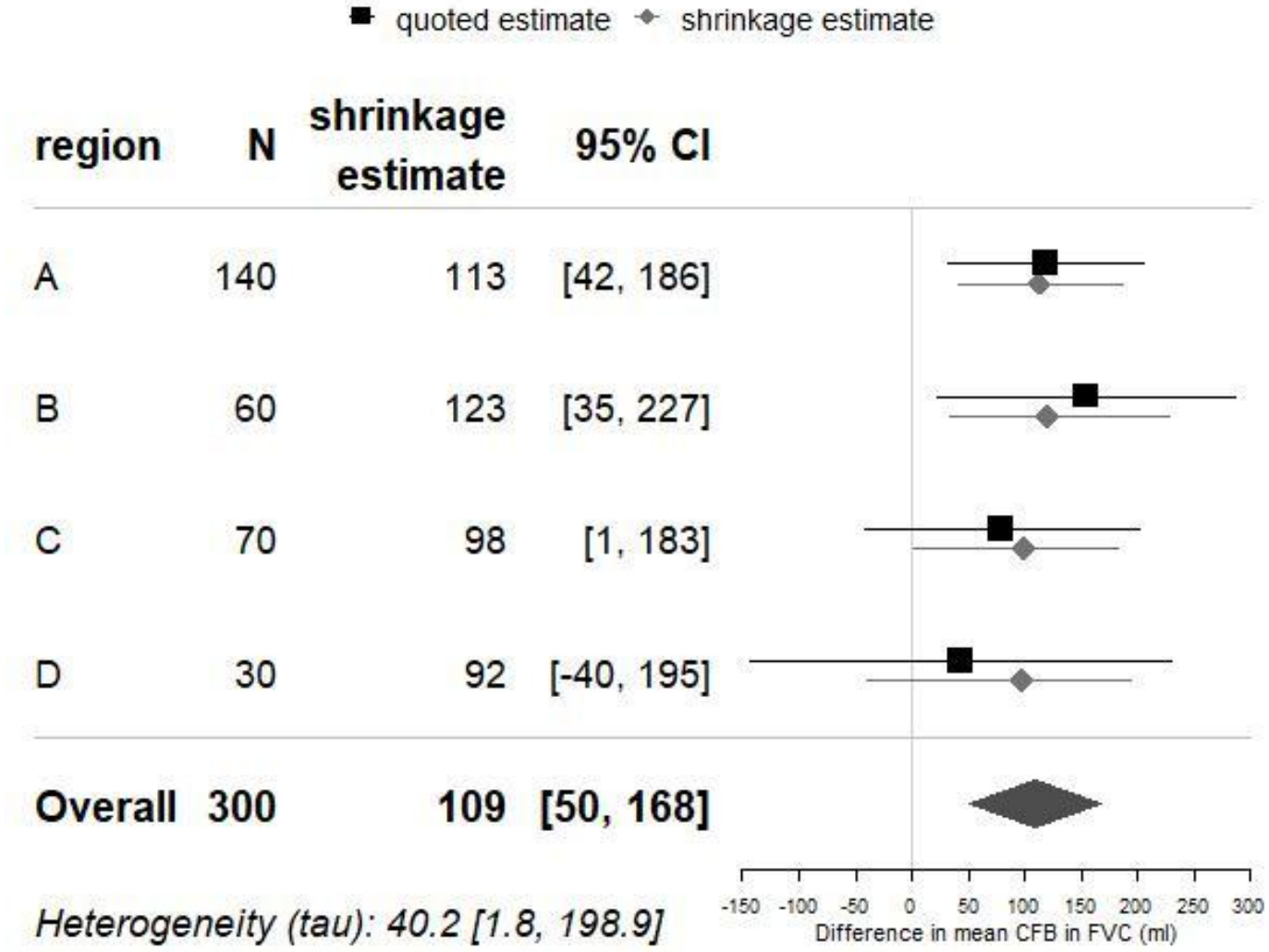
Posterior of τ



prior – – – –

posterior median

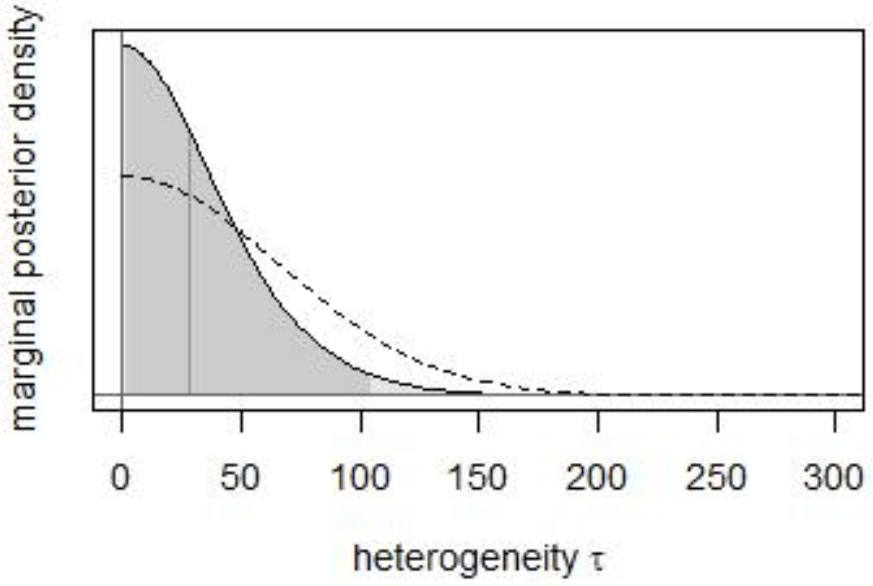
posterior 95% CrI



Reporting

Sensitivity analysis using $\tau \sim Half\ Normal(65)$ prior

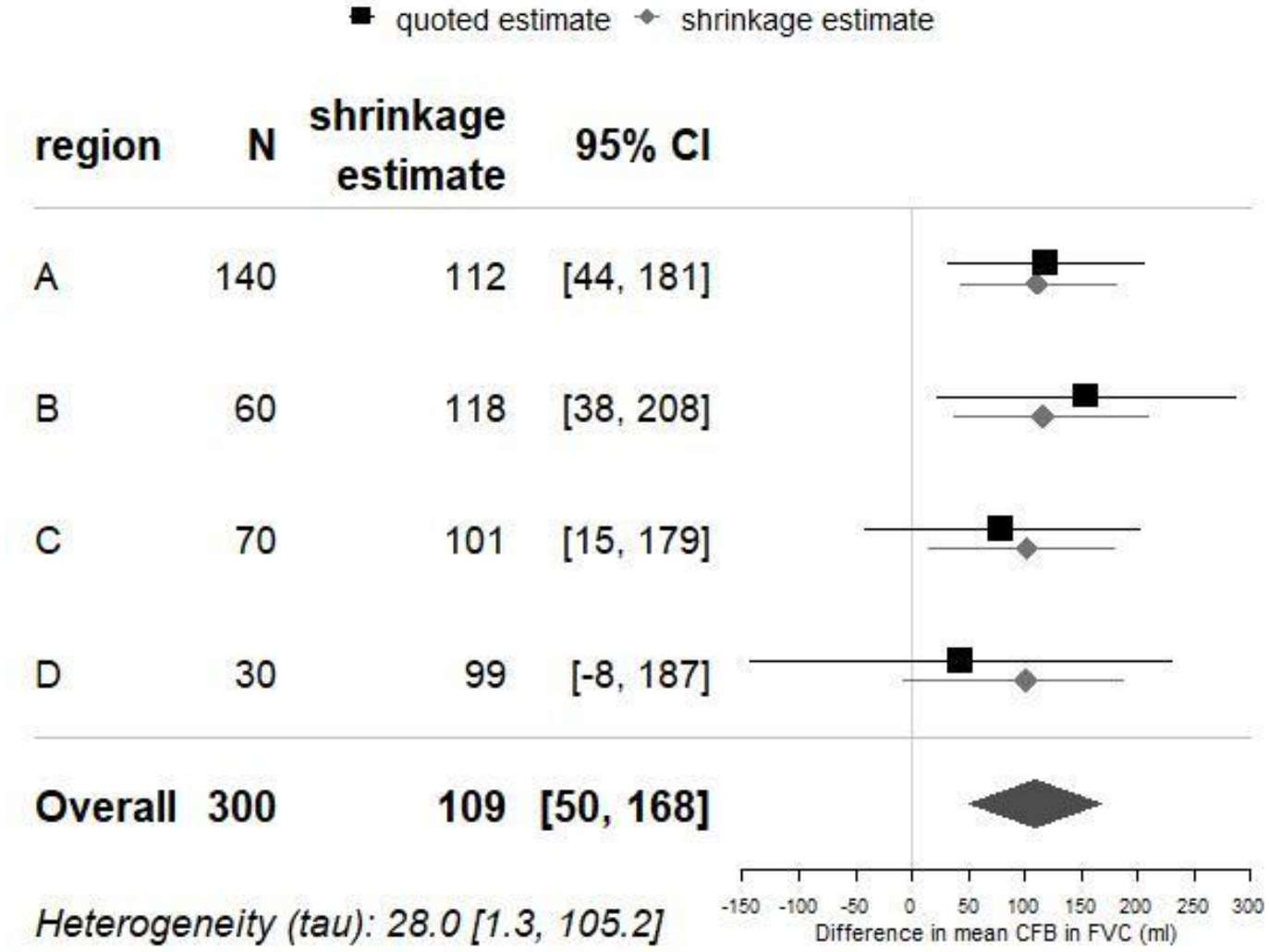
Posterior of τ



prior – – – –

posterior median

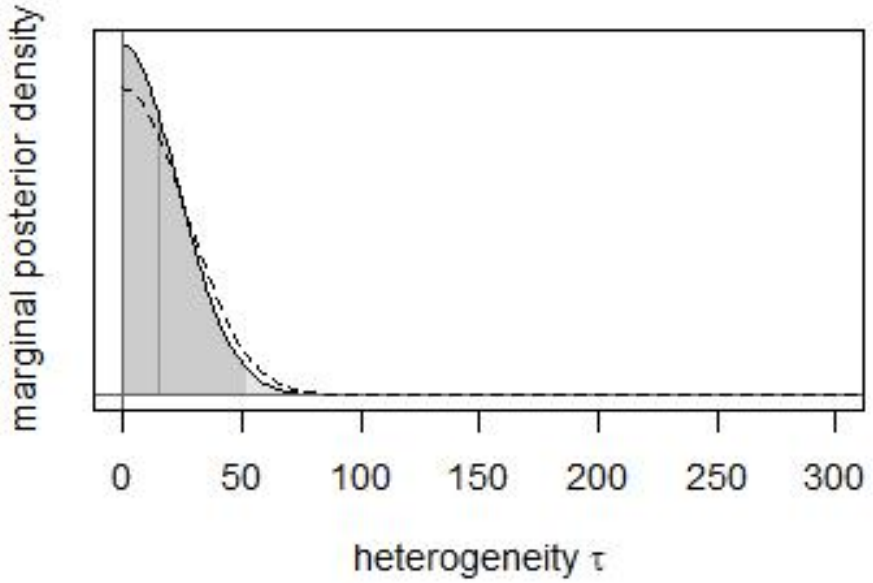
posterior 95% CrI



Reporting

Sensitivity analysis using $\tau \sim Half\ Normal(26)$ prior

Posterior of τ



prior – – – –

posterior median

posterior 95% CrI

quoted estimate shrinkage estimate

region	N	shrinkage estimate	95% CI
A	140	110	[47, 174]
B	60	113	[44, 185]
C	70	106	[35, 173]
D	30	105	[28, 177]
Overall	**300**	**109**	**[50, 168]**

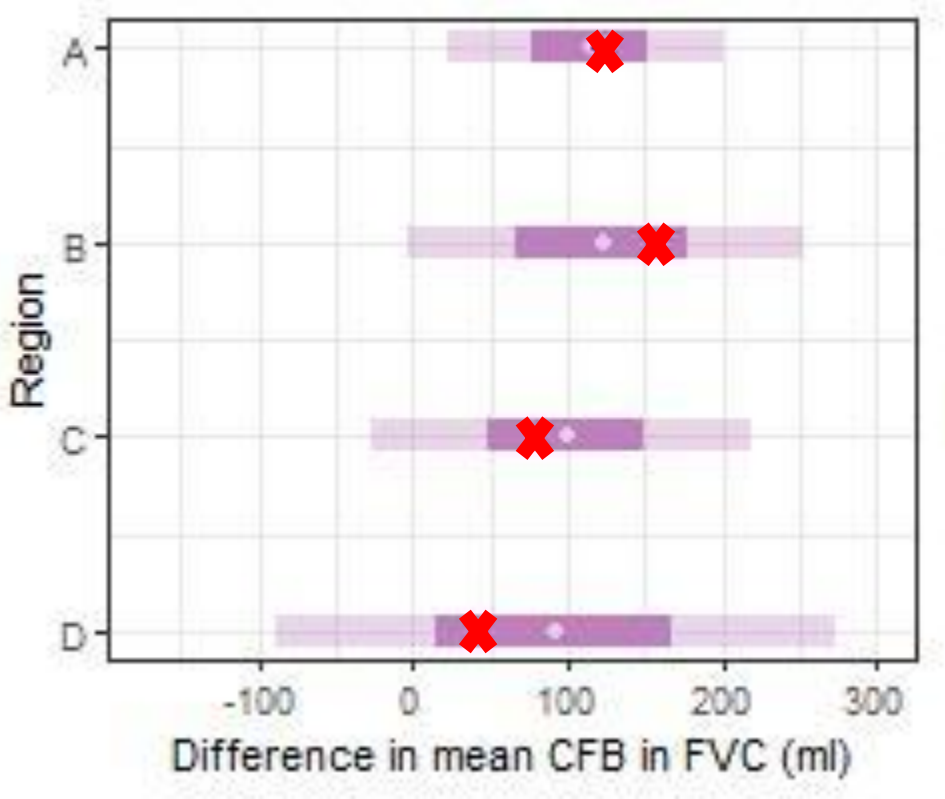
Heterogeneity (tau): 15.40 [0.71, 52.17]

-150 -100 -50 0 50 100 150 200 250 300
Difference in mean CFB in FVC (ml)

Reporting

Posterior predictive checks

$\tau \sim$ HN(130) prior



Actual effect observed in each subgroup

y

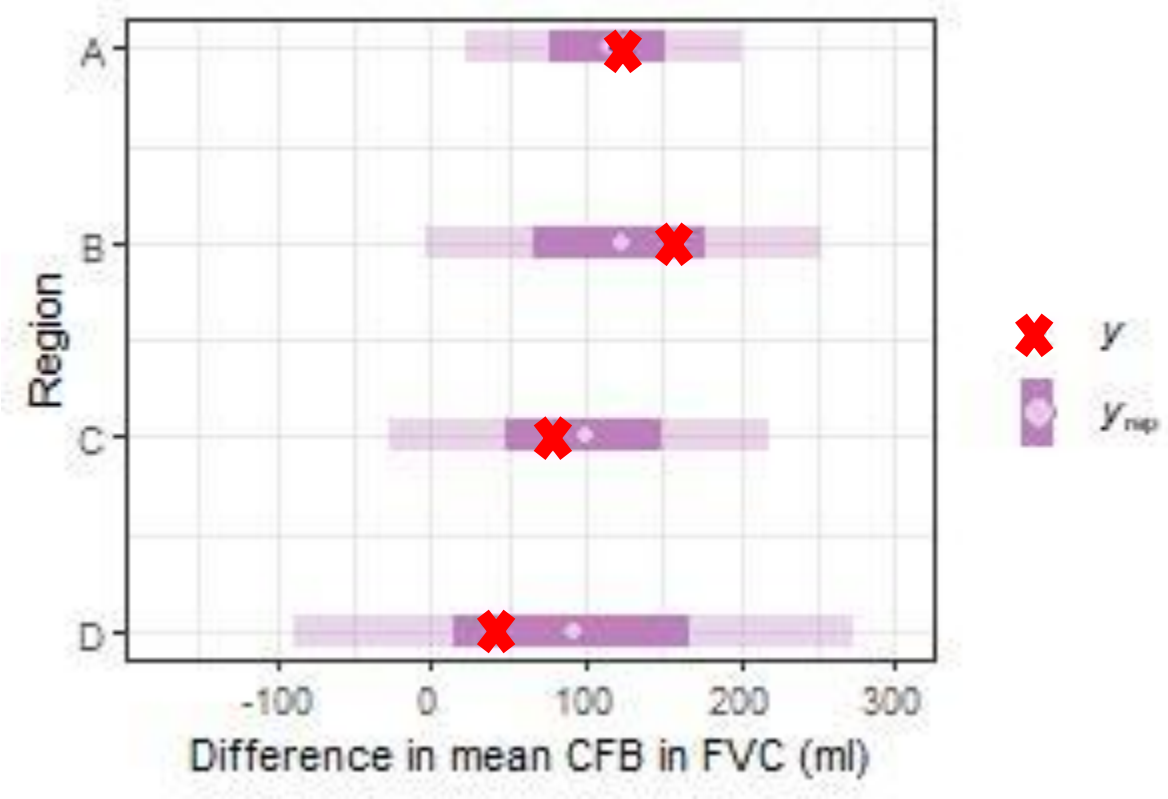
y_{rep}

Posterior predictive distribution for observed effect in each subgroup (median, 50% and 90% interval)

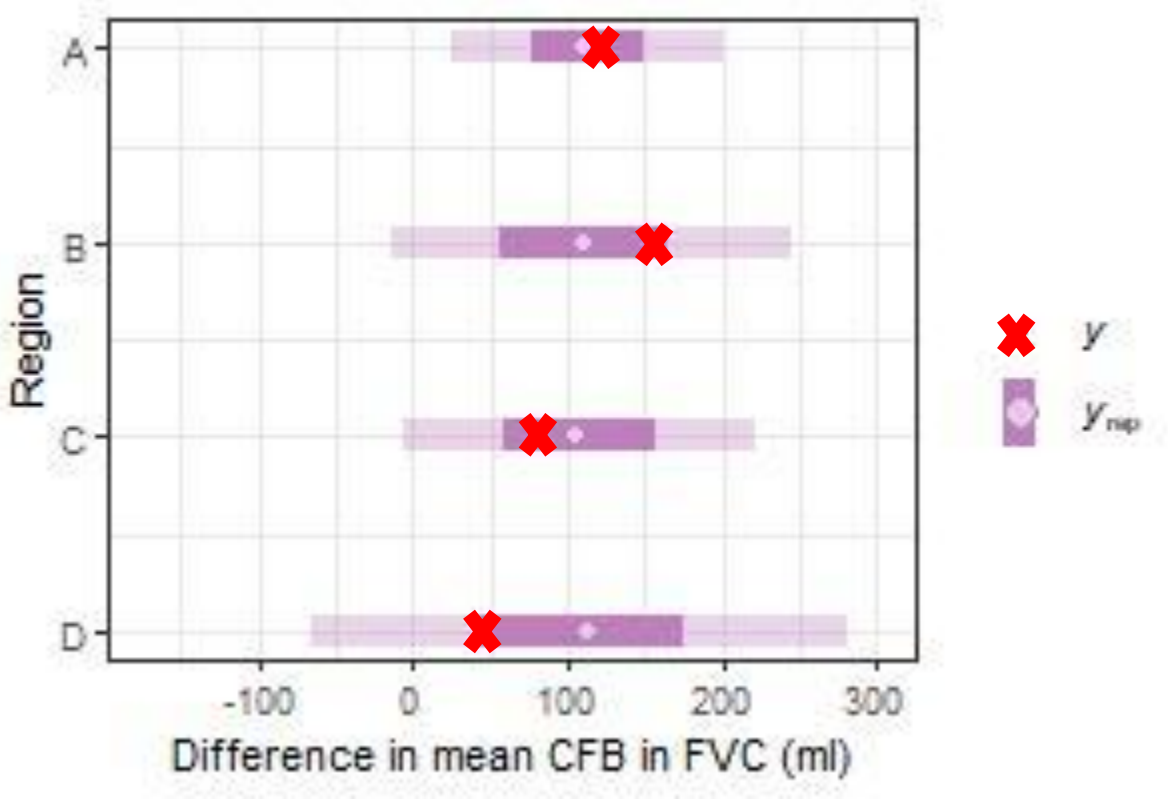
Reporting

Posterior predictive checks

$\tau \sim$ HN(130) prior



$\tau \sim$ HN(26) prior



Example 2: Bayesian Dynamic Borrowing in a Paediatric Lupus Trial

Context of Use: Clinical & Statistical Contexts

Context

- Rare disease with a **drug already approved in adults**
- Planned **paediatric** study: randomized controlled clinical trial
- External data available: two replicate **Phase 3 studies in adults**
- Motivation for using external data: supplement the planned pediatric sample size with adult data to **improve efficiency and increase precision of evidence for decision-making**

Context of Use: Clinical & Statistical Contexts

Context

- Rare disease with a **drug already approved in adults**
- Planned **paediatric** study: randomized controlled clinical trial
- External data available: two replicate **Phase 3 studies in adults**
- Motivation for using external data: supplement the planned pediatric sample size with adult data to **improve efficiency and increase precision of evidence for decision-making**

Clinical context

- **Very low feasibility** of recruiting pediatric patients
- **High unmet medical need** in the pediatric population
- Good **biological and clinical rationale for transportability** from adults to children

Context of Use: Clinical & Statistical Contexts

Context

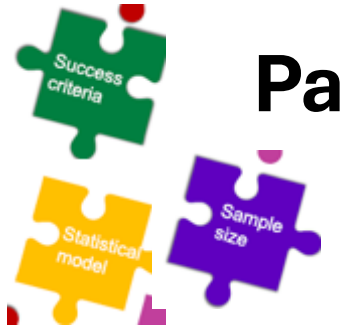
- Rare disease with a **drug already approved in adults**
- Planned **paediatric** study: randomized controlled clinical trial
- External data available: two replicate **Phase 3 studies in adults**
- Motivation for using external data: supplement the planned pediatric sample size with adult data to **improve efficiency and increase precision of evidence for decision-making**

Clinical context

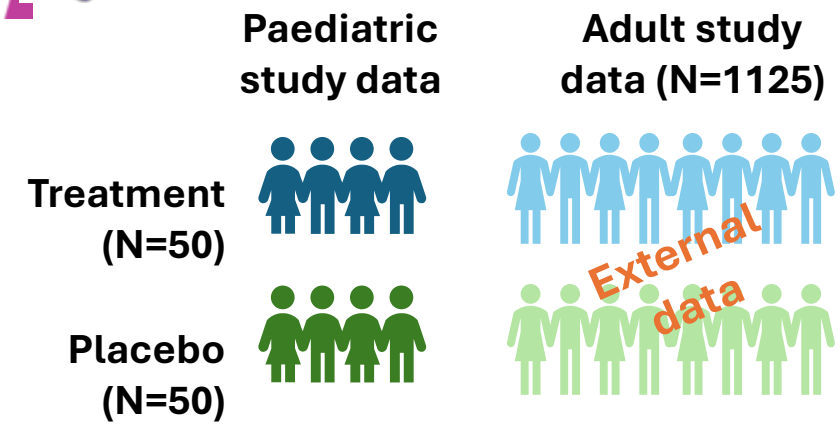
- **Very low feasibility** of recruiting pediatric patients
- **High unmet medical need** in the pediatric population
- Good **biological and clinical rationale for transportability** from adults to children

Statistical context

- **High-quality data from adults**
- **Similar trial design**, strata, endpoints etc
- **Subjective assumption of transportability** from adults to children



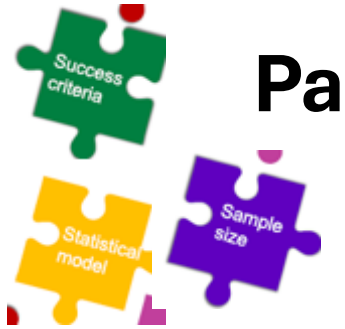
Paediatric Trial – Bayesian Dynamic Borrowing (BDB) Design



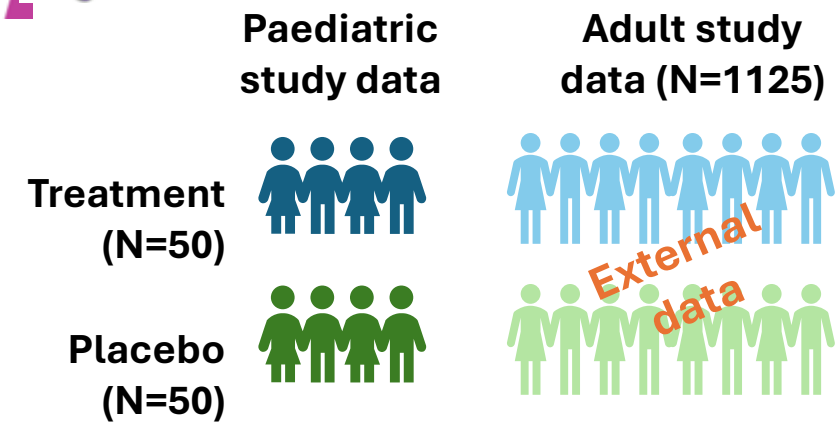
Primary endpoint: Disease activity responder index

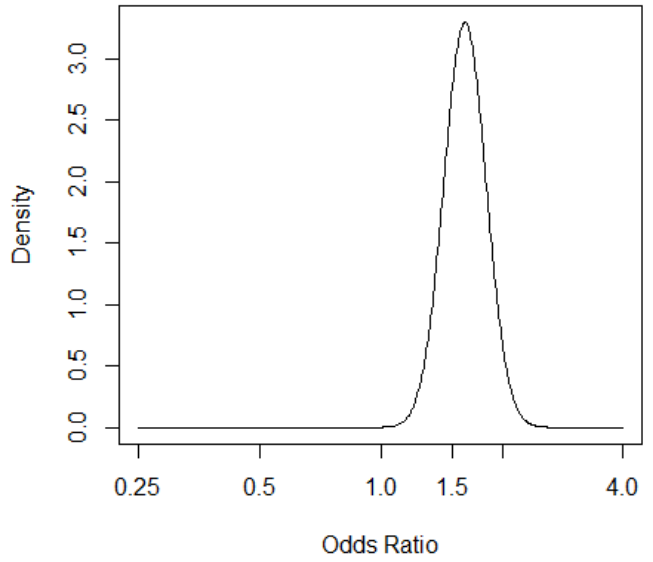
Treatment contrast: Odds Ratio for active v placebo (assumed Normally distributed on log scale)

Success: Pr(OR > 1 | data, prior) > 97.5%



Paediatric Trial – Bayesian Dynamic Borrowing (BDB) Design





Prior mean	1.62
Prior median	1.62
Prior 95% Credible Interval for OR	(1.27, 2.05)
Prior Probability OR > 1	0.99996

Primary endpoint: Disease activity responder index

Treatment contrast: Odds Ratio for active v placebo (assumed Normally distributed on log scale)

Success: Pr(OR > 1 | data, prior) > 97.5%

Robust mixture prior for paediatric log OR: weighted mixture of posterior distribution of treatment effect from **adult study (weight w)** and **vague distribution (weight 1-w)** centered on 0 with unit info variance

Elicitation of prior weight

- Multiple experts in relevant disease area with clinical experience of treating adults and/or children
- Review
  - available data from the adult studies, PK etc
  - comparability of study designs
  - similarities between adults and children based on experience and relevant literature

Elicitation of prior weight

- Multiple experts in relevant disease area with clinical experience of treating adults and/or children
- Review
  - available data from the adult studies, PK etc
  - comparability of study designs
  - similarities between adults and children based on experience and relevant literature

- How much confidence do you have in applying the adult clinical trial data to make decisions on treatment effect in children?

0	1	2	3	4	5	6	7	8	9	10
Ignore adult data as irrelevant to paediatric population										Fully trust adult data as applicable to paediatric population

Elicitation of prior weight

- Multiple experts in relevant disease area with clinical experience of treating adults and/or children
- Review
  - available data from the adult studies, PK etc
  - comparability of study designs
  - similarities between adults and children based on experience and relevant literature

- How much confidence do you hav…nical trial data to make decisions on treatment effec…

Average score = 7

Prior weight on adult data = 0.7

0	1	2	3	4	…	9	10
Ignore adult data as irrelevant to paediatric population							Fully trust adult data as applicable to paediatric population

Assessing (in)correctness of decisions

Metric		Comments	Paradigm
Pr(+ve Decision | Truth = null)	Type 1 error	**Hypothetical probabilities** of making future decisions given **fixed truths**	Frequentist
Pr(+ve Decision| Truth = MCID)	Power	**Hypothetical probabilities** of making future decisions given **fixed truths**	Frequentist

See Sections 1.3 and 9.3 of Frank Harrell's online course https://hbiostat.org/bayes/bet

Assessing (in)correctness of decisions

Metric		Comments	Paradigm
Pr(+ve Decision | Truth = null)	Type 1 error	Hypothetical probabilities of making future decisions given fixed truths	Frequentist
Pr(+ve Decision| Truth = MCID)	Power	Hypothetical probabilities of making future decisions given fixed truths	Frequentist
Pr(+ve Decision)	Assurance	**Predicted probability** of making a positive decision	Hybrid – requires **design (sampling) prior for true effect**
Pr(+ve Decision AND Truth = null)*	Joint probability that null is true and that a positive decision is made	**Predicted probability** of making a false positive decision	Hybrid - requires **design (sampling) prior for true effect**

*Best et al (2024) Beyond classical type I error: Bayesian metrics for Bayesian Designs using Informative Priors. *J Biopharm Stats*

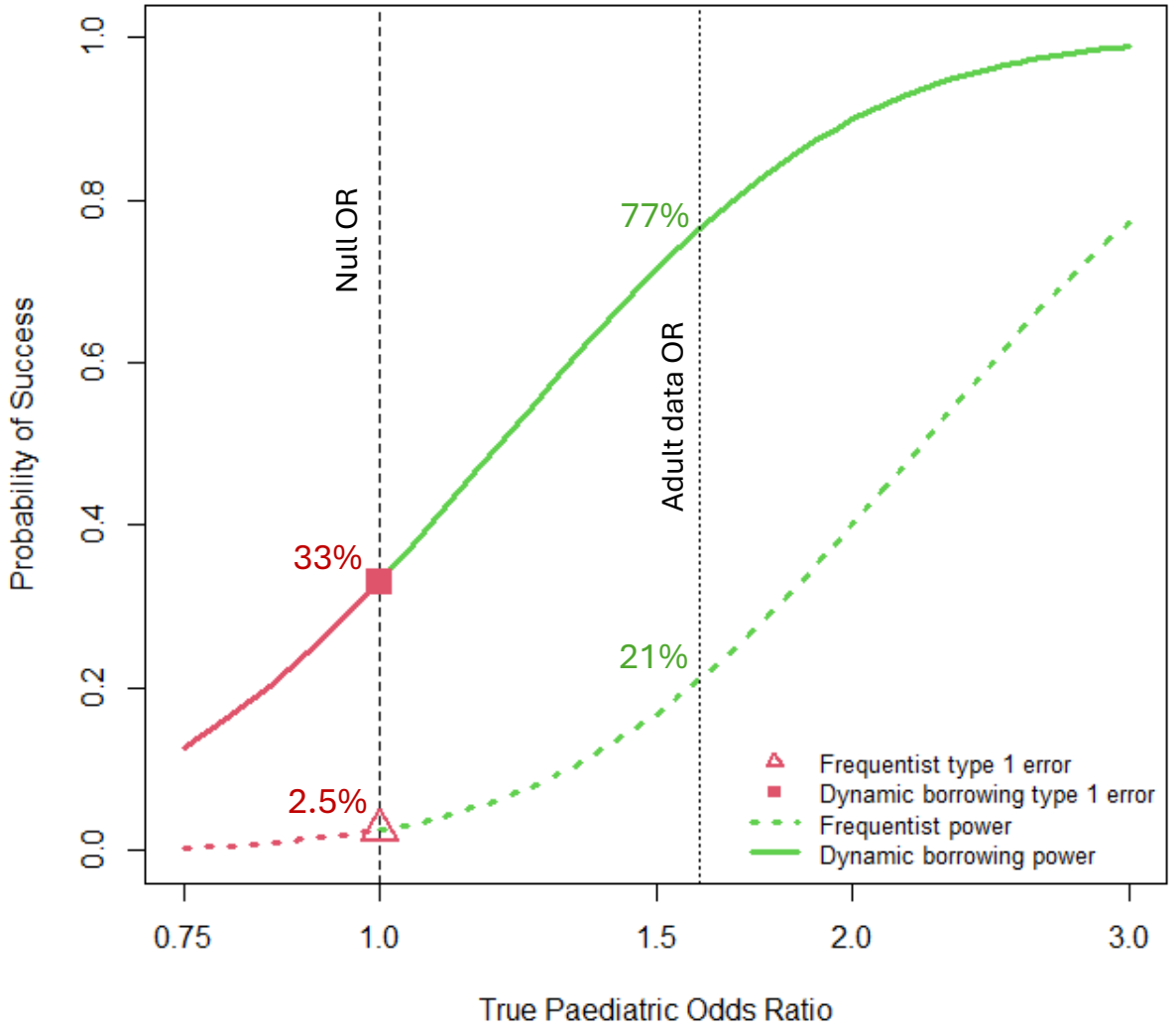
Assessing (in)correctness of decisions

Metric		Comments	Paradigm
Pr(+ve Decision | Truth = null)	Type 1 error	Hypothetical probabilities of making future decisions given fixed truths	Frequentist
Pr(+ve Decision| Truth = MCID)	Power	Hypothetical probabilities of making future decisions given fixed truths	Frequentist
Pr(+ve Decision)	Assurance	Predicted probability of making a positive decision	Hybrid – requires design (sampling) prior for true effect
Pr(+ve Decision AND Truth = null)	Joint probability that null is true and that a positive decision is made	Predicted probability of making a false positive decision	Hybrid - requires design (sampling) prior for true effect
Pr(Truth = null | +ve Decision)	Probability that a positive decision is incorrect (i.e. decision is a false positive)	**Judgement of incorrectness of decision at time decision is made**. Equals (1 – posterior prob of efficacy) if success rule is met	Bayesian

See Sections 1.3 and 9.3 of Frank Harrell's online course https://hbiostat.org/bayes/bet

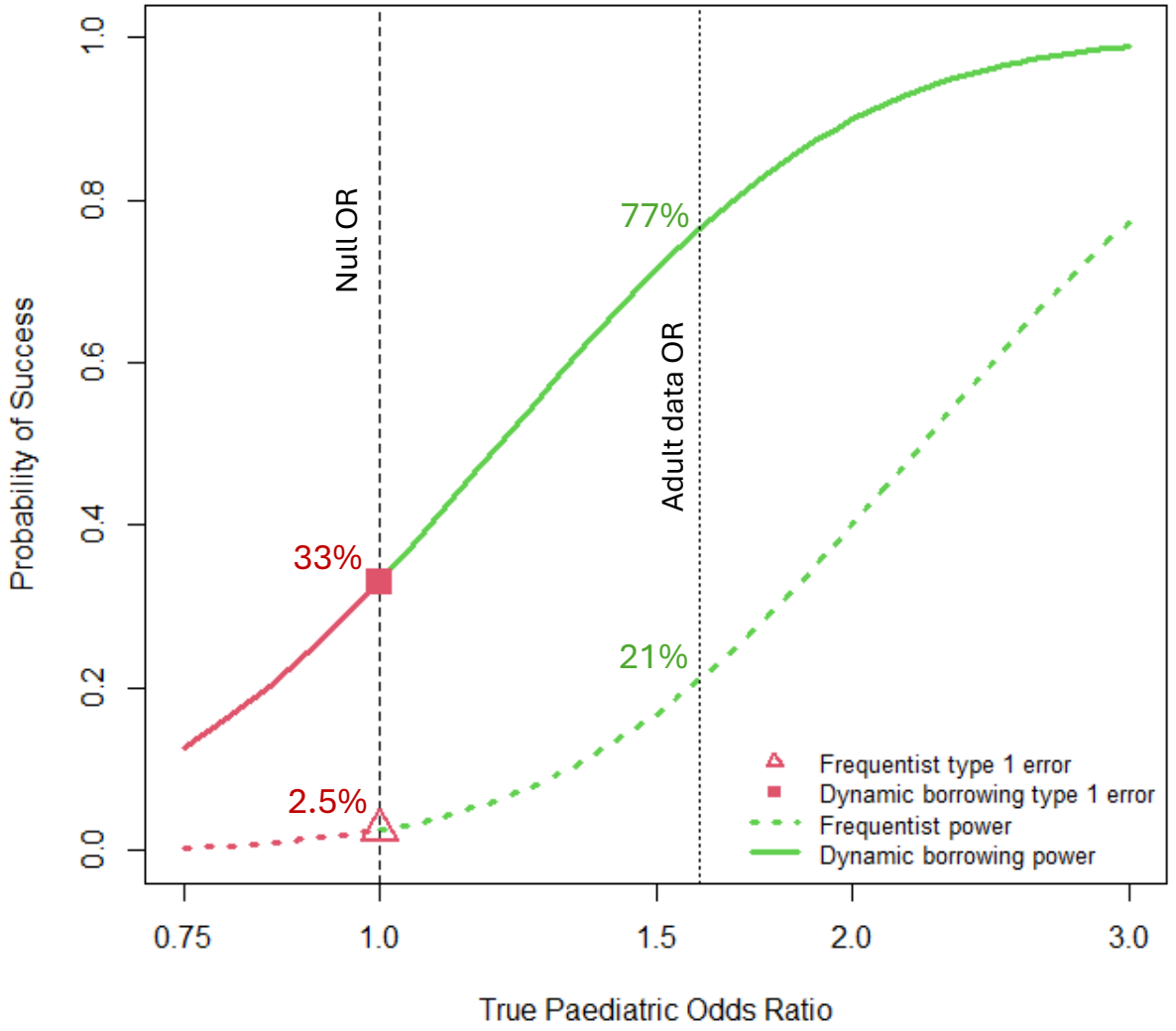
Frequentist operating characteristics

- Borrowing (+ve) information on the treatment effect inflates type 1 error



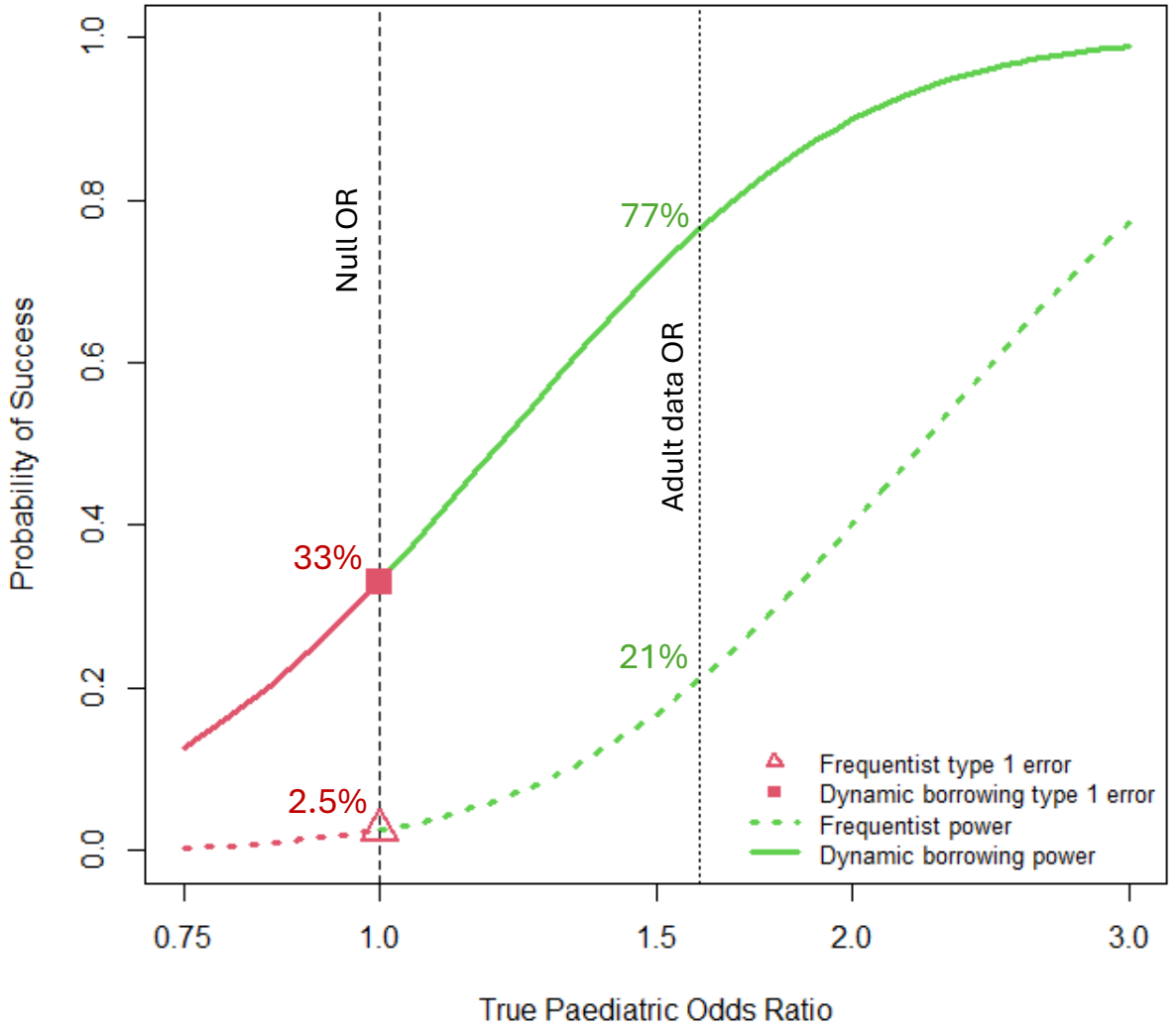
Frequentist operating characteristics

- Borrowing (+ve) information on the treatment effect inflates type 1 error
- Kopp-Schneider et al (2024): calibrate (frequentist) test without borrowing to type 1 error of borrowing design
  - ➢No power gains are possible
  - ➢But still potential gains in other OC (next slide)



Frequentist operating characteristics

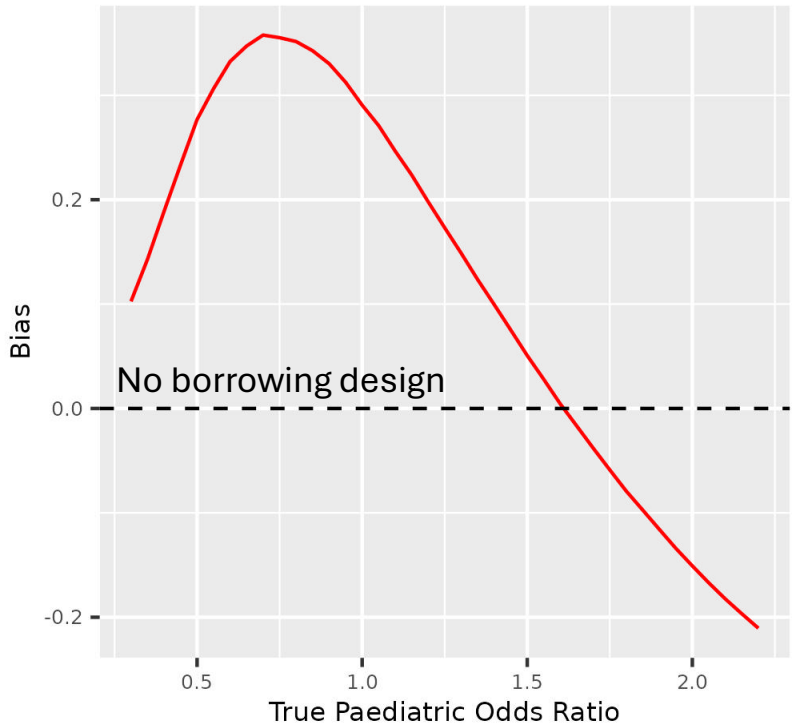
- Borrowing (+ve) information on the treatment effect inflates type 1 error
- Kopp-Schneider et al (2024): calibrate (frequentist) test without borrowing to type 1 error of borrowing design
  - ➢No power gains are possible
  - ➢But still potential gains in other OC (next slide)
  - ➢Assumes all possible values of parameter space are equally important
  - ➢Power gains from borrowing possible if we are willing to consider some regions of the parameter space as more important than others
    - ➢restrict (or weight) operating characteristics to that region



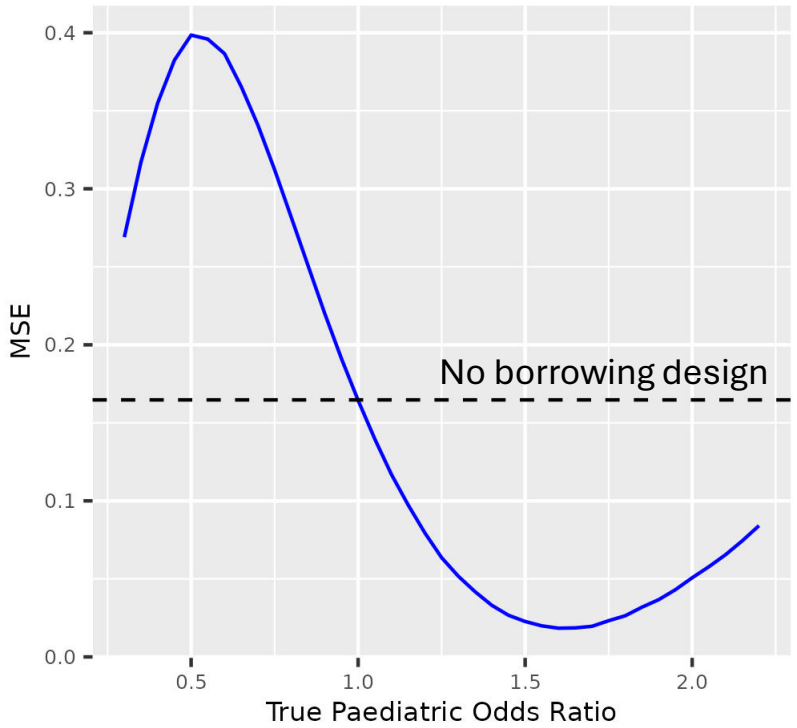
Frequentist operating characteristics

Precision

Bias



MSE



Design (sampling) priors

2 types of prior

- **Analysis prior (A):** pre-specified analysis prior for treatment effect parameter
- **Sampling prior (S):** design (or simulation) prior
  - Mechanism for generating data scenarios to evaluate operating characteristics of trial designs
  - If S ≠ A → can be used to judge accuracy of decisions under a prior which is different from the analysis (i.e. sponsor's) prior

Design (sampling) priors

2 types of prior

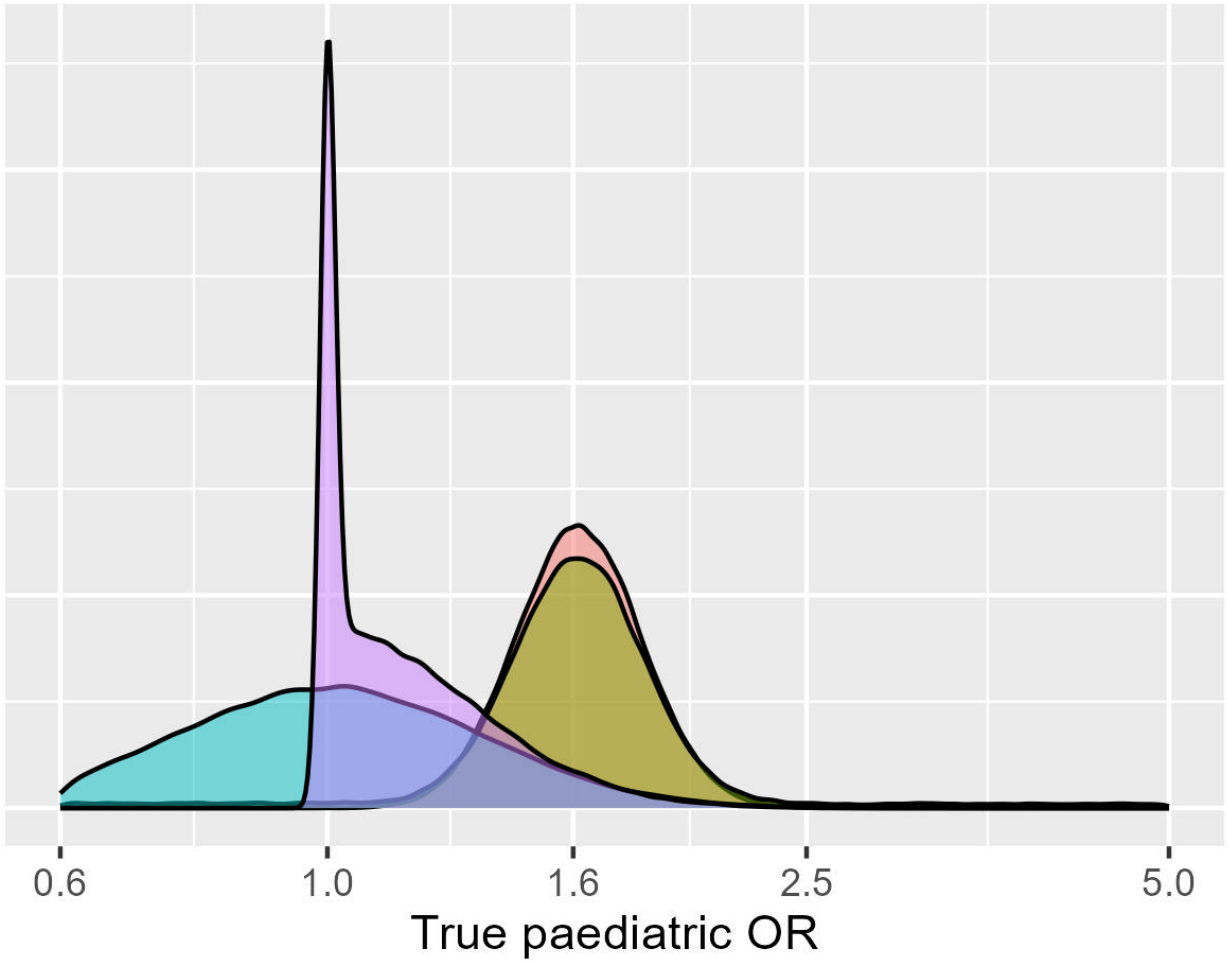
- **Analysis prior (A):** pre-specified analysis prior for treatment effect parameter
- **Sampling prior (S):** design (or simulation) prior
  - Mechanism for generating data scenarios to evaluate operating characteristics of trial designs
  - If $S \neq A \rightarrow$ can be used to judge accuracy of decisions under a prior which is different from the analysis (i.e. sponsor's) prior
- **Choosing S**
  - If solid agreement between sponsors and regulators about the prior, then may be sufficient to choose $S = A$
  - In general, consider various $S \neq A$ to assess how easy it is for accuracy of conclusions to be below acceptable level

Design (sampling) priors

2 types of prior

- **Analysis prior (A):** pre-specified analysis prior for treatment effect parameter
- **Sampling prior (S):** design (or simulation) prior
  - Mechanism for generating data scenarios to evaluate operating characteristics of trial designs
  - If S ≠ A → can be used to judge accuracy of decisions under a prior which is different from the analysis (i.e. sponsor's) prior
- **Choosing S**
  - If solid agreement between sponsors and regulators about the prior, then may be sufficient to choose S = A
  - In general, consider various S ≠ A to assess how easy it is for accuracy of conclusions to be below acceptable level
  - S often more sceptical than A – can be based on:
    - Data, e.g. select least favourable previous trial, or shift mean downwards
    - Expert elicitation
    - "Reference sceptical prior" (Spiegelhalter at al 1994), e.g. mean 0, small prob of treatment effect > alternative
    - Note: S = point mass (at null or alternative) → standard type 1 error and power calculations

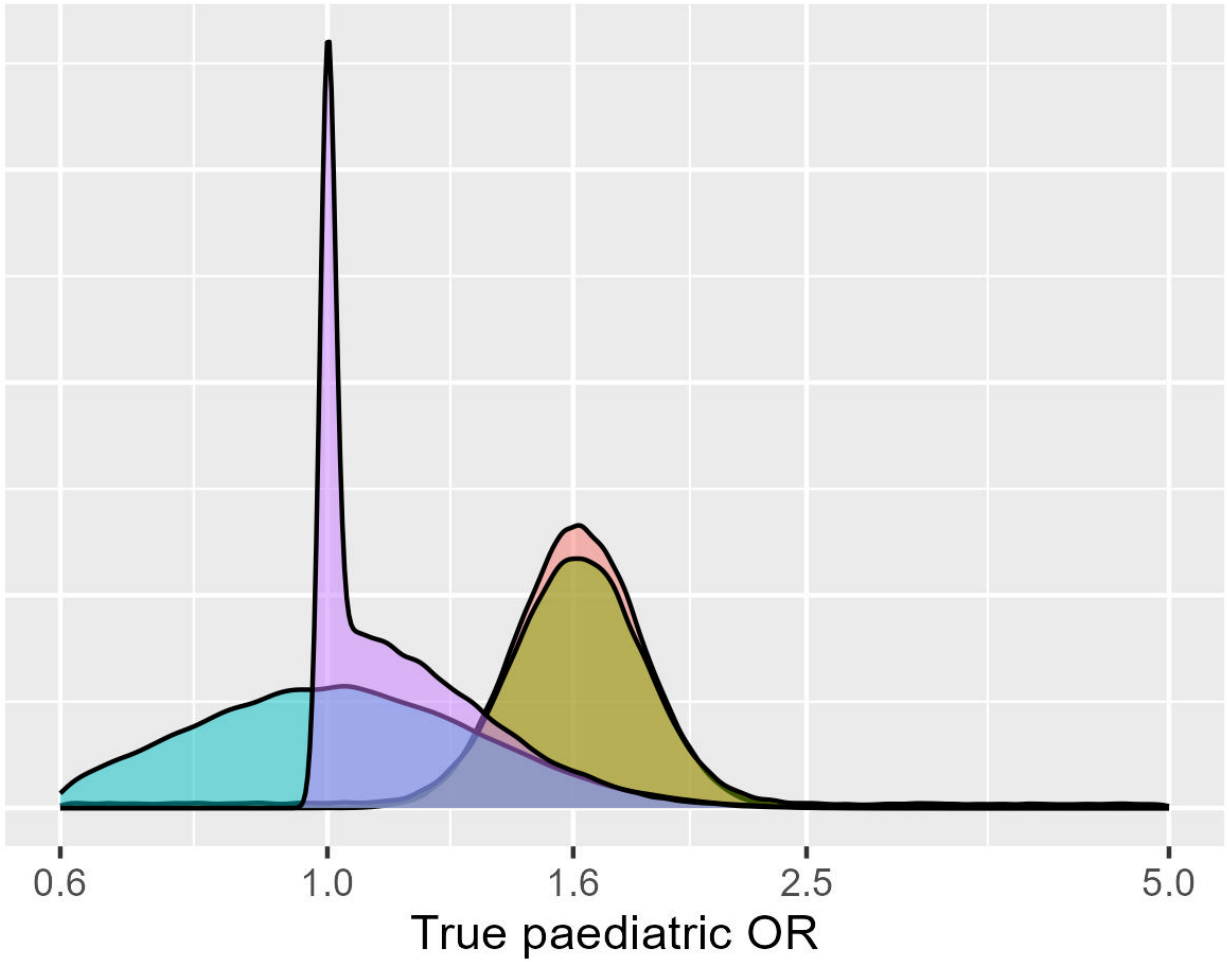
Design priors for paediatric example



Design prior

Adult	**Adult prior fully relevant**
RMP	**S = A**
Sceptical 1	**Mean 0, Pr(OR>1.6) = 0.05**
Sceptical 2	**Pr(OR=1) = 0.3, Pr(OR>1.6) = 0.05**

Design priors for paediatric example

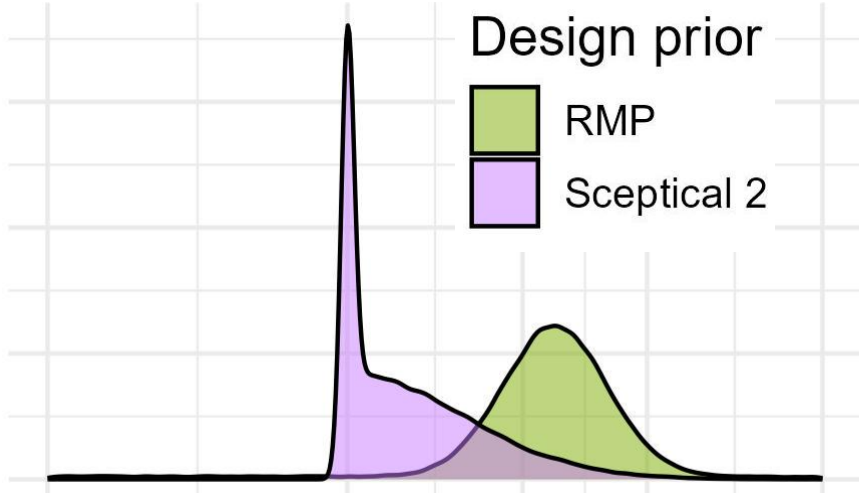


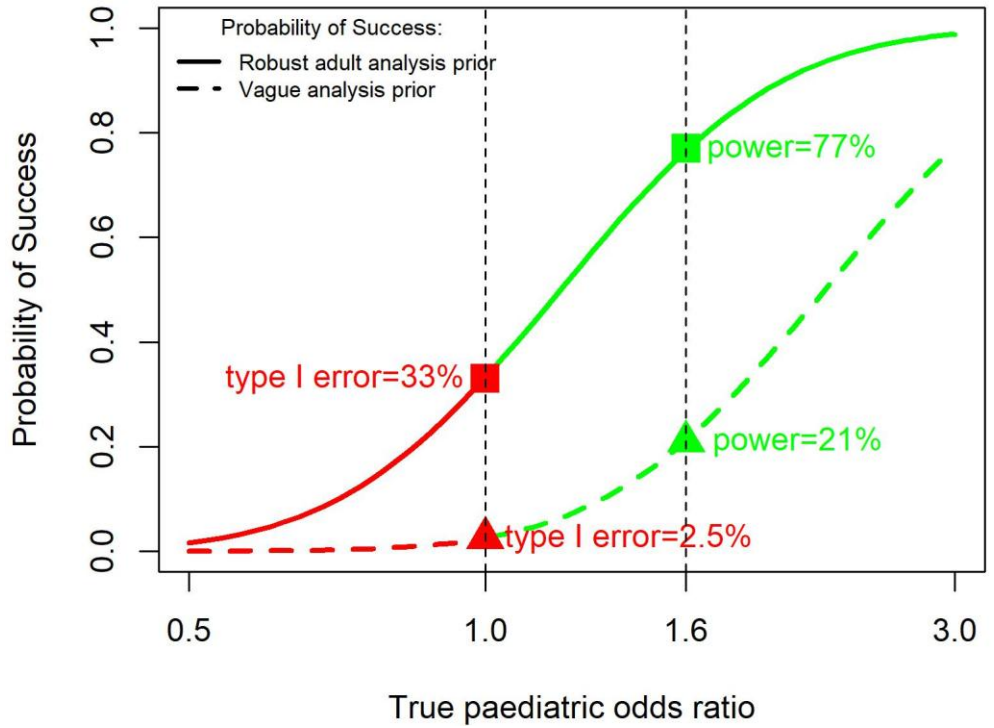
Design prior

Adult	**Adult prior fully relevant**
RMP	**S = A**
Sceptical 1	**Mean 0, Pr(OR>1.6) = 0.05**
Sceptical 2	**Pr(OR=1) = 0.3, Pr(OR>1.6) = 0.05**

✓ **Upfront discussion and alignment between sponsor and regulator to agree what design scenarios are possible**

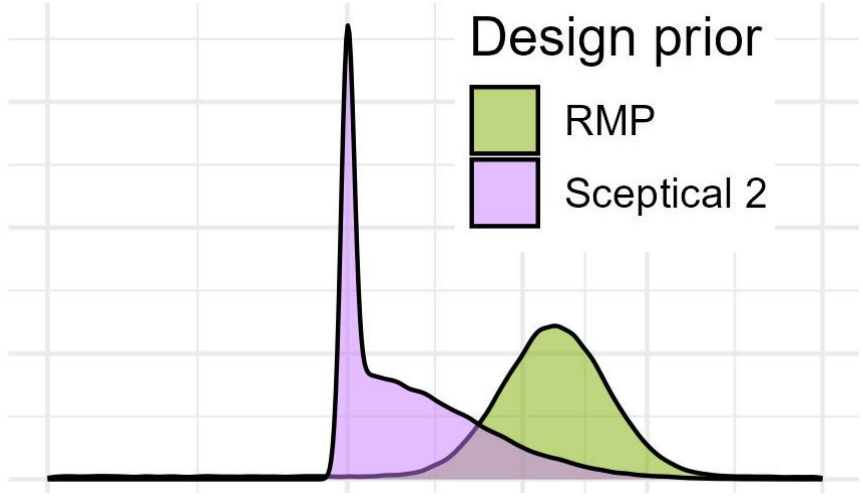
Assessing (in)correctness of decisions: paed example

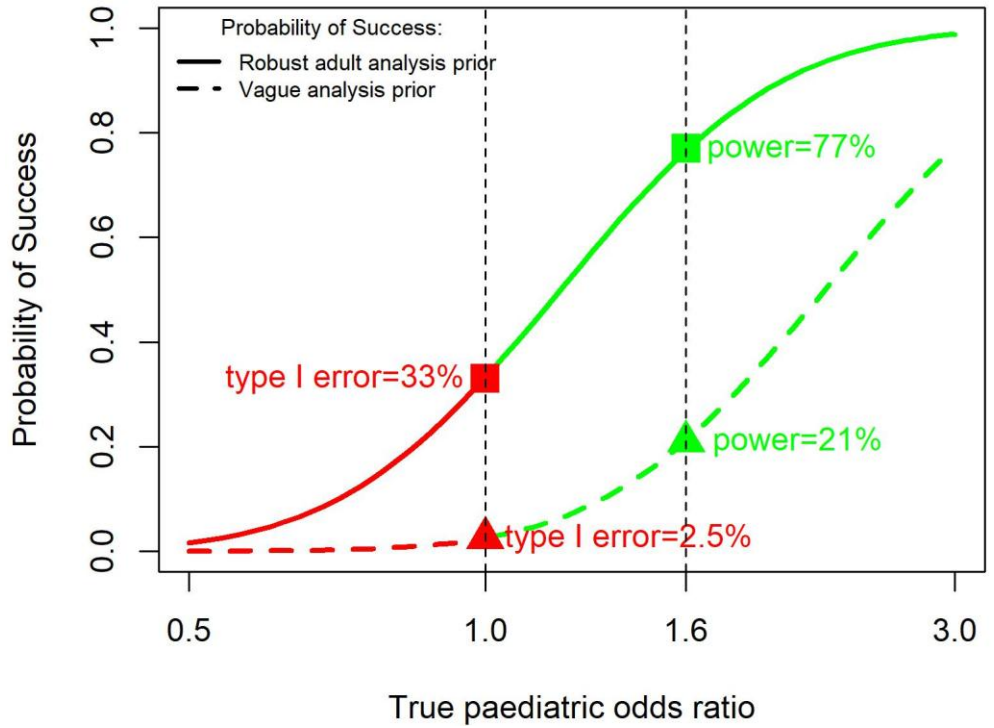




Assessing (in)correctness of decisions: paed example

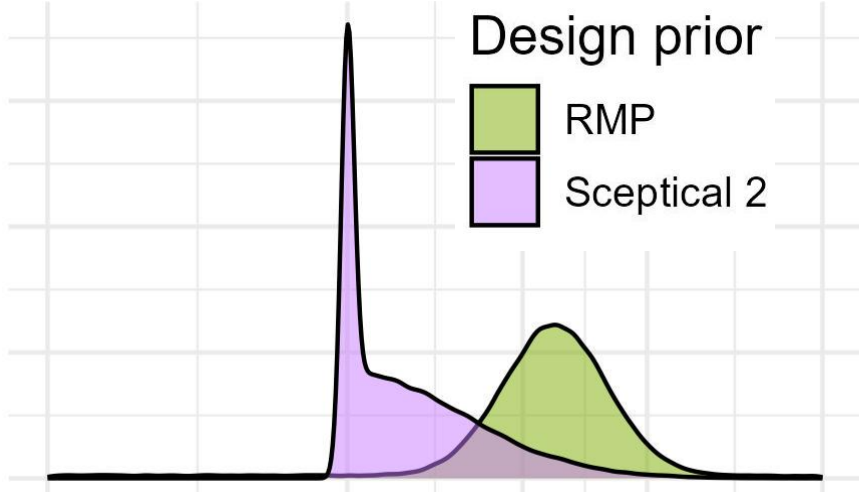
Metric	Design prior	Analysis prior	
		RMP	Vague
Type 1 error	**Point mass at OR=1**	**33%**	**2.5%**
Power	**Point mass at OR=1.6**	**77%**	**21%**
Prior probability of no treatment benefit: Pr(True OR ≤ 1)	**Robust mixture** / **Sceptical 2**	**15%** / **30%**	

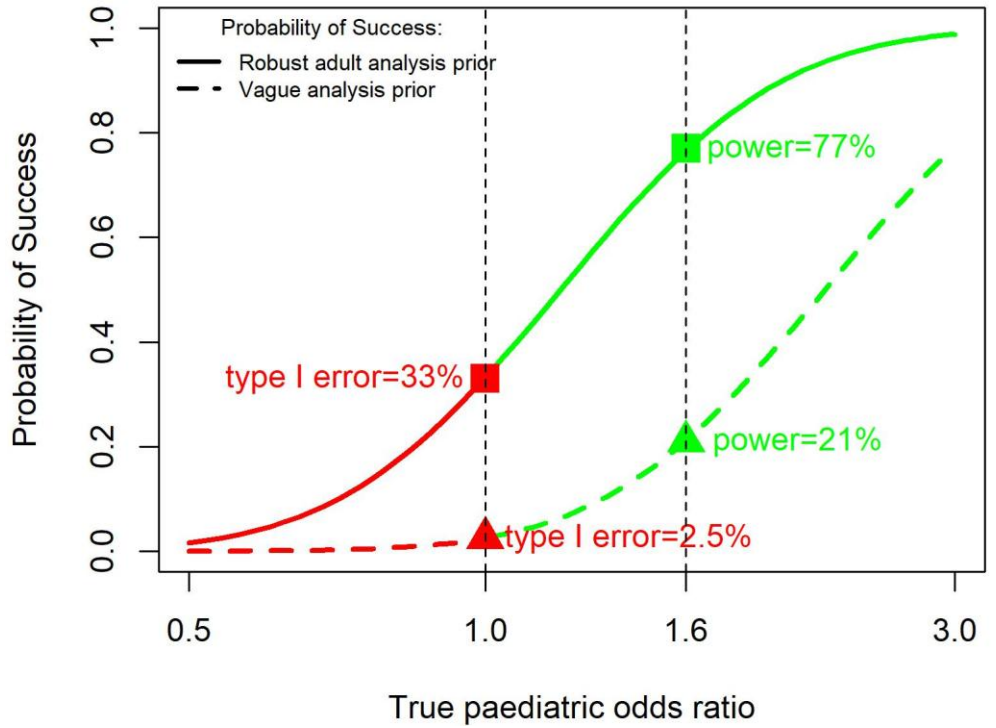




Assessing (in)correctness of decisions: paed example

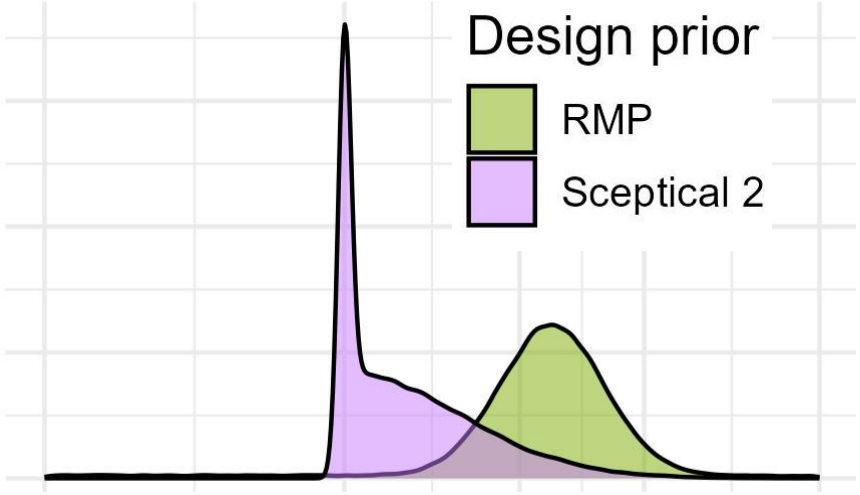
Metric	Design prior	Analysis prior	
		RMP	Vague
Type 1 error	Point mass at OR=1	33%	2.5%
Power	Point mass at OR=1.6	77%	21%
Prior probability of no treatment benefit: Pr(True OR ≤ 1)	Robust mixture	15%	
	Sceptical 2	30%	
Predicted probability of positive results (assurance): Pr(success)	Robust mixture	67%	27%
	Sceptical 2	48%	7%

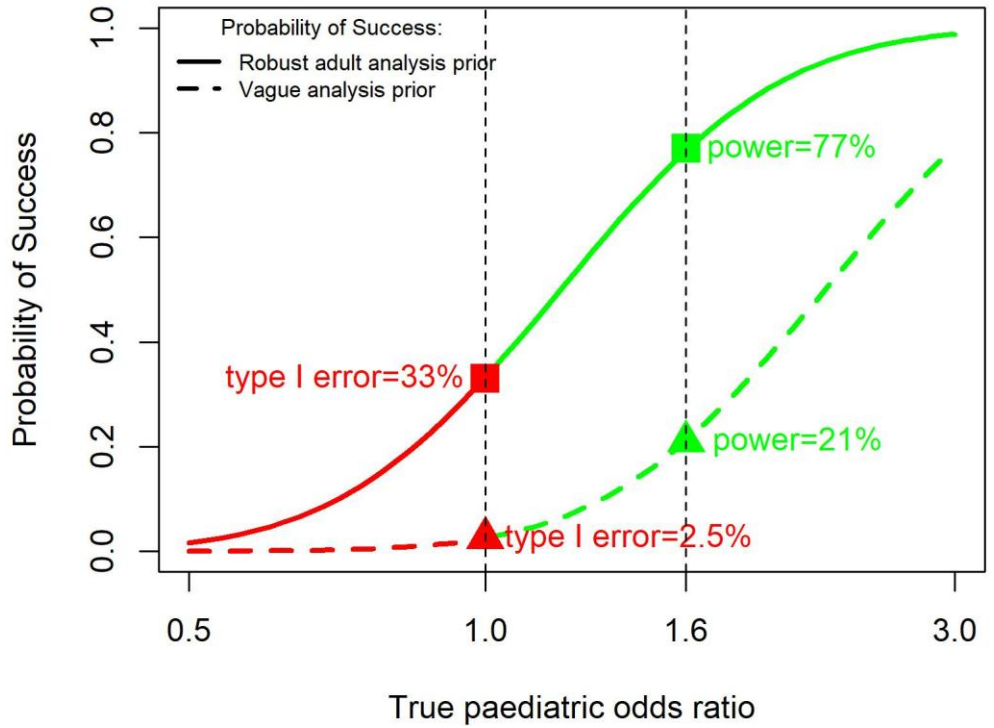




Assessing (in)correctness of decisions: paed example

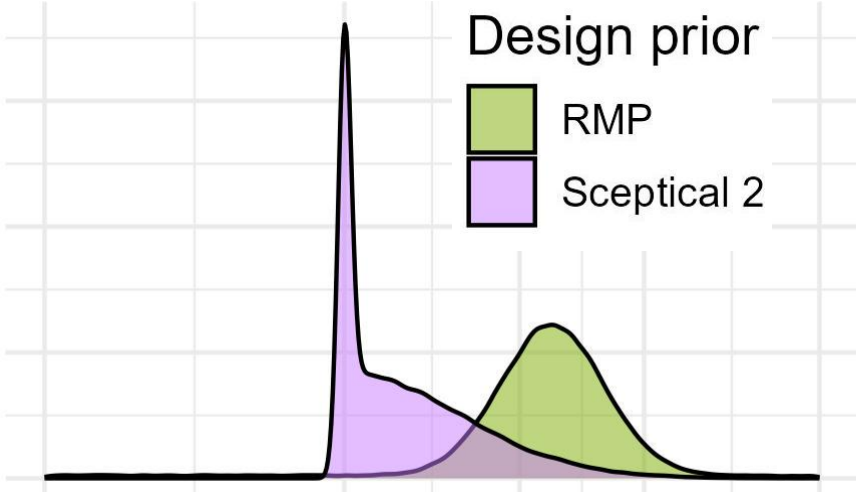
Metric	Design prior	Analysis prior	
		RMP	Vague
Type 1 error	Point mass at OR=1	33%	2.5%
Power	Point mass at OR=1.6	77%	21%
Prior probability of no treatment benefit: Pr(True OR ≤ 1)	**Robust mixture**	15%	
	Sceptical 2	30%	
Predicted probability of positive results (assurance): Pr(success)	Robust mixture	67%	27%
	Sceptical 2	48%	7%
Predicted probability of obtaining a false positive result: Pr(True OR ≤ 1 AND success)	**Robust mixture**	5%	0.4%
	Sceptical 2	10%	0.75%

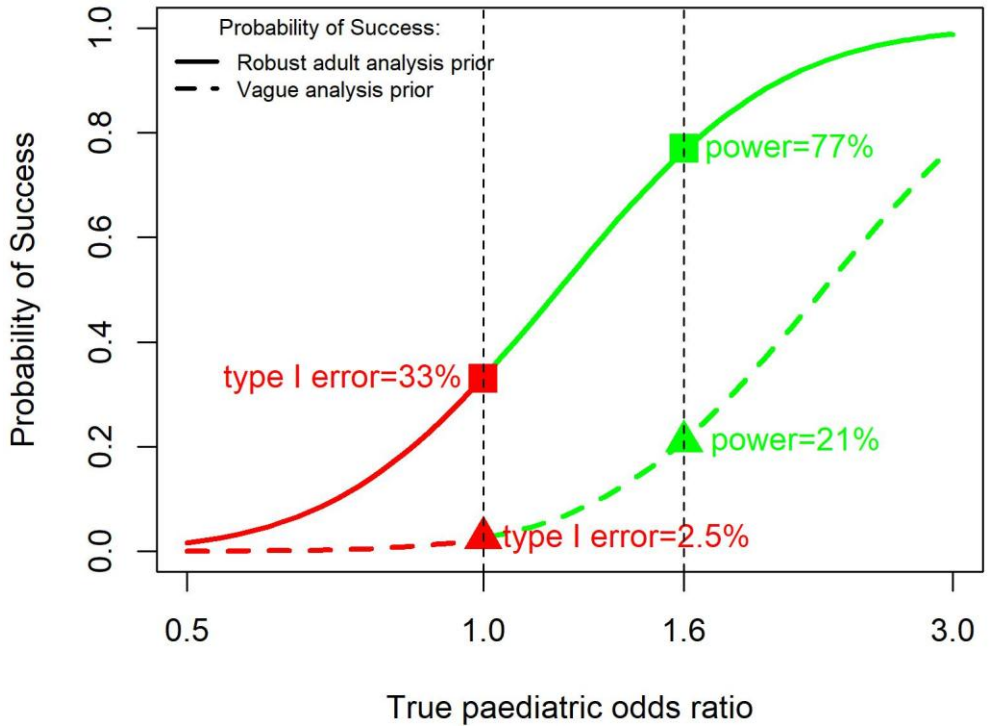




Assessing (in)correctness of decisions: paed example

Metric	Design prior	Analysis prior	
		RMP	Vague
Type 1 error	Point mass at OR=1	33%	2.5%
Power	Point mass at OR=1.6	77%	21%
Prior probability of no treatment benefit: Pr(True OR ≤ 1)	Robust mixture	15%	
	Sceptical 2	30%	
Predicted probability of positive results (assurance): Pr(success)	Robust mixture	67%	27%
	Sceptical 2	48%	7%
Predicted probability of obtaining a false positive result: Pr(True OR ≤ 1 AND success)	Robust mixture	5%	0.4%
	Sceptical 2	10%	0.8%
If positive result is observed, probability it is incorrect: Pr(True OR≤ 1 | success)	Robust mixture	0.6%	<0.1%
	Sceptical 2	29%	28%





Design characteristics: Illustrative results under "what if" data scenarios

How likely are we to end up in different scenarios, and what would the impact be on decisions or inferences?

"What if" value of observed OR in paediatric study	Prior predicted probability of value < observed under different design priors		Proposed BDB design (w =70%)		Frequentist design
	RMP design prior	Sceptical 2 design prior	Posterior weight on evidence informed by adult data	Point estimate (posterior mean) of OR in paediatrics [95% CrI]	Point estimate of OR in paediatrics [95% CI]
0.60	0.13	0.06	0.51	0.96 [0.31, 1.81]	0.60 [0.27, 1.33]
1.00	0.24	0.37	0.89	1.48 [0.75, 1.96]	1.00 [0.48, 2.10]
1.19	0.32	0.53	0.92	1.54 [1.00, 2.00]	1.19 [0.54, 2.64]
1.40	0.42	0.67	0.94	1.58 [1.18, 2.05]	1.40 [0.68, 2.87]
1.60	0.51	0.77	0.94	1.61 [1.23, 2.10]	1.60 [0.79, 3.25]
1.80	0.59	0.84	0.94	1.64 [1.26, 2.17]	1.80 [0.89, 3.64]

Prior-data conflict →

Minimum detectable effect →

Consistent with adult data →

Observed paediatric OR:

***Worse** than in adult data*

***Better** than in adult data*

How much *observed* "drift" is acceptable?

- OR_{obs} = observed OR in paediatric trial (N=100)
- OR_{BDB} = posterior mean OR from BDB analysis of paediatric trial (N=100 + robust mixture prior)

$$DIFF_{BDB} = |OR_{obs} - OR_{BDB}|$$

absolute difference between BDB result and observed result in paeds

Acknowledgement: Matt Psioda

How much *observed* "drift" is acceptable?

- OR_{obs} = observed OR in paediatric trial (N=100)
- OR_{BDB} = posterior mean OR from BDB analysis of paediatric trial (N=100 + robust mixture prior)

$$DIFF_{BDB} = |OR_{obs} - OR_{BDB}|$$

absolute difference between BDB result and observed result in paeds

- OR_{FULL} = observed OR in fully powered trial (N=500) with true OR = OR_{obs}

$$DIFF_{FULL} = |OR_{obs} - OR_{FULL}|$$

differences in OR we might observe due to **sampling variation** if we were to continue collecting data on sufficient children to have a fully powered trial, assuming true OR= OR_{obs} (conservative)

Acknowledgement: Matt Psioda

How much *observed* "drift" is acceptable?

- OR_{obs} = observed OR in paediatric trial (N=100)
- OR_{BDB} = posterior mean OR from BDB analysis of paediatric trial (N=100 + robust mixture prior)

$$DIFF_{BDB} = |OR_{obs} - OR_{BDB}|$$

absolute difference between BDB result and observed result in paeds

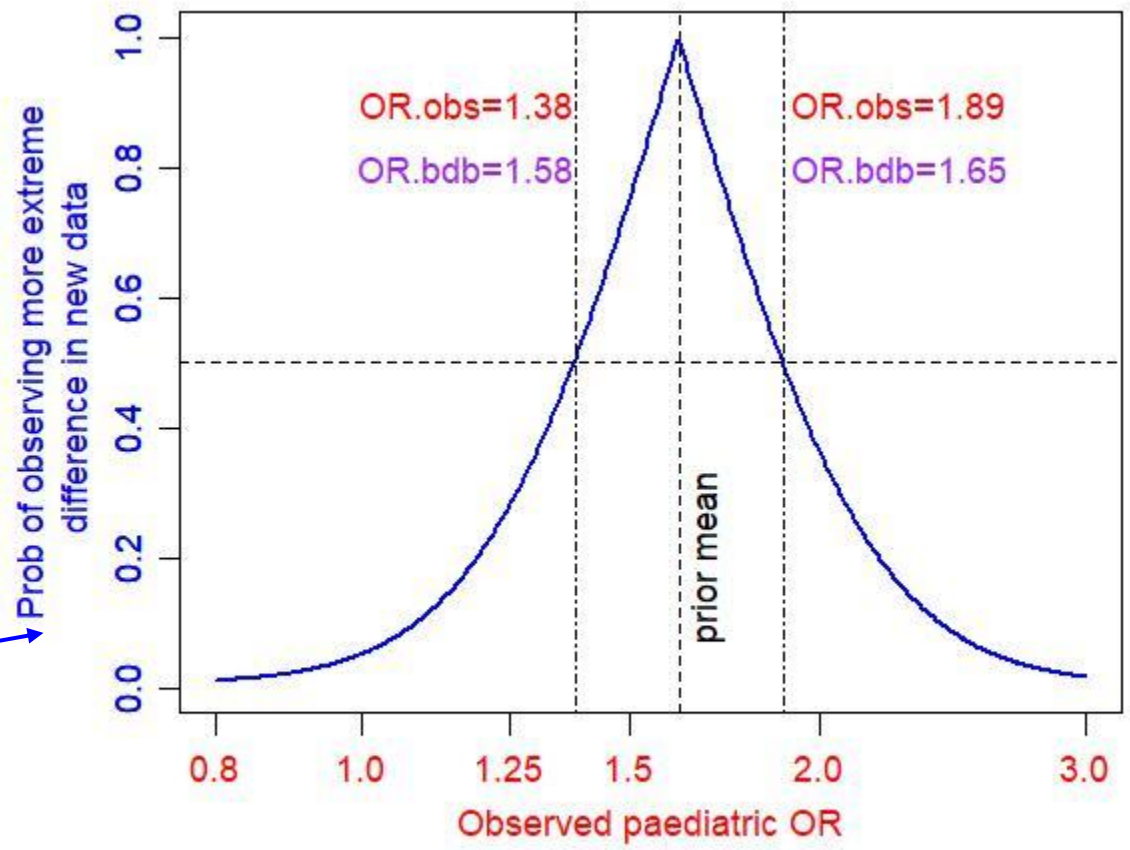
- OR_{FULL} = observed OR in fully powered trial (N=500) with true OR = OR_{obs}

$$DIFF_{FULL} = |OR_{obs} - OR_{FULL}|$$

differences in OR we might observe due to **sampling variation** if we were to continue collecting data on sufficient children to have a fully powered trial, assuming true OR= OR_{obs} (conservative)

- Calculate $\mathbf{Pr}(DIFF_{FULL} > DIFF_{BDB} | OR_{obs})$ for possible values of OR_{obs} in paediatric trial
  - **High probabilities suggest observed drift ($DIFF_{BDB}$) may be reasonable**

$DIFF_{FULL}$ likely to be larger than $DIFF_{BDB}$ for OR_{obs} in this range



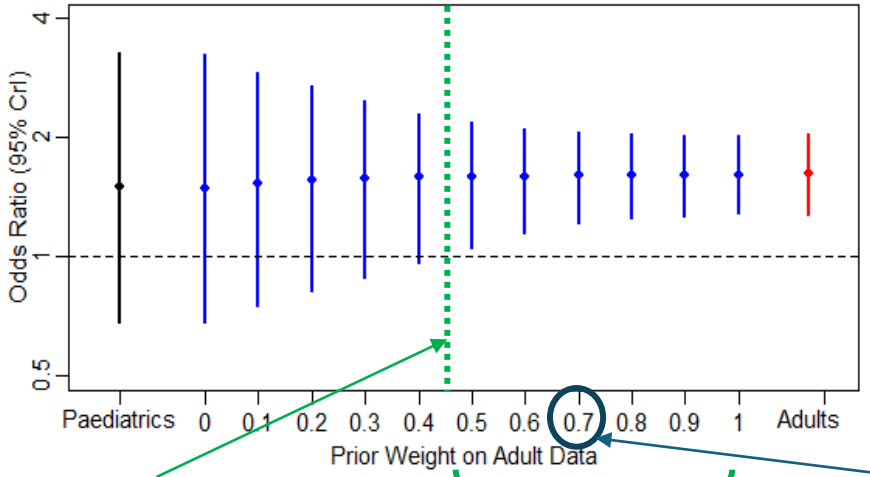
Acknowledgement: Matt Psioda

Reporting

Primary analysis: Posterior summary

Evidence source	Odds Ratio (95% CrI)	
Primary analysis - ***posterior***	1.61	(1.21, 2.07)
Paediatric study only	1.50	(0.68, 3.29)
Adult ***prior*** only	1.62	(1.28, 2.05)
Robust adult ***prior*** only	1.60	(0.02, 52.6)

Sensitivity analysis: Tipping point analysis to varying prior weight

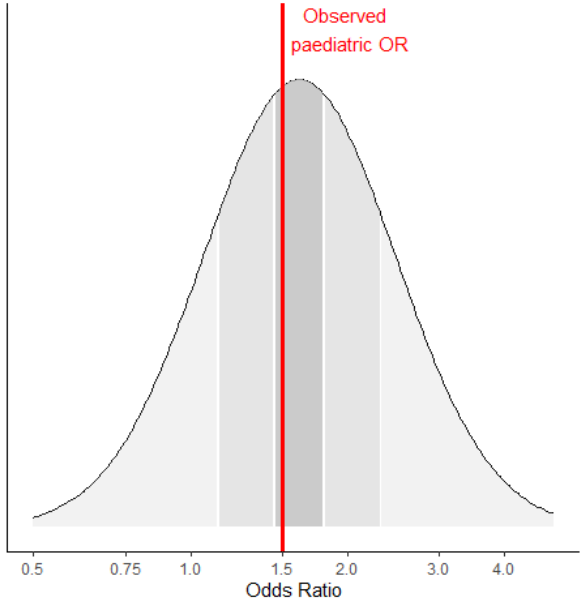


Tipping point

Pre-specified prior weight for primary analysis

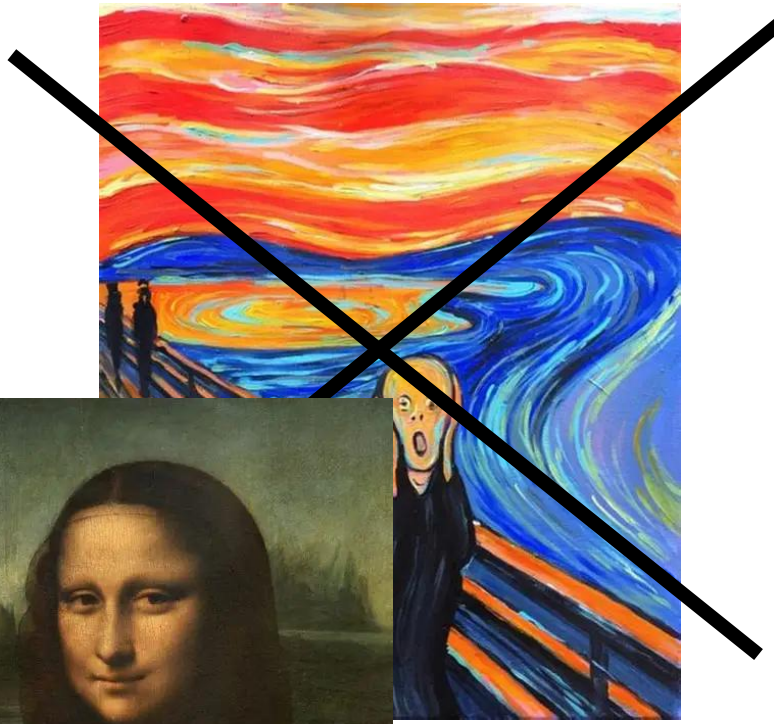
Range of prior weights for which the study conclusion (met primary success rule) is robust

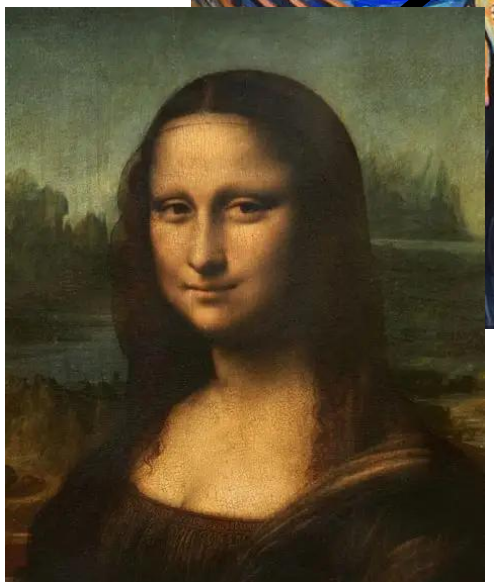
Prior predictive distribution for observed OR



Some recommendations

- Always consider priors on interpretable scale
- Prior or posterior predictions of observables are helpful
- Visualise the prior (static or dynamic)
- Design (sampling) priors are useful to:
  - guide choice of scenarios of interest/concern
  - assess impact of realistic prior-data conflict on
- Priors can be based on data, expert elicitation, or archetypal positions (e.g. sceptical, optimistic)
- Report sensitivity analyses to reasonable alternative priors





Final reflections

Katrina and Florian presentation:

Need for self-standing evidence

- It is **common understanding** that the standard basis for approval is **self-standing evidence** usually generated by two confirmatory RCTs with (strong) Type I error control
  - However, no guideline seems to **specifically** requiring this!

ChatGPT

Self-standing Evidence: This refers to evidence that, **on its own, is sufficient to establish a fact or prove a point**. In other words, even if you removed all other evidence, this piece would still carry enough weight to support the conclusion.

Final reflections

Katrina and Florian presentation:

Need for self-standing evidence

- It is **common understanding** that the standard basis for approval is **self-standing evidence** usually generated by two confirmatory RCTs with (strong) Type I error control
  - However, no guideline seems to **specifically** requiring this!

ChatGPT

Self-standing Evidence: This refers to evidence that, **on its own, is sufficient to establish a fact or prove a point**. In other words, even if you removed all other evidence, this piece would still carry enough weight to support the conclusion.

Compelling Evidence: This term emphasizes how persuasive or convincing the evidence is. It suggests that the **evidence is so strong that it leaves little room for doubt** or counter-argument. **Compelling evidence may combine multiple pieces** or exhibit such clarity and reliability that it forces a decision in favor of one conclusion over another.

Final reflections

Katrina and Florian presentation:

Need for self-standing evidence

- It is **common understanding** that the standard basis for approval is **self-standing evidence** usually generated by two confirmatory RCTs with (strong) Type I error control
  - However, no guideline seems to **specifically** requiring this!

ChatGPT

Self-standing Evidence: This refers to evidence that, **on its own, is sufficient to establish a fact or prove a point**. In other words, even if you removed all other evidence, this piece would still carry enough weight to support the conclusion.

Compelling Evidence: This term emphasizes how persuasive or convincing the evidence is. It suggests that the **evidence is so strong that it leaves little room for doubt** or counter-argument. **Compelling evidence may combine multiple pieces** or exhibit such clarity and reliability that it forces a decision in favor of one conclusion over another.

Is it time to shift the basis for approval to requiring compelling evidence?

References and Useful Resources

- Best, N., Ajimi, M., Neuenschwander, B., Saint-Hilary, G., & Wandel, S. (2024). Beyond the Classical Type I Error: Bayesian Metrics for Bayesian Designs Using Informative Priors. *Statistics in Biopharmaceutical Research*, *17*(2), 183–196. https://doi.org/10.1080/19466315.2024.2342817
- Röver C, Bender R, Dias S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Res Syn Meth*. 2021; 12: 448–474. **https://doi.org/10.1002/jrsm.1475**
- Spiegelhalter D, Myles J, Jones D, Abrams K. Bayesian methods in health technology assessment: a review. Health Technol Assess 2000;4 (38). https://doi.org/10.3310/hta4380
- Spiegelhalter, D.J., Freedman, L.S. and Parmar, M.K.B. (1994), Bayesian Approaches to Randomized Trials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 157: 357-387. https://doi.org/10.2307/2983527
- Wang, Y., Tu, W., Koh, W., Travis, J., Abugov, R., Hamilton, K., Zheng, M., Crackel, R., Bonangelino, P. and Rothmann, M. (2024), Bayesian Hierarchical Models for Subgroup Analysis. Pharmaceutical Statistics, 23: 1065-1083. https://doi.org/10.1002/pst.2424
- Frank Harrell online course: Introduction to Bayes for Evaluating Treatments (https://hbiostat.org/bayes/bet)
- Prior Choice Recommendations · stan-dev/stan Wiki · GitHub (https://github.com/stan-dev/stan/wiki/)
- Applied Modelling in Drug Development (https://opensource.nibr.com/bamdd)