Beyond the classical type I error: Bayesian metrics for Bayesian designs using informative priors

Nicky Best (GSK), Maxine Ajimi (AZ), Beat Neuenschwander (Novartis), Gaëlle Saint-Hilary (Saryga), Simon Wandel (Novartis)
on behalf of EFSPI/PSI Historical Data SIG

- Regulators increasingly open to use of **external data** in particular scenarios, e.g.
  - ➢FDA's Complex Innovative Designs (CID) initiative includes several projects using external data in pivotal studies
  - ➢ICH E11A Pediatric Extrapolation Draft Guideline to use external/reference data
  - ➢Several examples of drug approvals granted based on non-randomized studies using external controls[6]
- **Bayesian methods** offer an appealing approach to incorporate external evidence via the use of informative prior distributions
  - ➢Common practice to evaluate Bayesian designs: using simulations to understand frequentist operating characteristics, including the classical type I error
  - ➢Classical type I error cannot be strictly controlled[10,11] in a Bayesian design with informative priors, and may be above, below or equal to its nominal level
  - ➢The FDA[8] recommends that for Bayesian designs using informative priors, *appropriate alternative trial characteristics should be considered*.
- **We present several alternative Bayesian (i.e. fully probabilistic) metrics to evaluate the risk of a Bayesian trial producing false positive conclusions**

Notation: θ_t , θ_c = true treatment effects on active, control arms; $\delta = \theta_t - \theta_c$ = treatment contrast; Study Success = $I\{(\Pr(\delta > \delta_{null} \mid y) > 1-\alpha\}$

- We define the following metric $\mathbf{M_1} = \int Pr(Study\ Success \mid \delta)\ p(\delta) d\delta$ **(1)**
  where $p(\delta)$ is a suitable probability distribution describing values of the true treatment contrast
- Several common metrics are special cases of M_1:
  - ➢**Classical type 1 error**: $p(\delta)$ = Dirac measure with point mass at δ_{null} ⇒ $M_1 = Pr(Study\ Success \mid \delta = \delta_{null})$
  - ➢**Classical power**: $p(\delta)$ = Dirac measure with point mass at δ_{alt} ⇒ $M_1 = Pr(Study\ Success \mid \delta = \delta_{alt})$
  - ➢**Assurance** [15] (average power): $p(\delta)$ = design prior reflecting our uncertainty around hypothesized treatment effect

Analysis Priors and Design Priors

- Analysis Prior – used in analysis of the current trial and represents best reflection of the evidence and the corresponding uncertainty
- Design Prior – used for design evaluation to calibrate Bayesian designs under different assumptions about the true parameter value(s)

Metrics when borrowing information on controls

- Under the null, $\theta_t = \theta_c + \delta_{null}$, leading to the following version of metric (1):
  $\mathbf{M_2} = \int Pr(Study\ Success \mid \theta_c, \theta_t = \theta_c + \delta_{null})\ p(\theta_c)\, d\theta_c$ **(2)**
  - ➢Classical type 1 error is a "pointwise" rate, depending on true value of θ_c
  - ➢The Bayesian metric (2) is the average (unconditional, or marginal) of this classical type 1 error *wrt* the design prior $p(\theta_c)$
- For data generated under a normal likelihood, the average type I error defined in (2) is **strictly controlled at level α** if the **analysis prior is also used as the design prior $p(\theta_c)$**; asymptotically controlled at level α for any likelihood (proof: appendix of ArXiv pre-print)

Metrics when borrowing information on the treatment contrast

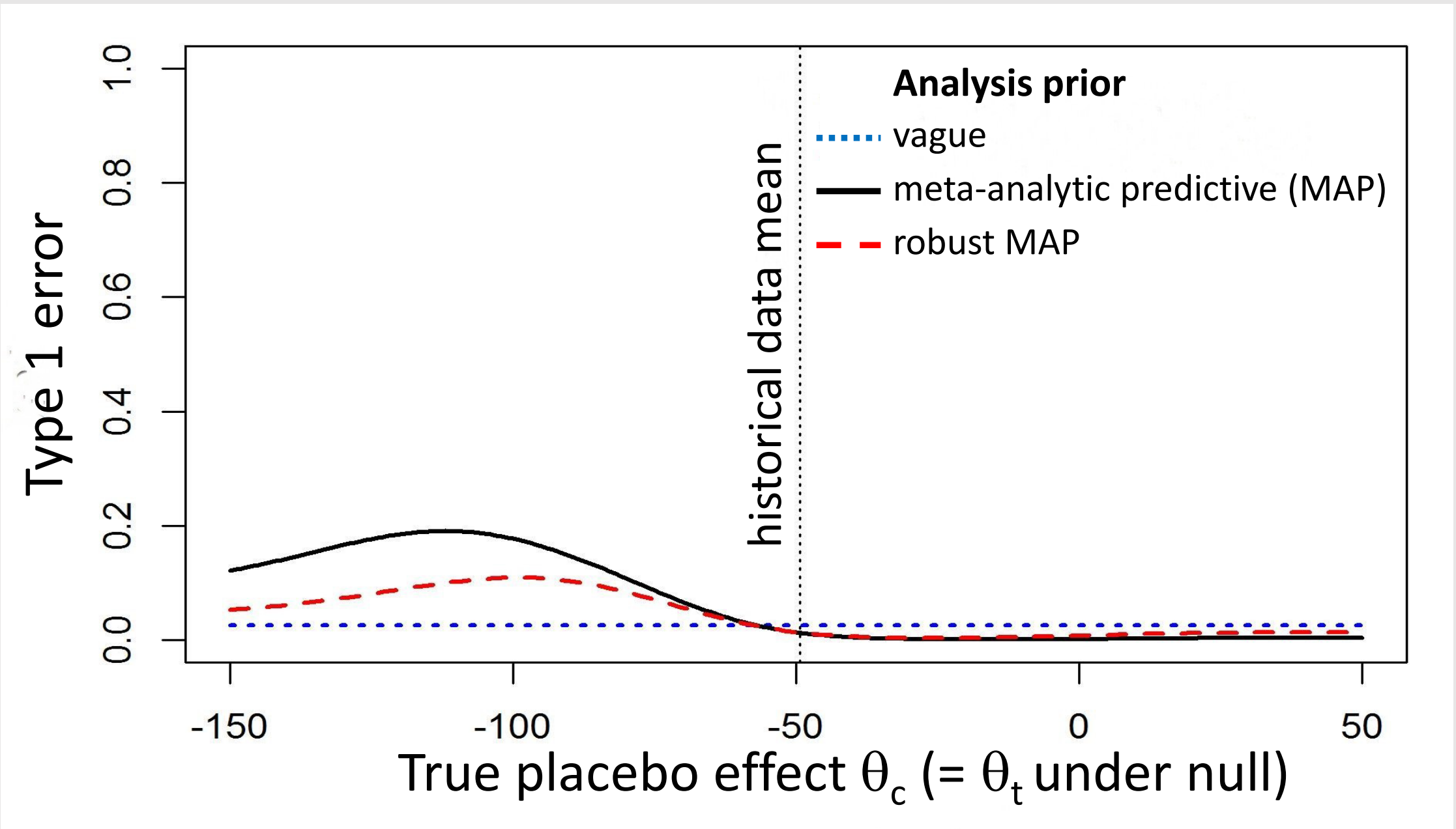
- For a Bayesian design with prior information on the treatment contrast, we need a design prior, $p(\delta)$, that is consistent with the assumed null treatment effect
- Usually, analysis prior supports a positive effect of the investigational treatment ⇒ **prior in conflict with null** treatment effect ⇒ **inflated classical type 1 error** [10,11]
- We propose an **alternative metric M₃** as follows

$$\mathbf{M_3} = \underbrace{\int \Pr(Study\ Success|\delta) \frac{p(\delta) I\{\delta \le \delta_{null}\}}{\Pr(\delta \le \delta_{null})} d\delta}_{Average\ type\ 1\ error\ under\ null\ (truncated)\ design\ prior} \times \underbrace{\Pr(\delta \le \delta_{null})}_{Prob\ treatment\ effect\ is\ null\ or\ harmful}$$

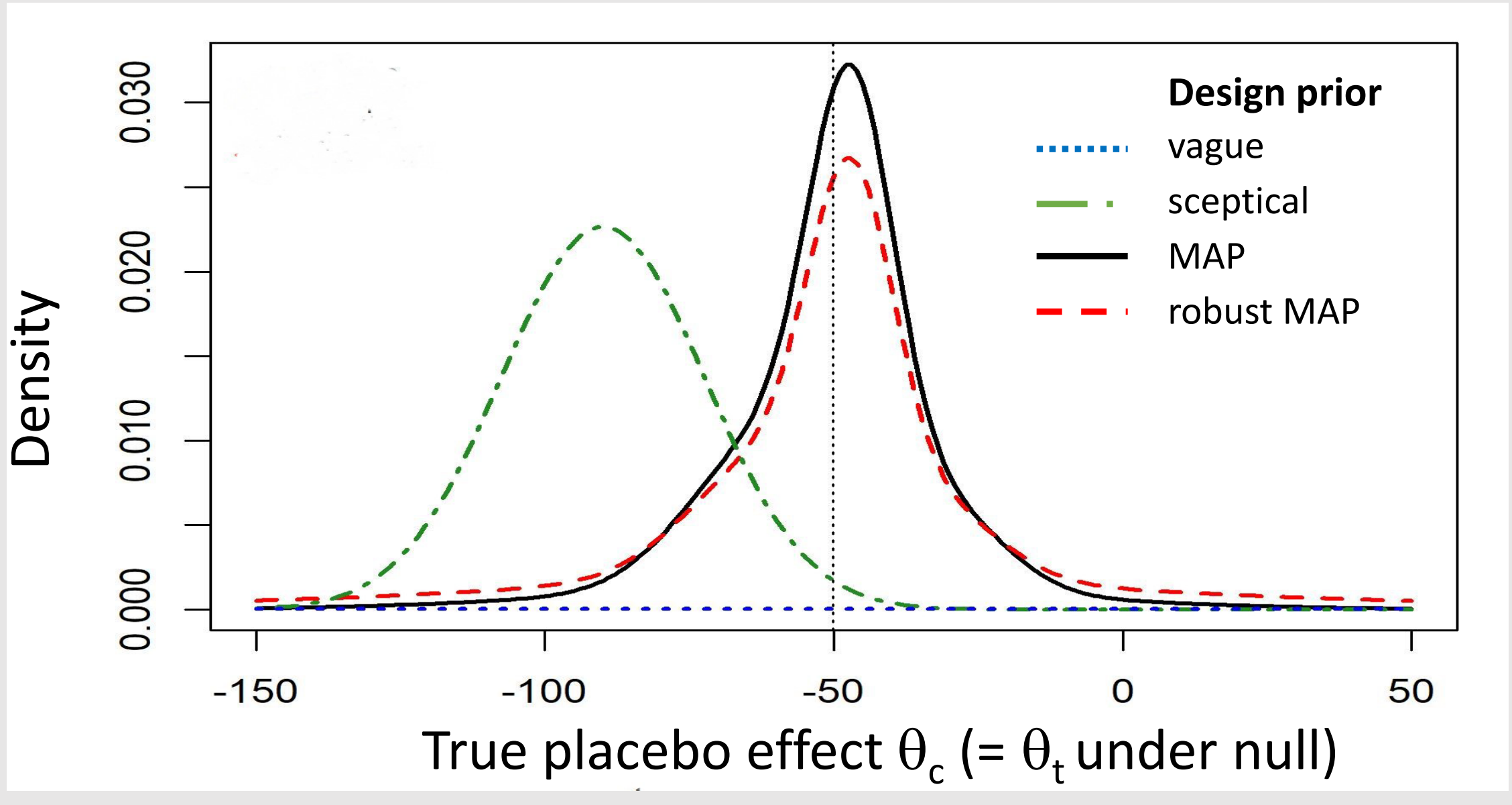
$$= \underbrace{\int_{\delta \le \delta_{null}} \Pr(Study\ Success|\delta)\ p(\delta) d\delta}_{Prob\ false\ positive\ result\ =\ Joint\ prob\ that\ trial\ is\ success\ and\ true\ treatment\ effect\ is\ null\ or\ harmful}$$

Case Study 1: Borrowing historical placebo data (example in Crohn's Disease[24,25])

Classical type 1 error for different placebo Analysis priors



Placebo Design priors used for evaluating Bayesian average type 1 error

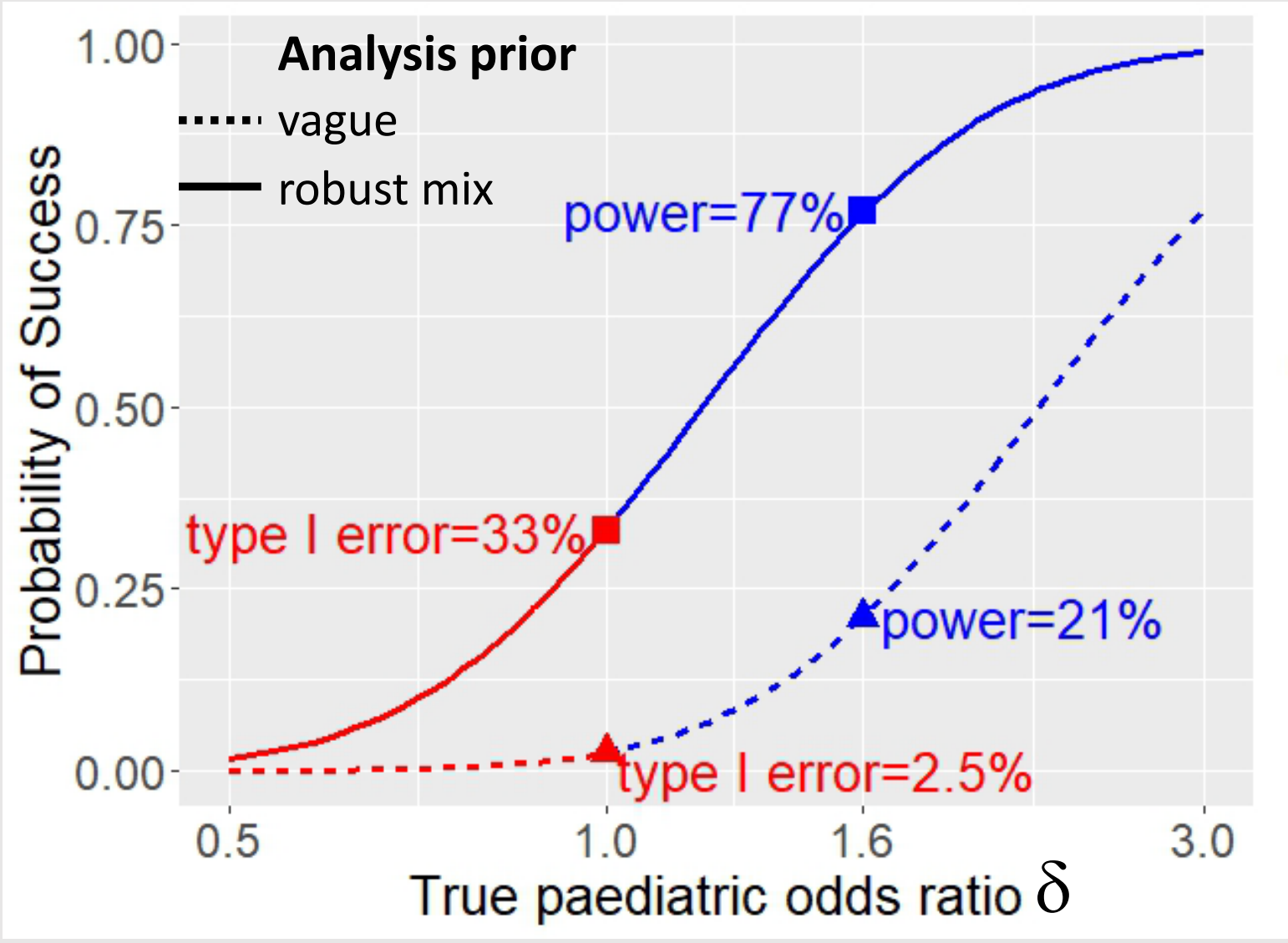


Bayesian average («unconditional») type I error (Metric M_2)

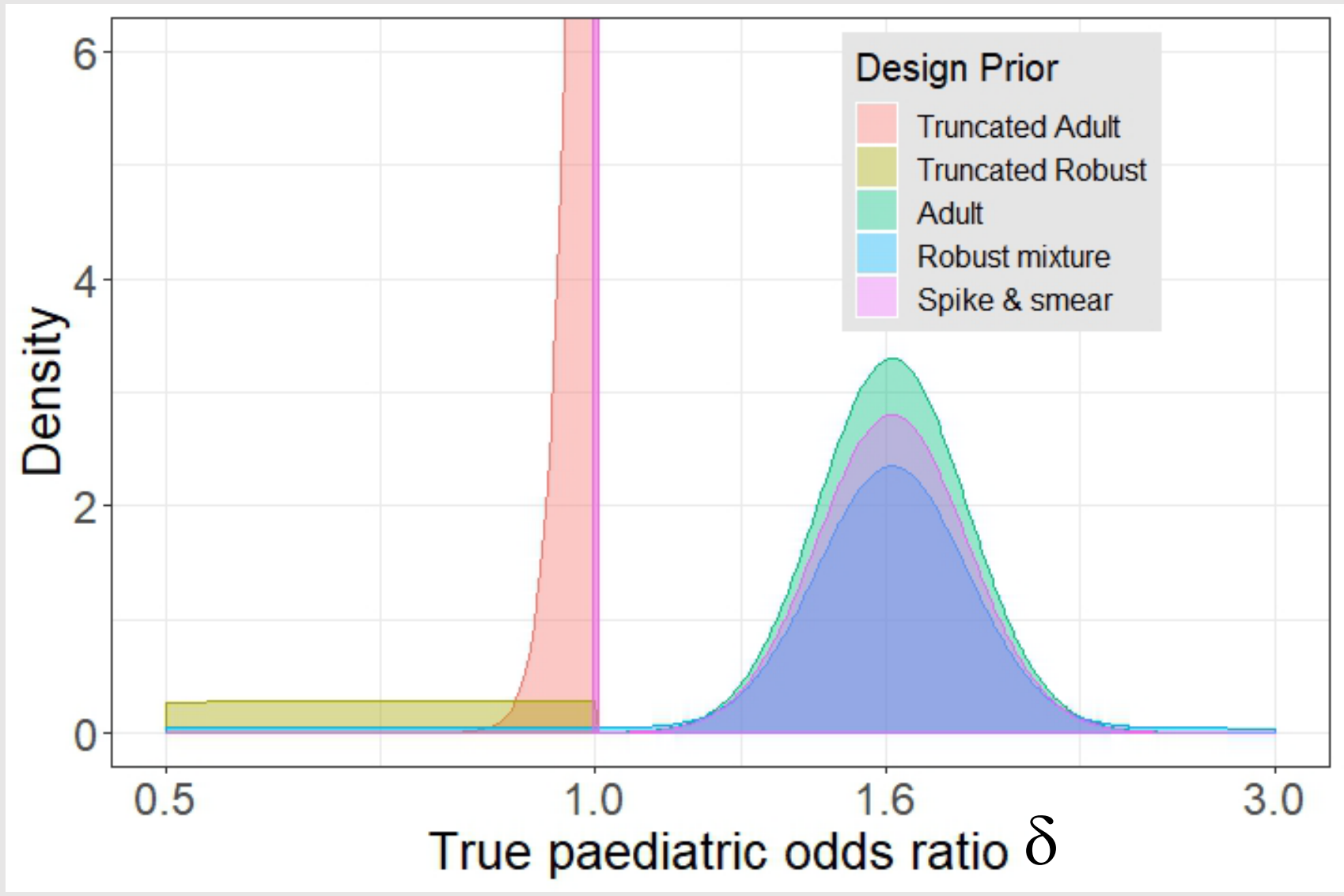
Placebo Analysis prior	Design prior for placebo effect			
	Vague	Sceptical	MAP	Robust MAP
Vague	**2.5%**	2.5%	2.5%	2.5%
MAP	48.5%	13.4%	**2.5%**	3.2%
Robust MAP	45.6%	8.8%	2.2%	**2.5%**

Case Study 2: Borrowing historical data on treatment contrast (Paediatric example[28])

Probability of success curves for two analysis priors



Design priors for treatment contrast used to evaluate Bayesian metrics



Bayesian metrics for different analysis and design priors

Metric	*Analysis prior for treatment difference*	Design prior for treatment difference		
		Truncated adult	Truncated robust mix	Point mass at 0*
Average type 1 error (metirc M_1)	***Vague***	2.1%	0.1%	2.5%
	Robust mixture	30.8%	2.5%	33.2%
		Adult	Robust mix	Spike & smear
Design prior prob of no benefit	-	0.004%	15.003%	15%
Prob of false +ve (metric M_3)	***Vague***	<0.001%	0.015%	0.375%
	Robust mixture	0.001%	0.375%	4.982%

*gives special case of M_1 = classical type 1 error

- **Strict control of the classical (frequentist) type 1 error is not possible when leveraging prior information in a Bayesian clinical trial design**
- We propose that **average type I error** (which is analogous to assurance under the null hypothesis) is also a relevant metric to inform decision-makers
- In designs where information is borrowed on the treatment contrast, we also recommend calculation of the **probability of actually declaring a false positive result**
- The strong focus on classical (frequentist) type 1 error control for pivotal studies has emphasized consideration of the bias question only. We argue that a more holistic viewpoint is required to judge designs that, by construction, aim at optimizing the **bias-variance trade-off**.

References and further details available in ArXiv pre-print: http://arxiv.org/abs/2309.02141