ROC N^{O} 2: EMA/2020/46/TDA, LOT 3
STUDY OF OPERATING CHARACTERISTICS OF BAYESIAN METHODS FOR EXTRAPOLATION OF TREATMENT EFFECTS

SIMULATION STUDY REPORT (DELIVERABLE 4)

Quinten Health
8 rue Vernier, Paris, France

February 5, 2025

Contents

Notations

$\Gamma(\alpha, \beta)$ The gamma distribution with parameters α and β.

$\Gamma^{-1}(\alpha, \beta)$ The inverse gamma distribution with parameters α and β.

$\mathbb{E}(Z)$ Expectation of the random variable Z.

$\mathbb{V}(Z)$ Variance of the random variable Z.

$\mathbf{D}$ Available data, made of input-observation pairs $(\boldsymbol{x}, y)$.

$\mathcal{B}(n, p)$ The binomial distribution with parameters n and p.

$\mathcal{B}(p)$ The Bernoulli distribution with parameter p.

$\mathcal{HN}(\sigma)$ The half normal distribution with scale parameter σ.

$\mathcal{N}(\mu, \sigma^2)$ The normal distribution with mean μ and variance σ^2.

$\mathcal{U}(a, b)$ Uniform distribution over $[a, b]$.

$\pi(\cdot)$ Prior distribution

$\theta_i^{(\text{true})}$ True treatment effect in study i, used to generate data in the simulation.

θ_S Source study treatment effect.

θ_T Target study treatment effect

N_S Source study sample size

N_T Target study sample size

$p(Z \mid W)$ Probability density function or probability mass function of the random variable Z conditioned on W.

$p(z)$ Probability density function or probability mass function of the random variable Z evaluated in z.

$Pr(A)$ Probability of event A.

y An observation of random variable Y.

For simplicity, we abuse notation by using Z to denote both random quantities and the arguments of their probability density functions, unless necessary for clarity, in which case capital letters indicate random variables, whereas lower case letters indicate the observations from these random variables. Moreover, unless necessary for understanding, we do not use distinct notations for probability density functions associated with different random variables.

Acronyms

CPP Conditional Power Prior.

ELIR Expected local-information-ratio.

EMA European Medicines Agency.

ESS Effective Sample Size.

FDA US Food and Drugs Administration.

NPP Normalized Power Prior.

PDCCPP Prior-data conflict calibrated power priors.

PP Power Prior.

PPP Prior-predictive p-value.

PTtP Predictive Test-then-Pool.

TIE Type I Error.

UMP Uniformly Most Powerful.

1 Introduction

1.1 Context

Though the use of Bayesian approaches has been considered in drug development for decades through the ICH E9 guideline (CPMP/ICH/363/96), the recent US 21st Century Cures Act highlighted the use of complex clinical trial designs, such as Bayesian designs (US House of Representatives 2015). This served as a catalyst for discussions regarding the opportunities and challenges of using external information in the design and analysis of clinical trials. Regulatory agencies are increasingly willing to consider the use of methods for borrowing information where appropriate. The European Medicines Agency (EMA) has issued over the last years a concept paper (EMA/129698/2012), a draft reflection paper on the extrapolation of efficacy and safety in medicine development (EMA/129698/2012), and a final reflection paper on the use of extrapolation in the development of medicines in pediatrics (EMA/189724/2018). The latter established a framework for generating evidence for regulatory assessment of marketing authorization applications in target populations, particularly in pediatrics. This approach involves the use of existing information from one or more source populations, such as adults, through quantitative methods such as Bayesian methods among others. Similarly, the US Food and Drug Administration (FDA) has issued guidance on the use of Bayesian methods, in particular through the guidance for the use of Bayesian statistics in medical device clinical trials (US Food and Drug Administration 2010). While the FDA's guidance documents on adaptive designs for medical device clinical studies and clinical trials of drugs and biologics (US Food and Drug Administration 2016; US Food and Drug Administration 2019) are less Bayesian-oriented, they also do consider (adaptive) Bayesian statistical methodologies. ICH E11 (R1) states that "A fundamental principle in pediatric drug development requires that children should not be enrolled in a clinical study unless necessary to achieve an important pediatric public health need." The ICH published a guideline on the clinical investigation of medicinal products in the pediatric population (EMA/CPMP/ICH/2711/1999) which also suggests the incorporation of external information in the design and analysis of clinical trials. This guideline was followed by a guideline on pediatric extrapolation (EMA/CHMP/ICH/205218/2022) which provides recommendations, in particular, for the use of Bayesian statistics in trial design and analysis in the pediatric context. Overall, these guidelines emphasize the need for harmonization of methodologies for extrapolation in drug development.

It is not clear from published literature what are the underlying operating characteristics of statistical methods that borrow treatment effects in the design and analysis of clinical trials. More specifically, it remains unclear how these operating characteristics depend on the setting, in particular, the drift between source and target study treatment effect, defined as the difference between the true treatment effect in the target study and the estimate of the treatment effect in the source study (Viele et al. 2018; Lim et al. 2020; Best et al. 2023). It is also unclear how operating characteristics depend on extrapolation methods' parameters, and how these methods compare to each other.

This stems from the fact that these borrowing methods have been proposed over the last two decades, and therefore no unified framework exists to evaluate and compare them. Moreover, existing simulation studies tend to limit reporting to the main results, which may result in some lack of clarity over the simulation study design, or other details (e.g. no details on Monte Carlo uncertainty). This is a general issue noted previously with simulation studies in statistics (Morris et al. 2019).

This report presents the result of a large-scale simulation study aiming at addressing these caveats.

1.2 Objectives

As explained in the EMA tender technical specifications, the simulation study targeted the following objectives :

- Perform a large-scale simulation study to better understand the relationship between parameters for any given model. Underlying simulation parameters that needs to be varied include but are not limited to: the amount of information that is to be borrowed from the source population; the sample size of the clinical trial in the target population; the magnitude of the treatment effect; the link function between observed outcome and statistical model for the treatment effect, as well as the varying parameters needed to specify the models. Key outputs of interest are:
  - the unconditional type 1 error rates, (which are expected to vary depending on the choice of parameters);
  - the statistical power in at least two scenarios: (i) in the event that the true effect in the target population is the same as the source population i.e. the data can be extrapolated, and (ii) when the effect in the target population is half as big as the source population ("moderate" effect). In

this latter case, where the treatment effect in the target population is not the same as the source population but is non-zero, bias should also be measured. It should be investigated whether results can be meaningfully summarised in terms of the drift away from the true value of the prior, conditional on whether a simulation study is feasible;
  - plots of parameter choices versus unconditional type 1 error, similar to Figure 9 of Viele et al. (2014) are important outputs;
  - when the amount of borrowing is not fixed but instead model-dependent, the effective sample size of the final model should also be calculated.
- Finally, provide a comparison between models, in particular, to answer the question as to which models offer the most power at comparable type 1 error control. A comparison should also be made against the power of a frequentist approach that uses different measures of unacceptable Type I Error rate rather than the standard 5%. For example, if a particular Bayesian model under specific parameterizations has a maximum unconditional type 1 error rate of 12%, then the power of this model should be compared with the power of a frequentist approach using 12% as the cut-off to declare efficacy. Using 5% as the cut-off would not allow an assessment as to whether the Bayesian method had any additional power beyond that gained from simply using a method that increases the traditional frequentist l type 1 error. Other operating characteristics of the Bayesian approach have been proposed, for example, the average type 1 error and the maximum type 1 error in a pre-specified range should also be calculated. The purpose of this final exercise is to ascertain whether the improved power is simply bought at the expense of type 1 error Control, and if so, which, if any of the models outperform frequentist approaches with an explicitly greater type 1 error. The parameters and distributions chosen for the simulation study, specifically for the magnitude of the adult treatment effect, should reflect the types of data seen in such proposals.

Following the above objectives, key outputs of interest are:

- Type I error rate/Power/MSE/Bias/Precision as a function of drift, defined as the difference between the true treatment effect in the new study and the estimate of the source treatment effect, $\delta = \theta_T^{(\text{true})} - \hat{\theta}_S$.
- Type I error rate/Power/MSE/Bias/Precision as a function of the relevant model parameters, in the congruent (i.e. no or small drift) and non-congruent (i.e. significant drift) scenarios,
- Prior Effective Sample Size (ESS). See Section 8.2 for details.
- For dynamic borrowing priors, the posterior value (or a relevant summary from the posterior distribution, e.g. the posterior mean/median and 95% credible interval) of the parameter(s) that governs how much weight is given to the source data as a function of the true treatment effect.

These outputs were analyzed and reported so that they could be used to inform evaluation and assess specific cases.

2 Simulation settings

The key steps considered in the design of the simulation study are summarized in Figure 1. The simulation study was designed, coded, and analyzed following good practices detailed in Morris et al. (2019).

To mimic the situation of paediatric extrapolation, where information on the treatment effect in adults may be used to inform trials in paediatrics, we focus on scenarios where non-concurrent data sources could be used to inform the design and analysis of a target clinical trial. Importantly, no covariates were included.

2.1 Scenarios considered to study the selected methods

2.1.1 Endpoints and summary measures

We considered four types of endpoints (i.e., the variable collected at an individual level): continuous, binary, time-to-event, and recurrent event endpoints. These endpoints were the same in the source and target studies.

For simplicity, and because this is the most standard setting, we always considered that the source and target data likelihoods belong to the same family of distributions.

Continuous endpoints and associated summary measures were considered normally distributed. For binary endpoints, we included one case study (Aprepitant case study, see 10.2) in which the summary measure was the difference in response rate between the two arms, modeled using binomial likelihoods for

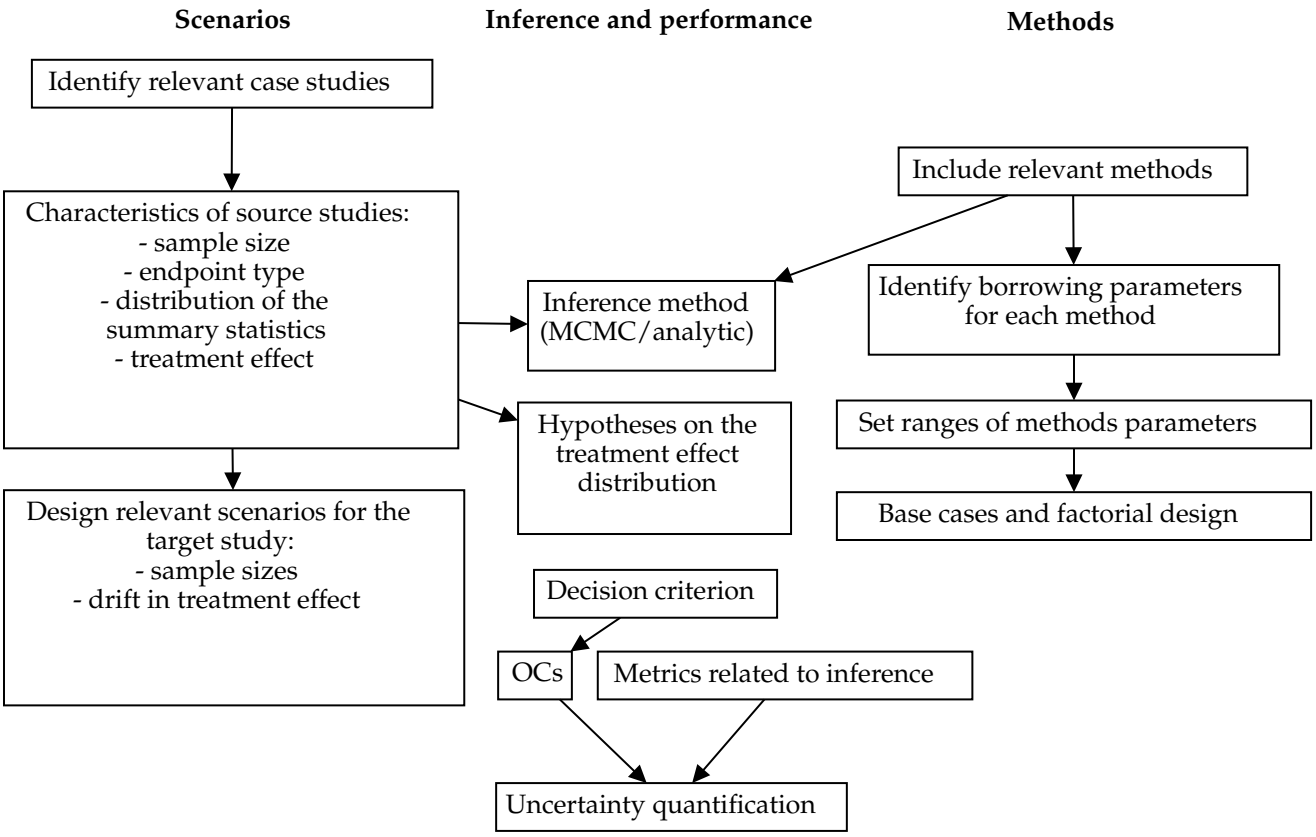

Figure 1: Steps taken in the design of the simulation study

each arm's data, and one case in which the summary measure was the log odds ratio, modeled on the log scale using a normal distribution (Belimumab case study, see 10.2). Similarly, we used normal likelihoods in the cases of time-to-event and recurrent event endpoints, where the associated summary measures were the log hazard ratio and the log event rate ratio respectively.

2.1.2 Procedure for the selection of case studies

To ensure the scenario considered in the simulation study are realistic, we relied on existing studies in adults and paediatrics to inspire the design of the scenarios. In particular, we used historical adults aggregate data instead of simulated data when building priors for the target study treatment effect.

We searched for studies where the efficacy of treatment was assessed in similar settings in adults and in paediatrics. We targeted studies that satisfied the following criteria listed in descending order of priority:

- Studies where extrapolation from adults was proposed or submitted to a regulatory health authority.
- Examples where the study in paediatrics did not demonstrate a positive treatment effect, yet a post-hoc analysis was published using Bayesian borrowing from adults.
- Examples where the study in paediatrics did not demonstrate a positive treatment effect, and at least one related study exists in adults, which was not used in the analysis of the paediatric data.
- Examples where the study in paediatrics demonstrated a positive treatment effect and at least one related study exists in adults, which were not used in the analysis of the paediatric data. Even if the paediatric study was sufficiently powered in this case, we can still investigate the scenario in which sample sizes would be smaller in the paediatric study in simulations.

To identify such studies, we screened papers included as case studies in the Bayesian borrowing literature, the paediatrics.eu - EPARs database, as well as Google Scholar and Pubmed using keywords such as "Bayesian borrowing paediatrics". This resulted in a first selection, which was narrowed down to cover a variety of endpoints, summary measures, disease areas, and sample sizes. This selection is summarized in Table 1, and additional context on these case studies is given in Section 10.

Disease	Lower limb spasticity	Type-2 diabetes	Postoperative nausea and vomiting	Systemic Lupus Erythematosus (SLE)	Multiple Sclerosis	Severe Eosinophilic Asthma
Drug	**Botox vs placebo**	**Dapagliflozin vs placebo (+ Metformin)**	**Aprepitant vs ondansetron**	**Belimumab vs placebo**	**Teriflunomide vs placebo**	**Mepolizumab vs placebo**
Endpoint	Disease severity score	Glycated hemoglobin HbA1c	Absence of vomiting and rescue therapy 0-24h after surgery	SLE Responder Index	Time to first relapse	Number of clinically significant exacerbations.
Endpoint type	**Continuous**	**Continuous**	**Binary**	**Binary**	**Time to event**	**Recurrent event**
Summary measure	Difference in mean scores between the two arms	Difference between the two arms in change in HbA(1c) scores from baseline to week 24/26	Difference in response rates between the two arms	Log odds ratio for active treatment compared to placebo	Log hazard ratio for active treatment compared to placebo	Log exacerbation rate ratio for active treatment compared to placebo
Treatment effect distribution	Normal	Normal	Integral of the product of binomials	Normal (approximation for the log OR)	Normal (approximation for the log HR)	Normal (approximation for the log rate ratio)
N_T**: ctrl/trt/tot**	130/126/256	76/81/157	52/55/107	39/53/92	57/109/166	NA/NA/25
N_S**: ctrl/trt/tot**	235/233/468	134/133/267	293/280/573	562/563/1125	752/731/1483	NA/NA/551
$\mathbf{y_T}$**/Data**	0.10 (0.10)	1.03 (95% CI, 0.49-1.57) (at week 26)	Treatment : 48/55, control: 42/52	Treatment : 28 /53 Placebo: 17/39	HR : 0.66 (95% CI, 0.39-1.11)	Rate ratio : 0.67 (0.17, 2.68)
$\mathbf{y_S}$**/Data**	0.20 (0.10)	0.36 (0.102) (at week 24)	Treatment : 184/293 Control : 154/280	Treatment: 285/563 Placebo: 218/562	HR : 0.68 (95% CI, 0.58-0.79)	Rate ratio : 0.50 (0.39, 0.64)
Reference	Wang et al. (2022)	Shehadeh et al. (2023), Bailey et al. (2010)	Jin et al. (2021), Salman et al. (2019), Diemunsch et al. (2007)	Best et al. (2023), Psioda and Xue (2020), Brunner et al. (2020), Brunner et al. (2021)	Bovis et al. (2022)	Best et al. (2021), MENSA trial (Ortega et al. 2014), Keene et al. (2020)

Table 1: Table summarizing the case studies used to inspire the simulation study design

2.1.3 Sample sizes

For a given case study, the source data sample size N_S was kept fixed across scenarios, but the target data sample size N_T varied in a range of values where the maximum is the same as N_S (for benchmark purposes, even if it is unrealistic that a trial in paediatric population will have the same size as the trial in adults), and the minimum is a much lower value, but still being realistic for a trial in paediatrics. When there is more than one source trial, we varied N_T in a range of values where the maximum is the largest sample size among the source studies.

As a consequence, we included cases where $N_T = N_S$, $N_T = N_S/2$, $N_T = N_S/4$ and $N_T = N_S/6$ with a minimum of 20 subjects per arm. The corresponding sample sizes for each case study are given in Table 2.

The sample sizes in each arm of the target study were equal. In several cases, we reused methods implementations, which did not allow to directly take into account different sample sizes in the arms of the source study. In these cases, we computed an equivalent sample size per arm as $n_S = 2\frac{n_S^{(c)} n_S^{(t)}}{n_S^{(c)} + n_S^{(t)}}$.

N_T	Botox	Dapagliflozin	Aprepitant	Belimumab	Teriflunomide	Mepolizumab
N_S	468	267	573	577	761	551
$N_S/2$	234	133	286	289	381	275
$N_S/4$	117	66	143	144	190	137
$N_S/6$	78	44	95	96	95	91

Table 2: Table summarizing the total sample sizes considered for the target study, in each case study.

2.1.4 Drift in treatment effect

The drift in treatment effect, defined as $\delta = \theta_T^{(\text{true})} - \hat{\theta}_S$ (Viele et al. 2014; Lim et al. 2020; Best et al. 2023), is the key driver of bias when using extrapolation. We focused in particular on three scenario categories :

1. the true effect in the target population is the same as the observed treatment effect in the source population ("consistent treatment effect"),
2. the true effect in the target population is half that observed in the source population ("partially consistent treatment effect"),
3. there is no treatment effect in the target population.

From a regulatory perspective, we are particularly interested in drift values corresponding to target treatment effect $\theta_T^{(\text{true})} \in [\theta_0, \hat{\theta}_S]$, that is, a drift in $[\theta_0 - \hat{\theta}_S, 0]$. This 'critical' interval should always be covered when exploring how OCs vary with drift. However, with an adaptive borrowing method, the probability of meeting the decision criterion, $\Pr(\text{Study success}|\mathbf{D}_T = d_T, \mathbf{D}_S = d_S)$, is expected to reach a maximum at some drift value beyond which source data start being discarded. For the study of adaptive borrowing methods, it is thus important to select a range of drift wide enough for this discarding phenomenon to be observed.

To determine the range of drift to consider for a given case study, we propose the following rationale when the treatment effect follows a normal distribution: one may consider that if the overlap between the posterior distribution of the treatment effect in the source study $p(\theta_S|\mathbf{y}_S)$ and the target study $p(\theta_T|\mathbf{y}_T)$ is very small, the source study should be discarded. To include this idea in our simulation framework, we analytically determine, for a given value of θ_T, the Hellinger distance between $\mathcal{N}\left(\hat{\theta}_S, \sigma^2_{\theta_S}\right)$, where σ_{θ_S} is the standard error on θ_S, and $\mathcal{N}\left(\hat{\theta}_S + \delta, \sigma^2_{\theta_T}\right)$, where σ_{θ_T} is the standard error on θ_T derived from the observed target study data alone. The Hellinger distance $H(f,g)$ between two probability distributions f and g is defined as :

$$H^2(f,g) = \frac{1}{2}\int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx = 1 - \int \sqrt{f(x)g(x)}dx$$

We determine the value of the negative drift for which the Hellinger distance reaches 0.9, and use this as the lower boundary of the drift ranges considered. Beyond such an extreme value for the observed drift,

borrowing from source data can be considered futile. Note that, for simplicity, we used the same drift range for all scenarios in a given case study, irrespective of later changes introduced in the denominator of source ratio-like summary measures or target study sampling standard deviation.

Note that in case where the posterior predictive $p(\overline{y}_T|\theta_T, \sigma^2_{\theta_T})$ is very wide, it may not be guaranteed that the range $[\theta_0 - \hat{\theta}_S, 0]$ is included within the drift range obtained with the above method (noted $\mathscr{R}$). Although this case may happen in very rare cases given the quite conservative threshold of 0.9 considered, we used the range $\mathscr{R} \cup [\theta_0 - \hat{\theta}_S, 0]$ for the drift in practice. We observed that $[\theta_0 - \hat{\theta}_S, 0] \subset \mathscr{R}$ in all case studies considered.

When we model the distribution of the treatment effect using a difference of rates, the likelihood is : $p(\mathscr{D}_T|p_T^{(c)}, p_T^{(t)}) = B(n_T^{(c)}|N_T^{(c)}, p_T^{(c)})B(n_T^{(t)}|N_T^{(t)}, p_T^{(t)})$ where :

- n_T^a : number of responders in arm a (c : control, t : target) of the target trial.
- N_T^a: number of subjects in arm a of the target trial.
- $p_T^{(c)}$: response rate in arm a of the target trial.

The likelihood is therefore a product of binomials, and the treatment effect $\theta_T = p_T^{(t)} - p_T^{(c)}$ spans the range $[-1, 1]$, therefore the drift spans the interval $[-1 - \hat{\theta}_S, 1 - \hat{\theta}_S]$. Moreover, we need to ensure that $p_T^{(t)}$ and $p_T^{(c)}$ are within $[0, 1]$. Since we assume $p_T^{(c)} = \hat{p}_S^{(c)}$, we have:

$$\begin{aligned} p_T^{(t)} &= \theta_T + p_T^{(c)} \\ &= \delta + \hat{\theta}_S + \hat{p}_S^{(c)} = \delta + \hat{p}_S^{(t)} \end{aligned} \tag{1}$$

This implies the following constraint: $-\hat{p}_S^{(t)} \leq \delta \leq 1 - \hat{p}_S^{(t)}$. By combining these two constraints, the drift interval is $\mathscr{R} = [\max(-1 - \hat{\theta}_S, -\hat{p}_S^{(t)}), \min(1 - \hat{\theta}_S, 1 - \hat{p}_S^{(t)})]$.

The corresponding drift ranges considered for each case study are listed in Table 3. We considered evenly spaced values in the range of drift. Note that, for computational cost reasons, we did not use the same number of drift values for each method and each case study. We used between 23 and 33 drift values in total. Note that a fine enough resolution is important for the later computation of Bayesian operating characteristics (see Section 7).

Case study	**Drift range**	**Drift with $\theta_T^{(true)} = \theta_0$**	**Treatment effect range**
Belimumab	[-1.02,1.02]	-0.48	[-0.541,1.5]
Botox	[-0.365,0.365]	-0.2	[-0.165,0.565]
Dapagliflozin	[-0.707,0.707]	-0.36	[-0.347,1.07]
Mepolizumab	[-1.53,1.53]	0.693	[-2.23,0.839]
Aprepitant	[-0.657,0.343]	-0.132	[-0.526,0.474]
Teriflunomide	[-0.588,0.588]	0.411	[-0.999,0.177]

Table 3: Drift ranges considered for each case study.

2.1.5 Changes in the denominator of source ratio summary measures

We intended to determine if changes in the denominator value of a ratio-like summary measure (i.e. RR, OR, HR) have an impact on the operating characteristics. To do so, two additional values are considered for the denominator of the source study summary measures: 1/2 and 3/2 of the original study value, while keeping the value of the treatment effect in the source study constant. Such change implies a change in the standard error on the treatment effect in the source study.

2.1.6 Data generation and sampling approximations

The target study data were generated as follows :

1. For ratio-like summary measures, we computed the standard error on the source study treatment effect (which, because of the potential change of denominator, will differ from the original value in the source study).

2. For ratio-like summary measures, set the value of the control arm response rate in the target study equal to the control arm in the source study (we assume that differences between the source and target study treatment effects come from the treatment arms).
3. Set the value of the treatment effect in the target study, depending on the drift, and compute the corresponding value for the target treatment arm response rate.
4. Compute the corresponding standard error on the treatment effect estimate in the target study.
5. Sample aggregate target study data.

When generating aggregate data for simulated trials, two alternatives can be considered: a first approach is to generate aggregate data following the true data-generating mechanism. Another approach, computationally more efficient in some cases, is to generate the summary aggregate data by assuming a sampling mechanism that matches the likelihood used at the analysis stage (later referred to as "approximate sampling"). Below, we detail the approaches used for sampling aggregate data for each case study.

2.1.7 Data generation for continuous endpoints

For continuous endpoints, we simply sampled patient-level data from $\mathcal{N}\left(\hat{\theta}_S + \delta, \sigma_T^2\right)$. The corresponding summary measures (estimate of the mean and standard error on the mean) were then computed. The target data sampling variance σ_T^2 was set as a scenario parameter. Note that this is not the variance used at the analysis stage. At the analysis stage, we assumed that the target data variance is known, and equal to the empirical variance in the target data sample, $\hat{\sigma}_T^2$.

2.1.8 Data generation for binary endpoints

We assumed that the response rates are the same in the source and target studies control arms. So, for drift δ, the response rate in the target study treatment arm is: $p_T^{(t)} = \frac{e^{\delta}}{e^{\delta}+1/\text{odds}_S}$, where odds_S is the observed odds in the source study.

- True data-generating process: To generate summary measures that are log odds ratio, we sampled $n_T^{(c)} \sim \mathcal{B}(n_T^{(c)}|N_T^{(c)}, p_T^{(c)})$ and $n_T^{(t)} \sim \mathcal{B}(n_T^{(t)}|N_T^{(t)}, p_T^{(t)})$, where :
  - n_T^a : number of responders in arm a (c : control, t : target) of the target trial.
  - N_T^a: number of subjects in arm a of the target trial.
  - $p_T^{(c)}$: response rate in arm a of the target trial.

  Then, we computed the corresponding estimated rates : $\hat{p}_T^a = \frac{n_T^a}{N_T^a}$, and finally, the summary measure of the treatment effect: $\hat{\theta}_T = \log\left(\frac{\hat{p}_T^{(t)}/(1-\hat{p}_T^{(t)})}{\hat{p}_T^{(c)}/(1-\hat{p}_T^{(c)})}\right)$. Additionally, we estimated the standard error on the treatment effect as: $\hat{\sigma}_{\theta_T} = \sqrt{\frac{1}{n_T^{(c)}} + \frac{1}{N_T^{(c)}-n_T^{(c)}} + \frac{1}{n_T^{(t)}} + \frac{1}{N_T^{(t)}-n_T^{(c)}}}$.
- In practice, we did not use approximate sampling in this case. However, the code allows sampling normally distributed patient-level data from $\mathcal{N}\left(\hat{\theta}_S + \delta, \frac{1}{p_T^{(c)}N_T^{(c)}} + \frac{1}{(1-p_T^{(c)})N_T^{(c)}} + \frac{1}{p_T^{(t)}N_T^{(t)}} + \frac{1}{(1-p_T^{(t)})N_T^{(t)}}\right)$, and compute the corresponding aggregate summary measures.

2.1.9 Data generation for time-to-event endpoints

When the endpoint is a time-to-first event, the rates in the control arm and treatment arm of the target study are $\lambda_T^{(c)}$ and $\lambda_T^{(t)}$ respectively. We assume $\lambda_T^{(c)} = \lambda_S^{(c)}$, so that $\lambda_T^{(t)} = e^{\delta}\lambda_S^{(t)}$.

- True data-generating process: We will assume that event times follow a Poisson distribution. Therefore, we sample times-to-first event from an exponential distribution, with arm-specific rates $\lambda_T^{(c)}$ and $\lambda_T^{(t)}$. Due to maximum follow-up time, we additionally apply right-censorship. We then perform a survival analysis using an exponential regression model to estimate the rates in each arm. The

treatment effect is defined as $\hat{\theta}_T = \log(\hat{\lambda}_T^{(t)} / \hat{\lambda}_T^{(c)})$, where the λs are rate parameters of an exponential distribution. The standard error on the log rate ratio is approximately $\sqrt{\frac{1}{n_T^{(t)}} + \frac{1}{n_T^{(c)}}}$, where n_T^a is the number of events observed in arm a.

- Sampling from a Gaussian: To limit computational time, we used approximate sampling in this case. To do so, we first sample a number of events in each arm a, $n_T^{(a)}$, from $\mathcal{P}(\lambda_T^{(a)} \Delta t N_T^{(a)})$, where Δt is the maximum follow-up time. We then sample summary measures of the treatment effect from $\mathcal{N}\left(\log(\lambda_T^{(t)} / \lambda_T^{(c)}), \sqrt{\frac{1}{n_T^{(t)}} + \frac{1}{n_T^{(c)}}}\right)$. Note that we do not sample directly from $\mathcal{N}\left(\log(\lambda_T^{(t)} / \lambda_T^{(c)}), \sqrt{(\lambda_T^{(c)} \Delta t N_T^{(c)})^{-1} + (\lambda_T^{(t)} \Delta t N_T^{(t)})^{-1}}\right)$ as we observed that this does not provide an accurate approximation to the true data-generating process. However, when comparing the power of a frequentist t-test for comparison with Bayesian methods, we assume that the standard error on the log rates ratio is $\sqrt{(\lambda_T^{(c)} \Delta t N_T^{(c)})^{-1} + (\lambda_T^{(t)} \Delta t N_T^{(t)})^{-1}}$ in this case.

The Teriflunomide case study (time-to-event endpoints) is the only case study for which we used approximate sampling in order to gain computational speed.

2.1.10 Data generation for recurrent event endpoints

- True data-generating process: The original study used a negative binomial regression, so we do the same by sampling IPD from a negative binomial distribution, and then estimating the parameters of this distribution from the data.

  The negative binomial distribution can be parameterized using its mean μ and the dispersion parameter k. The mean μ is the expected number of failures before achieving k successes. For this parameterization, the mean μ and variance σ^2 are related as follows: $\mu = \frac{k(1-p)}{p}$, $\sigma^2 = \frac{k(1-p)}{p^2}$, where p is a success probability.

  From the mean equation, we get $p = \frac{k}{k+\mu}$, and substituting p back into the variance equation: $\sigma = \sqrt{\mu + \frac{\mu^2}{k}}$. Assuming a normal distribution for the mean, the standard error of the mean is therefore :

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{k(1-p)}{np^2}}$$

  Therefore, using the delta method, the standard error of the log event rate ratio is approximated as:

$$\text{SE}\left(\log\left(\frac{\lambda_t}{\lambda_c}\right)\right) \approx \sqrt{\left(\frac{\text{SE}(\lambda_t)}{\lambda_t}\right)^2 + \left(\frac{\text{SE}(\lambda_c)}{\lambda_c}\right)^2}$$

  So that :

$$\text{SE}\left(\log\left(\frac{\lambda_t}{\lambda_c}\right)\right) \approx \sqrt{\frac{1}{n_t} \cdot \left(\frac{\sqrt{\mu_t + \frac{\mu_t^2}{k}}}{\mu_t}\right)^2 + \frac{1}{n_c} \cdot \left(\frac{\sqrt{\mu_c + \frac{\mu_c^2}{k}}}{\mu_c}\right)^2}$$

- Approximate sampling: Again, in practice, we did not use approximate sampling in this case. However, the code allows sampling directly from :

$$\mathcal{N}\left(\log\left(\frac{\lambda_t}{\lambda_c}\right), \sqrt{\frac{1}{n_t} \cdot \left(\frac{\sqrt{\mu_t + \frac{\mu_t^2}{k}}}{\mu_t}\right)^2 + \frac{1}{n_c} \cdot \left(\frac{\sqrt{\mu_c + \frac{\mu_c^2}{k}}}{\mu_c}\right)^2}\right)$$

Note that this approach could, in case of very small sample sizes in the target study, lead to large variability in the standard error on the log rate ratio.

2.1.11 Decision criterion

We considered a one-sided null hypothesis $\theta_T \leq \theta_0$ for all case studies except for the Teriflunomide and the Mepolizumab case studies (see section 10), for which the null hypothesis was $\theta_T \geq \theta_0$. For all scenarios considered, we chose $\theta_0 = 0$ (commonly used to show the superiority of the experimental treatment over the control treatment).

We denote Θ_0 the null hypothesis space. Given observed data d_S and d_T in the source and target study respectively, it was concluded that $\theta_T \notin \Theta_0$ if the posterior probability $\Pr(\theta_T \notin \Theta_0 | \mathbf{D}_T = d_T, \mathbf{D}_S = d_S) > \eta$, with $\eta = 0.975$. This critical value η is chosen as it is equivalent to requiring the lower limit of the 95% posterior credible interval calculated with the equal-tail method (i.e. with limits corresponding to the quantiles 2.5% and 97.5% of the posterior distribution) for the treatment effect to be outside Θ_0.

3 Inference in the target trial

3.1 Hypotheses for the outcome distributions

3.1.1 Variance in the target study

Estimating the within-group heterogeneity in the target trial can be challenging given the typically small sample size in the target population. It would thus be of interest to (partially) borrow the variance from adults. In particular, at the design stage, it may be possible to assume that the standard deviation of the outcome distribution is the same as in the source study. To our knowledge, however, the topic of borrowing outcome variance from a source population has only been considered by (Hobbs et al. 2011), who introduced the location-scale commensurate prior. In particular, the effect of potential drift between observed variance in the source population and variance in the target population has not been investigated. However, investigating this problem would add significant complexity to the study, and slightly depart from our initial objectives, therefore we leave this topic for future research. As a consequence, when assuming a Gaussian likelihood when analyzing the data, we assumed that the standard error of the summary measure in the target population is known, and we set the standard deviation to the sample standard deviation in the target study, as is often done in meta-analytic approaches and in Bayesian borrowing (Weber et al. 2018; Best et al. 2021).

However, in practice, variance of the individual outcome may be substantially larger in the target study. For example, paediatric populations tend to be less homogeneous compared to adults because, for instance, of change in weight with age, organ maturation, and body composition differences (Kern 2009). Therefore, we included an additional simulation scenario for the case studies with continuous endpoints (Botox and Dapagliflozin, see section 10) where the simulated variance in the paediatric data is two times larger than the variance observed in adults.

3.1.2 Likelihood

For all case studies except the Aprepitant case study, we will assume that the summary measure of the target study is normally distributed. Therefore, in these cases, $p(\hat{\theta}_T | \theta_T, \sigma^2_{\theta_T}) = \mathcal{N}(\hat{\theta}_T \mid \theta_T, \sigma^2_{\theta_T})$, where $\sigma^2_{\theta_T}$ is the standard error on the target treatment effect, which, as explained above, is assumed known and estimated based on the target data sample. Since we assume that the likelihood from both the target and source study are from the same family, we also have: $p(\hat{\theta}_S | \theta_S, \sigma^2_{\theta_S}) = \mathcal{N}(\hat{\theta}_S \mid \theta_S, \sigma^2_{\theta_S})$.

In the Aprepitant case study, by contrast, the source data consists of $N_S^{(c)}$ (resp. $N_S^{(t)}$) Bernoulli trials with $y_S^{(c)}$ (resp. $y_S^{(t)}$) successes in the control arm (resp. the treatment arm), that is :

$$\begin{aligned} y_T^{(c)} \mid p_T^{(c)} &\sim \text{Bin}(p_T^{(c)}, N_T) \\ y_T^{(t)} \mid p_T^{(t)} &\sim \text{Bin}(p_T^{(t)}, N_T^{(t)}) \end{aligned} \tag{2}$$

The model structure is described in Figure 2.

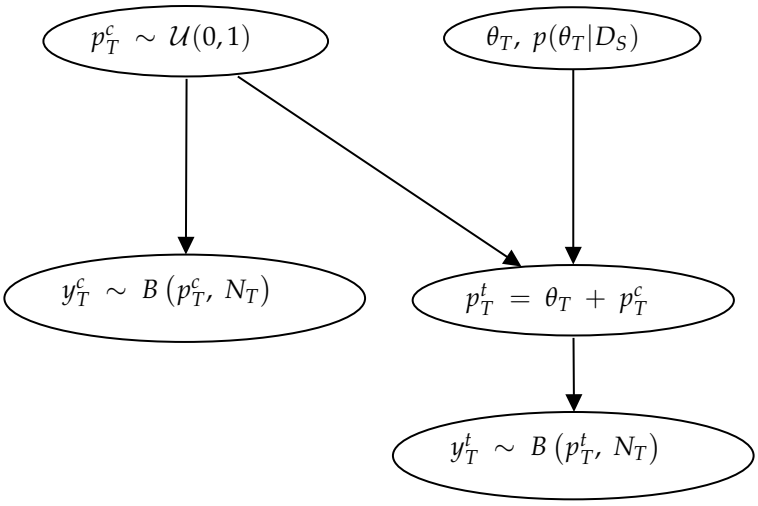

Figure 2: Structure of the model in case where the likelihood is a product of binomials.

The likelihood $\mathcal{L}(\theta_T|\mathbf{D}_T)$ is therefore :

$$\begin{aligned} p(\mathbf{D}_T|\theta_T) &= \int_{p_T^{(c)}=0}^{1} \int_{p_T^{(t)}=\max(\theta_T,0)}^{\min(1+\theta_T,1)} p(\mathbf{D}_T|\theta_T, p_T^{(t)}, p_T^{(c)}) p(p_T^{(t)}, p_T^{(c)}|\theta_T) dp_T^{(t)} dp_T^{(c)} \\ &= \int_0^1 p(\mathbf{D}_T|p_T^{(t)} = \theta_T + p_T^{(c)}, p_T^{(c)}) p(p_T^{(c)}) dp_T^{(c)} \\ &= \int_0^1 \text{Bin}\left(y_T^{(c)}|p_T^{(c)}, N_T^{(c)}\right) \text{Bin}\left(y_T^{(t)}|\theta_T + p_T^{(c)}, N_T^{(t)}\right) p(p_T^{(c)}) dp_T^{(c)} \end{aligned} \tag{3}$$

Here, we do not simply borrow the treatment effect, but the response rates in each arm. Indeed, we cannot use a binomial likelihood for the data without specifying each rate. However, only the prior on the treatment effect explicitly incorporates the source study data.

Note also that, although we assumed in the target data generating process that $p_T^{(c)} = p_S^{(c)}$, we do not make this assumption when analyzing the data and put a uniform prior on $p_T^{(c)}$ instead.

3.2 Use of source studies data in inference

Where Bayesian methods are applied, they are conditioned on the observed result in the source clinical trial(s) d_S. Indeed this corresponds to the situation of interest for the project where historical data are known and can be used to inform the analysis of a new trial. Most previous simulation studies in the field used fixed source datasets sampled from known distributions: either many source datasets (see e.g. Shi et al. (2023), Pan et al. (2022), Pan et al. (2017), Chu and Yuan (2018), Kaizer et al. (2018), Holzhauer et al. (2018), Brard et al. (2019), Rosmalen et al. (2018), Jiao et al. (2019), Su et al. (2022), and Gravestock and Held (2019)), or single datasets (Gravestock et al. 2017; Hupf et al. 2021; Viele et al. 2014; Feißt et al. 2020; Han et al. 2017). By contrast, in this study, we considered existing datasets as source studies.

When multiple source studies were selected for a given target study, for simplicity, we aggregated their results by simply pooling them. This is only the case for the Belimumab and Teriflunomide studies, where adults data come from two studies with identical designs.

3.3 Prior on the source study treatment effect

Even if the source data are kept fixed, several Bayesian borrowing methods need an initial prior $\pi_0(\theta_S)$ (i.e. prior before extrapolation) to be specified. This is the case, for example, with the family of power priors.

We used "noninformative" or weakly informative initial priors on the treatment effect in both the source and target populations, in a way that matches the summary statistics. That is:

- for normally distributed treatment effect, when using a normal approximation on the log(OR) or log(HR), defining a prior is slightly involved. Note that, for two random variables X and Y such that

$X = \log(Y)$, $p(Y) = |\frac{\partial X}{\partial Y}| p(\log(Y))$. Accordingly, if $p(X) \propto \mathbb{I}(X \in [a, b])$, then $p(Y) \propto \frac{\mathbb{I}(\log(Y) \in [a,b])}{Y}$. Therefore, if we put, for example, a Gaussian prior on the treatment effect on the log scale, then this corresponds to putting a lot of probability mass on small values on the raw scale. However, it is common in Bayesian analysis to put a wide normal prior on the log scale (see for example Smith et al. (1995) and Al Amer et al. (2021)), and this has been done also in the context of Bayesian borrowing (Weber et al. (2018), for example, use $\mathcal{N}(0, 10)$, Nikolakopoulos et al. (2018) assumed a flat prior). An advantage is that this allows for conjugate analysis with a number of borrowing methods, hence allowing analytical derivation of the posterior. The fact that much more weight is put, on the natural scale, on small values of the treatment effect compared to large values, can however be seen as problematic. We conducted some pilot tests to compare results obtained with either normal or uniform priors, for each case study and without borrowing, to determine a sensible proper normal prior that does not put too much probability mass on large negative values that, when backtransformed to the original scale, will lead to very small values. We opted for $\mathcal{N}(0, 1000)$, as larger variance would lead to putting almost no mass on moderate values of the treatment effect on the raw scale, whereas smaller variance would lead to non-vague priors. Such a prior would lead to virtually no difference compared to using a flat prior. However, when using the Normalized Power Prior (Duan et al. 2006), the Empirical Bayes Power Prior (Gravestock et al. 2017), and PDCCPP Nikolakopoulos et al. (2018), either because we relied on existing implementation or on analytic derivations, we had to assume flat initial priors.

- for normally distributed treatment effect and normally distributed endpoints, we used, for consistency, a similar vague normal prior on the treatment effect : $\mathcal{N}(0, 1000)$.

- when the likelihood was defined on the rates in each arm (in the Aprepitant case study), we used uniform priors on the control rate, $p_c \sim \mathcal{U}(-1, 1)$, and a uniform prior on the treatment effect, $\theta_T | p_c \sim \mathcal{U}(-p_c, 1 - p_c)$.

3.4 Estimation of posterior distribution of model parameters

Markov Chain Monte Carlo In the Bayesian framework, all information about the target treatment effect is summarized in the posterior distribution $p(\theta_T \mid \mathbf{D}_S = d_S, \mathbf{D}_T = d_T)$.

In many cases, however, this posterior distribution cannot be computed analytically, but several methods exist to approximate it. In the simulation study, when possible, we relied on numerical integration to compute the posterior distribution, or on Markov chain Monte Carlo (MCMC) simulation techniques to draw approximate samples from the posterior distribution. These samples then allowed us to estimate quantities of interest, such as the posterior mean, median, and other quantiles.

We used the probabilistic programming language Stan for running MCMC. Given that all parameters in the models are continuous, we used the default sampler in Stan, the No-U-Turn Sampler (NUTS, Hoffman and Gelman (2011)), an advanced and highly efficient MCMC sampling algorithm. Unless required because of convergence issues or strong autocorrelation, we used Stan's default parameters for NUTS .

Number of chains Using multiple chains with random initial values makes the convergence diagnostic more accurate (see Section 3.4), and is safer in situations where the posterior distribution is multi-modal. That is, it mitigates the risk of having the chain circumscribed around a mode, and potentially allows identifying multimodality. This can lead to a better approximation of the posterior, even if between-chain mixing is not achieved. As a consequence, we used 4 chains.

Initial values Treatment effect parameters and hyperparameters were initialized by taking samples from their respective prior (or hyperprior) distributions.

Thinning In the past, it was common to "thin" MCMC simulations, discarding all but every k^{th} sampled values to reduce autocorrelation while reducing memory burden. However, several renowned MCMC experts such as Pr Christian Robert or Dr Thomas Wiecki advocated against using thinning - except maybe in very specific situations. Indeed, in practice, it is still better to keep autocorrelated samples than to remove them, and samplers such as NUTS tend to have very low auto-correlation. Therefore, we did not use thinning.

MCMC Effective Sample Size The MCMC effective sample size (MCMC ESS) represents the number of independent samples from the posterior distribution that provide the same amount of information as the correlated draws generated by MCMC. In other words, it quantifies the efficiency of the MCMC algorithm

Method	Fixed parameters/Priors	Parameters to vary	Range of variation
Test-then-pool, equivalence test	None	Significance level of the equivalence test η. Equivalence margin λ.	$\eta \in \{0.1, 0.5\}$, $\lambda \in \{0.1, 0.5, 0.8\}$
Test-then-pool, difference test	None	Significance level of the difference test η	$\eta \in \{0.1, 0.5\}$
Conditional power prior (PP)	Initial prior on θ	Power parameter γ	$\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$
Normalized PP	Initial prior on θ $\gamma \sim Beta(\xi_\gamma/\omega_\gamma, (1-\xi_\gamma)/\omega_\gamma)$. $\xi_\gamma = 0.5$	ω_γ	ω_γ is varied so that the standard deviation of the Beta prior ranges from 0 to 0.50
Empirical Bayes PP	Initial prior on θ	None	None
p-value based PP	Initial prior on θ	Shape parameter k	$k \in \{0.01, 0.1, 1, 10, 20\}$
Commensurate PP	Initial prior on θ_S	Prior on the commensurability parameter τ	See the text
Robust mixture prior	Variance of the vague component	Mixture weight w.	w in a grid of values ranging from 0 to 1 in steps of 0.1.

Table 4: Methods and parameters considered in the simulation study. When the method is based on a consistency assumption ($\theta_T = \theta_S$), we denote the treatment effect as θ.

Parameter	**Heterogeneity prior family**	**Heterogeneity prior**
τ^2	Inverse Gamma	$\Gamma^{-1}(1/3, 1)$
τ^2	Inverse Gamma	$\Gamma^{-1}(1/7, 1)$
τ^2	Inverse Gamma	$\Gamma^{-1}(1/1000, 1)$
τ	Half Normal	$\mathcal{HN}(1)$
τ	Half Normal	$\mathcal{HN}(0.5)$
$log(\tau)$	Cauchy	$Cauchy(0, 30)$

Table 5: Priors on the heterogeneity parameter. Priors taken from Weber et al. (2018) and Hobbs et al. (2011).

in exploring the posterior distribution. An MCMC ESS is estimated for each parameter. Aiming for a sufficiently large MCMC ESS is crucial for reliable estimation.

Moreover, a crucial quantity estimated from MCMC draws is $\Pr(\theta_T \notin \Theta_0)$. Indeed, it is concluded that the treatment is effective if $\Pr(\theta_T \notin \Theta_0) > \eta$, with $\eta = 0.975$. Therefore, we need to ensure that we get a precise estimate of the 0.975th sample quantile.

The reasoning used to determine the standard deviation of sample quantiles is given in appendix A, allowing us to conclude that if we want the 0.975th sample quantile to be estimated with the same precision as the median, we would need $1/0.47 = 2.14$ times more samples.

Based on these considerations, we ensured that the MCMC effective sample size for the target trial treatment effect parameter θ_T is at least 10,000 and adapt the chains' length consequently. Assuming a posterior that is a standard normal distribution, this would correspond to a standard error on the median estimate of 0.0125, and a standard error on the 0.975th sample quantile estimate of 0.0267. Concretely, with $N_C = 4$ chains of length L, for each simulated data replicate, we computed the MCMC ESS for the target treatment effect ϵ_{θ_T}. We then adjusted the chain length so that $L \leftarrow 1.1 \times L \times \epsilon_{\theta_T}/\epsilon$, where ϵ is the target MCMC ESS of 10,000, and repeated the iteration until sufficient MCMC, that is, until $\epsilon_{\theta_T} > \epsilon$. To avoid an explosion of chains length, we capped L to 10,000. By contrast, for speed gains, we reduced chain lengths when $\epsilon_{\theta_T} > 1.1 \times \epsilon$, applying $L \leftarrow 1.1 \times L \times \epsilon_{\theta_T}/\epsilon$, and proceeded to the next data replicates.

Convergence diagnostics By definition, a Markov chain generates samples from the target distribution only after it has converged to equilibrium. In theory, convergence is only guaranteed asymptotically, therefore, in practice, diagnostics must be applied to monitor convergence for the finite number of draws actually available. Therefore, at the model development stage, when using MCMC, Markov Chains visual inspection was performed using tools such as trace plots and autocorrelation plots to verify that the MCMC chains have reached a stationary distribution. To automate the MCMC convergence diagnostic for each replicate, we used the Gelman and Rubin (1992) potential scale reduction statistic $\widehat{R}$ to monitor convergence. $\widehat{R}$ measures the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains. If all chains are at equilibrium, these will be the same and $\widehat{R}$ will be one, and greater otherwise. Gelman and Rubin's recommendation is that the independent Markov chains be initialized with diffuse starting values for the parameters and sampled until all values for $\widehat{R}$ are below 1.1. We also monitored the number of transitions ending with a divergence.

Execution of the code does not stop in case of issues with MCMC inference; rather, a warning is stored in the results table so that the pipeline is not interrupted. In case of convergence issues, we adapted the MCMC algorithm by increasing the acceptance probability of the sampler, the tuning period, and reparameterizing the distribution. These convergence diagnoses also allowed us to determine if some models have particular behaviors that need specific handling.

3.5 Sources of uncertainty

It is important, when reporting Bayesian estimates such as the posterior mean, to report the associated uncertainty. However, this uncertainty originates from two sources that we need to distinguish: on one hand, the uncertainty in the estimates which is due to the finite number of simulations and, when applicable, the finite length of Markov chains (which, taken together, constitute the Monte Carlo error); on the other hand, the posterior uncertainty inherent to any Bayesian estimation method (posterior uncertainty), which is due to the finite amount of data.

Therefore, for each operating characteristic estimate, we reported the uncertainty due to the finite number of simulations through the Monte Carlo Standard Error (MCSE) of Estimate (Morris et al. 2019) or Monte Carlo Confidence Intervals. For each MC estimate, the corresponding MCSE formula is given in Table 6 of Morris et al. (2019). In figures plotting the MC estimates, we displayed MC error as 95% confidence intervals. These confidence intervals were estimated using nonparametric bootstrap for metrics other than coverage and probability of success, for which we know the true underlying distribution. These confidence intervals were not estimated by assuming normality as we noticed this could lead to problematic confidence intervals for metrics such as the mean MSE despite the large number of replicates. The MCSE for the probability of study success is given by $\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{sim}}}$, where $\hat{p}$ is the estimate of the probability of success. The relative Monte Carlo standard error (RMCSE) is $\sqrt{\frac{1-\hat{p}}{n_{sim}\hat{p}}}$. For a number of replicates of 10, 000, and $\hat{p} = 0.025$, this gives an RMCSE of 6.3%, which can be considered reasonable. The MCSE and RMCSE were estimated for each MC estimate.

4 Adaptation of existing methods to the settings of interest

Normal likelihood When a normal likelihood was assumed, adapting existing methods that were developed to borrow the control arm only to borrow the treatment effect was straightforward. Indeed, we only had to define a prior on the treatment effect instead of the control arm summary measure and to use as likelihood $\mathcal{N}(\hat{\theta}_T \mid \theta_T, \sigma^2_{\theta_T})$ instead of $\mathcal{N}(\hat{p}_T^{(t)} \mid p_T^{(t)}, \sigma^2_{p_T^{(t)}})$.

Binomial likelihood Adapting methods that borrow the control arm with a binomial likelihood to borrow the treatment effect, with the model structure in Figure 2, is far from straightforward. In these cases, as described in 5, we sometimes did not adapt the method and used a normal likelihood instead.

To adapt methods that were developed to be used with normal likelihoods, we used the model structure described in 2, but using a truncated normal as the prior distribution $p(\theta_T|\mathbf{D}_S)$, so as to ensure that the treatment effect θ_T remains within valid range.

5 Selected statistical methods

The choice of statistical methods to be considered for the simulation study is based on an extensive literature review. Several existing methods can be shown to be equivalent in the setting we consider, and others are improved versions by the same authors. This led us to the selection of models described in this section. For each method, we varied the parameters that affect the amount of borrowing. These parameters are summarized in Table 4.

5.1 Separate analysis and pooling

For each borrowing method, a comparison was made against the power of frequentist analyses that use either full borrowing (pooling) or no borrowing (separate) at the nominal type 1 error rate of 2.5%. Note that, when using Bayesian methods, the empirical variance is estimated from the sample data. Therefore, when the likelihood is Gaussian, the corresponding frequentist test is a t-test. For each method of interest, the power of the t-test was evaluated at different significance levels that depend on the unconditional type 1 error rate of the borrowing method of interest. When the likelihood is given by Figure 2 (Aprepitant case study), we used a test of difference of proportions based on Cohen's h.

For comparison of other operating characteristics and inference metrics, we also implemented Bayesian analyses that pool the data or perform a separate analysis. With vague priors, these are equivalent to the frequentist analyses.

5.2 Conditional Power Prior

5.2.1 Method description

As a Bayesian baseline, and to investigate the effect of borrowing without adaptation to prior-data conflict, we started by investigating the effect of fixed borrowing with discounted adults posteriors as priors.

In order to incorporate a fixed amount of information from source studies into the prior for θ_T, Ibrahim and Chen (2000) introduced the power prior (also referred to as the conditional power prior (CPP) (Neelon and O'Malley 2010)):

$$\pi(\theta_T|\mathbf{D}_S, \gamma) \propto \mathscr{L}(\theta_T|\mathbf{D}_S)^{\gamma}\pi_0(\theta_T), \tag{4}$$

where $\gamma \in [0, 1]$, and $\pi_0(\theta_T)$ denotes the so-called "initial" prior distribution for θ_T. The main feature of the method is that the impact of source data on the posterior distribution can be controlled by choosing the value of the power parameter γ, thus providing a simple way of discounting prior information. When $\gamma = 1$, data from the source and target study are pooled, whereas if $\gamma = 0$, data from the source study are discarded. This power parameter allows smoothly changing the analysis from no borrowing to pooling. This method assumes that the parameter of interest θ_T is the same in the source and target studies. In the Normal-Normal model, this is equivalent to inflating the prior variance by a factor $1/\gamma$.

5.2.2 Parameters to be varied

The power parameter γ was taken in the set $\{0, 0.25, 0.5, 0.75, 1\}$.

5.2.3 Implementation

For normal likelihood, we used a custom implementation using the analytical posterior. In the Aprepitant case study, we used a custom implementation that relied on Stan for MCMC inference.

5.3 Frequentist test-then-pool

5.3.1 Method description

The most common frequentist borrowing method is test-then-pool. The idea is to assess the difference between source and target data before deciding whether to pool the data or not. In this test-then-pool approach (Viele et al. 2014), the hypothesis $H_0 : \theta_T = \theta_S$ is tested, where again, θ_T and θ_S denote the variable of interest (such as the response rate or treatment effect) for the target study and the source study, respectively. If H_0 is rejected, this indicates that the data should not be pooled, and should be analyzed independently. So the source data are either fully borrowed or not borrowed at all, depending on the test result. Liu (2018) argues that testing the difference between θ_S and θ_T through the p-value may not be the best approach to evaluate whether one can reasonably pool the data or not. Indeed, the power of the test depends not only on the difference $\theta_S - \theta_T$, but also on factors such as the sample sizes. For instance, a small sample size in the control group may result in a non-significant p-value, which could lead to systematically pooling the data even if there is a true difference between the source and target data. To address this issue, Liu (2018) proposed a more conservative approach: testing an equivalence hypothesis instead, with: $H_0 : |\theta_S - \theta_T| > \lambda$ versus $H_1 : |\theta_S - \theta_T| < \lambda$, where $\lambda > 0$ represents a predetermined equivalence margin. They compute the p-value as the maximum of the p-values for testing two one-sided hypotheses: $H_{0a} : \theta_S - \theta_T > \lambda$ and $H_{0b} : \theta_S - \theta_T < -\lambda$ (Schuirmann 1987). Under this approach, a significant p-value implies the rejection of the null hypothesis of non-equivalence. Thus, data was pooled if the p-value was less than a pre-specified level.

We investigated both of these approaches using t-tests. In our implementation, once the frequentist test indicated whether to pool the data or not, we used the Bayesian implementation of the separate analysis (or pooled analysis).

It is not clear, however, how to generalize the Test-then-Pool approach in the case where data are binary and no Gaussian approximation is used for the likelihood (Aprepitant case study). Indeed, for binary data, Viele et al. (2014) used a Fisher exact test to compare the control arms. If we were to borrow the two arms, a sensible approach would be to perform tests (either difference or equivalence tests) on each pair of arms, and pool if both tests agree that the data are consistent. However, it is not clear which frequentist test to use when we focus on the treatment effect, as in this case, we want to test for a difference in differences of proportions. One possible approach would be to compare the likelihood of two models: one in which the data from both studies come from the same distribution, and another in which they come from different distributions, and compute the corresponding Bayes Factor. Based on the Bayes Factor value, the data could then be pooled or analyzed separately. However, setting such a threshold may prove difficult and arbitrary. Therefore, letting this problem for future work, we decided to proceed as in the Gaussian likelihood case, and used a t-test. Once the decision to pool or not the data was taken, we performed the analysis assuming the likelihood in each arm is a binomial.

5.3.2 Parameters to be varied

Borrowing is determined by the significance level of the equivalence/difference test, and the equivalence margin. We chose a significance level of 0.10 and 0.50. Liu (2018) suggested an equivalence margin $\lambda = 0.1$ in proportions for binary data. We considered an equivalence margin of 0.10, 0.50, or 0.80.

5.3.3 Implementation

The test was performed using the BSDA package. Pooling or separate analysis relied on the implementation of the pooling and separate analysis. In the normal likelihood case, pooling and separate analysis were implemented using RBesT to derive the exact posterior distribution. In the Aprepitant case, we used a custom implementation based on Stan for MCMC inference.

5.4 Normalized Power Prior

5.4.1 Method description

In the power prior approach, the power prior parameter γ can be treated as a random variable subject to inference by making use of a prior $\pi(\gamma)$ in a hierarchical model. This gives rise to the normalized power prior (NPP, Duan et al. (2006) and Neuenschwander et al. (2009)). This is achieved by introducing a normalizing constant for $\pi(\gamma)$:

$$C(\gamma) = 1 \Big/ \int \mathscr{L}(\theta_T|\mathbf{D}_S)^\gamma \pi_0(\theta_T) d\theta_T. \tag{5}$$

The normalized prior on (θ_T, γ) is then defined as :

$$\pi(\theta_T, \gamma|\mathbf{D}_S) = C(\gamma)\mathscr{L}(\theta_T|\mathbf{D}_S)^\gamma \pi_0(\theta_T)\pi(\gamma). \tag{6}$$

Analytical derivation for the prior and posterior distribution obtained with a normalized power prior with a normal likelihood, a Beta prior on the power parameter $\gamma \sim \text{Be}(p, q)$, and known standard deviation, can be found in Pawel et al. (2023) or in appendix A of Gravestock et al. (2017). In this setting, the normalized power prior is:

$$\pi\left(\theta_T, \gamma \mid \mathbf{D}_S\right) = \frac{\mathcal{L}\left(\mathbf{D}_S \mid \theta_T\right)^\gamma \pi(\gamma)}{\int_{-\infty}^{+\infty} \mathcal{L}\left(\mathbf{D}_S \mid \theta'_T\right)^\gamma d\theta'_T} = \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \sigma^2_{\theta_S}/\gamma\right) \text{Be}(\gamma \mid p, q)$$

The marginal prior on θ_T is :

$$\begin{aligned}
\pi\left(\theta_T \mid \mathbf{D}_S\right) &= \int_0^1 \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \sigma^2_{\theta_S}/\gamma\right) \text{Be}(\gamma \mid p, q) d\gamma \\
&= \frac{1}{\text{B}(p,q)} \frac{1}{\sqrt{2\pi\sigma^2_{\theta_S}}} \int_0^1 \sqrt{\gamma} \exp\left(\gamma \frac{\left(\hat{\theta}_S - \theta_T\right)^2}{-2\sigma^2_{\theta_S}}\right) \gamma^{p-1}(1-\gamma)^{q-1} d\gamma \\
&= \frac{1}{\text{B}(p,q)} \frac{1}{\sqrt{2\pi\sigma^2_{\theta_S}}} \frac{\Gamma(p+1/2)\Gamma(q)}{\Gamma(p+q+1/2)} \text{M}\left(\frac{1}{2}+p, \frac{1}{2}+p+q, -\frac{\left(\hat{\theta}_S - \theta_T\right)^2}{2\sigma^2_{\theta_S}}\right) \\
&\propto \text{M}\left(\frac{1}{2}+p, \frac{1}{2}+p+q, -\frac{\left(\hat{\theta}_S - \theta_T\right)^2}{2\sigma^2_{\theta_S}}\right),
\end{aligned} \tag{7}$$

where $\text{M}(a, b, z) = 1/(\Gamma(a)\Gamma(b-a)) \int_0^1 e^{zt} t^{a-1}(1-t)^{b-a-1} dt$ is Kummer's confluent hypergeometric function, which is implemented in standard numerical mathematics libraries (note that the term $\Gamma(p+q+1/2)$ in the numerator is omitted in Gravestock et al. (2017)) .

Combining the joint prior $\pi\left(\theta_T, \gamma \mid \mathbf{D}_S\right)$ with the likelihood of the target study data produces a joint posterior for θ_T and γ, that is,

$$\begin{aligned}\pi\left(\theta_T, \gamma \mid \mathbf{D}_T, \mathbf{D}_S\right) &= \frac{\mathcal{L}(\mathbf{D}_T \mid \theta_T)\pi\left(\theta_T, \gamma \mid \mathbf{D}_S\right)}{\int_0^1 \int_{-\infty}^{\infty} \mathcal{L}\left(\mathbf{D}_T \mid \theta_T'\right) \pi\left(\theta_T', \gamma' \mid \mathbf{D}_S\right) d\theta_T' d\gamma'} \\ &= \frac{\mathcal{N}\left(\hat{\theta}_T \mid \theta_T, \sigma_{\theta_T}^2\right) \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \sigma_{\theta_S}^2/\gamma\right) \mathrm{Be}(\gamma \mid p, q)}{\int_0^1 \mathcal{N}\left(\hat{\theta}_T \mid \hat{\theta}_S, \sigma_{\theta_T}^2 + \sigma_{\theta_S}^2/\gamma'\right) \mathrm{Be}\left(\gamma' \mid p, q\right) d\gamma'},\end{aligned} \tag{8}$$

from which a marginal posterior for γ can be obtained by integrating out θ_T, that is,

$$\begin{aligned}\pi\left(\gamma \mid \mathbf{D}_T, \mathbf{D}_S\right) &= \int_{-\infty}^{+\infty} \pi\left(\theta_T, \gamma \mid \mathbf{D}_T, \mathbf{D}_S\right) d\theta_T \\ &= \frac{\mathcal{N}\left(\hat{\theta}_T \mid \hat{\theta}_S, \sigma_{\theta_T}^2 + \sigma_{\theta_S}^2/\gamma\right) \mathrm{Be}(\gamma \mid p, q)}{\int_0^1 \mathcal{N}\left(\hat{\theta}_T \mid \hat{\theta}_S, \sigma_{\theta_T}^2 + \sigma_{\theta_S}^2/\gamma'\right) \mathrm{Be}\left(\gamma' \mid p, q\right) d\gamma'} \\ &\propto \mathcal{N}\left(\hat{\theta}_T \mid \hat{\theta}_S, \sigma_{\theta_T}^2 + \sigma_{\theta_S}^2/\gamma\right) \mathrm{Be}(\gamma \mid p, q).\end{aligned} \tag{9}$$

The posterior distribution of the power parameter can therefore be approximated using numerical integration.

Gravestock et al. (2017) show that:

$$\begin{aligned}\pi\left(\theta_T \mid \mathbf{D}_T, \mathbf{D}_S\right) &= C(\gamma) \int_0^1 \mathcal{N}\left(\hat{\theta}_T \mid \theta_T, \sigma_{\theta_T}^2\right) \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \sigma_{\theta_S}^2/\gamma\right) \mathrm{Be}(\gamma \mid p, q) d\gamma \\ &= C(\gamma) \mathcal{N}\left(\hat{\theta}_T \mid \theta_T, \sigma_{\theta_T}^2\right) \int_0^1 \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \sigma_{\theta_S}^2/\gamma\right) \mathrm{Be}(\gamma \mid p, q) d\gamma \\ &= C(\gamma) \mathcal{N}\left(\hat{\theta}_T \mid \theta_T, \sigma_{\theta_T}^2\right) \pi\left(\theta_T \mid \mathbf{D}_S\right) \\ &\propto \exp\left(-\frac{(\hat{\theta}_T - \theta_T)^2}{2\sigma_{\theta_T}^2}\right) \mathrm{M}\left(\frac{1}{2} + p, \frac{1}{2} + p + q, -\frac{\left(\hat{\theta}_S - \theta_T\right)^2}{2\sigma_{\theta_S}^2}\right).\end{aligned} \tag{10}$$

Generalizing the Normalized Power Prior to borrow treatment effect in the Aprepitant case study is not straightforward. Therefore, in this case, we assumed a normal likelihood.

5.4.2 Parameters to be varied

We used a beta prior on the power parameter: $\gamma \sim Beta(p, q)$, which is a common choice (Gravestock et al. 2017; Shi et al. 2023). To better interpret this prior, we reparameterize it as $\gamma \sim Beta(\xi_\gamma/\omega_\gamma, (1-\xi_\gamma)/\omega_\gamma)$, where $\mathbb{E}[\gamma] = \xi_\gamma$ and $\mathbb{V}[\gamma] = \sigma_\gamma^2 = \frac{\omega_\gamma \xi_\gamma (1-\xi_\gamma)}{1+\omega_\gamma}$. We used $\xi_\gamma = 0.5$, and vary ω_γ so that the standard deviation of the Beta prior ranges from 0 to 0.50. We have: $\omega_\gamma = \frac{\sigma_\gamma^2}{\xi_\gamma(1-\xi_\gamma) - \sigma_\gamma^2}$.

5.4.3 Implementation

When implementing this model, we took inspiration from the code in Pawel et al. (2023), which relies on numerical integration instead of the full analytical expression that includes the confluent hypergeometric function. We noticed that computing the posterior using the full analytical expression was faster than using numerical integration. However, when using adaptive quadrature to compute the mean and variance of the distribution, using numerical integration to obtain the posterior density led to a much faster computation compared to using the full analytical expression, yet with similar accuracy. Therefore, we instead relied on numerical integration to compute the posterior distribution.

5.5 PDCCPP and empirical Bayes PP

5.5.1 Method description

Nikolakopoulos et al. (2018) suggested using a point estimate of the power parameter γ of the power prior that controls type 1 error. Their approach is based on the Box p-value (also called prior-predictive p-value (PPP)). The two-sided PPP for a test statistic T_T (noted t_T when corresponding to observed data) is defined as :

$$ppp(d_S, \gamma) = 2\min[\Pr(T_T \geq t_T | \mathbf{D}_S = d_S, \gamma), \Pr(T_T \leq t_T | \mathbf{D}_S = d_S, \gamma)], \tag{11}$$

where d_T denotes the observed target trial data.

They suggest choosing a fixed value for γ using :

$$\gamma = \min\left[\max_{\gamma \in [0,1]} \{\gamma : ppp(d_S, \gamma) \leq c\}, 1\right], \tag{12}$$

where $ppp(d_S, \gamma)$ indicates that the PPP is evaluated conditionally on d_S and γ, and c is an arbitrary constant. They show that, with normal outcomes, for a one-sided hypothesis test based on the posterior probability of θ_T with an appropriate choice of c, the PDCCPP controls the type 1 error rate at a pre-specified level. Nikolakopoulos et al. (2018) call this variant the prior-data conflict calibrated power prior (PDCCPP).

In the Normal-Normal model, $T_T = \hat{\theta}_T$, and $\hat{\theta}_T | \gamma \sim \mathcal{N}\left(\theta_S, \frac{\sigma_S^2}{\gamma N_S} + \frac{\sigma_T^2}{N_T}\right)$. So the PPP is :

$$\gamma = \begin{cases} \dfrac{\frac{\sigma_S^2}{N_S}}{\left[\left(\frac{\hat{\theta}_T - \hat{\theta}_S}{z_{1-c/2}}\right)^2 - \frac{\sigma_T^2}{N_T}\right]}, & \text{if } \hat{\theta}_T < \hat{\theta}_S + z_{c^*/2}\sigma_{pr} \vee \hat{\theta}_T > \hat{\theta}_S + z_{1-c^*/2}\sigma_{pr} \\ 1, & \text{if } \hat{\theta}_S + z_{c^*/2}\sigma_{pr} \leq \hat{\theta}_T \leq \hat{\theta}_S + z_{1-c^*/2}\sigma_{pr} \end{cases} \tag{13}$$

with $\sigma_{pr} = \sqrt{\frac{\sigma_T^2}{N_T} + \frac{\sigma_S^2}{N_S}}$.

Nikolakopoulos et al. (2018) also introduces a variant of PDCCPP, which is a form of Test-then-Pool based on the predictive p-value, Predictive Test-then-Pool (PTtP). In PTtP:

$$\gamma = \begin{cases} 0, & \text{if } \hat{\theta}_T < \hat{\theta}_S + z_{c/2}\sigma_{pr} \vee \hat{\theta}_T > \hat{\theta}_S + z_{1-c/2}\sigma_{pr} \\ 1, & \text{if } \hat{\theta}_S + z_{c/2}\sigma_{pr} \leq \hat{\theta}_T \leq \hat{\theta}_S + z_{1-c/2}\sigma_{pr} \end{cases}, \tag{14}$$

However, the original publication only includes the code to estimate the calibration parameter c for the PDCCPP, and given the close resemblance between the two methods, we focused on PDCCPP.

Importantly, with $c = 2(1 - \Phi(1))$, the estimator of γ in 12 is equivalent to the maximum likelihood estimator (empirical Bayes power prior) initially studied by Gravestock et al. (2017). Gravestock et al. (2017) (supplementary A) derives an analytical posterior for the empirical power prior in the case of a normal likelihood and a beta prior on γ:

$$\hat{\delta} = \frac{\sigma_{\theta_S}^2}{\max\left\{\left(\hat{\theta}_T - \hat{\theta}_S\right)^2, \sigma_{\theta_T}^2 + \sigma_{\theta_S}^2\right\} - \sigma_{\theta_T}^2}, \tag{15}$$

where the max is required to restrict $\hat{\delta} \leq 1$. The empirical Bayes posterior distribution,

$$p\left(\theta_T \mid \hat{\theta}_T, \hat{\theta}_S, \delta = \hat{\delta}\right) \propto \begin{cases} \mathcal{N}\left(\theta_T \mid \hat{\theta}_T, \sigma_{\theta_T}^2\right) \times \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \left(\hat{\theta}_T - \hat{\theta}_S\right)^2 - \sigma_{\theta_T}^2\right) & \text{if } \left(\hat{\theta}_S - \hat{\theta}_T\right)^2 > \sigma_{\theta_T}^2 + \sigma_{\theta_S}^2 \\ \mathrm{N}\left(\theta_T \mid \hat{\theta}_T, \sigma_{\theta_T}^2\right) \times \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \sigma_{\theta_S}^2\right) & \text{otherwise.} \end{cases} \tag{16}$$

We thus included PDCCPP, as well as the empirical Bayes power prior. Note, however, that PDCCPP and PTtP were only developed in the case of a normal likelihood. Thus, in the Aprepitant case study, for which other methods assume the model structure in Figure 2, we assumed a normal likelihood to use PDCCPP.

5.5.2 Parameters to be varied

Both the PDCCPP and the empirical Bayes power prior do not include free parameters.

5.5.3 Implementation

We adapted the code in Nikolakopoulos et al. (2018).

5.6 P-value based power prior

5.6.1 Method description

In a generalization of the test-then-pool approach, Liu (2018) proposed a method for selecting the power parameter γ in the conditional power prior based on the p-value of an equivalence test between the source and target data. The equation used to determine γ is as follows:

$$\gamma = \exp\left[\frac{k}{1-p}\ln(1-p)\right], \tag{17}$$

where k is a shape parameter that must be specified. This function was chosen so that more source data is borrowed when the p-value is close to 0 (i.e., the non-equivalence null hypothesis is strongly rejected). Larger values of k result in steeper curves (faster decrease from 0 to 1), that is more discounting will be applied to the source data for a given p-valuez. This method can be viewed as an extension of the test-then-pool approach, with the power parameter smoothly adjusting the amount of borrowing from no borrowing to pooling. Again, we used t-tests to compare the source and target studies. For the same reason we assumed a normal likelihood when performing the test in the Test-then-Pool approaches for the Aprepitant case study, again, we perform a t-test to compute the p-value, then analyze the data assuming the model structure in Figure 2.

5.6.2 Parameters to be varied

The shape parameter k determines the amount of borrowing for a given p-value. To understand its impact and give insights into its interpretation, we varied k along a grid of values between 1 and 20, similarly to Liu (2018).

5.6.3 Implementation

The test was performed using the BSDA package. We then reused our Conditional Power Prior implemented by plugging the power parameter value derived from the p-value.

5.7 Commensurate Power Prior

5.7.1 Method description

The commensurate power prior is given by (Hobbs et al. 2011):

$$\pi(\theta_T, \gamma, \tau|\mathbf{D}_S) = \int \pi(\theta_T|\theta_S, \tau)\frac{\mathcal{L}(\theta_S|\mathbf{D}_S)^\gamma \pi_0(\theta_S)}{\int \mathcal{L}(\theta_S|\mathbf{D}_S)^\gamma \pi_0(\theta_S)d\theta_S}d\theta_S \times p(\gamma|\tau)p(\tau) \tag{18}$$

where $\pi_0(\theta_S)$ is an initial prior for θ_S. Hobbs et al. (2011) chose the following distributions:

$$\theta_T|\theta_S, \tau \sim \mathcal{N}\left(\theta_S, \frac{1}{\tau}\right),$$

and

$$\gamma|\tau \sim Beta(g(\tau), 1),$$

where $g(\tau)$ is a positive function of τ that is small for τ closed to zero and large for large values of τ. When the evidence for commensurability is weak, τ is forced toward zero, increasing the variance of the commensurate prior for θ_T. So the amount of borrowing can be adapted in two ways: through the power prior parameter, or through the commensurability parameter.

Below is a detailed derivation of this prior:

$$\begin{aligned}\pi(\theta_T, \gamma, \tau|\mathbf{D}_S) &= \pi(\theta_T|\gamma, \tau, \mathbf{D}_S)\pi(\gamma, \tau|\mathbf{D}_S) \\ &= \int \pi(\theta_T|\gamma, \tau, \theta_S, \mathbf{D}_S)\pi(\theta_S, \gamma, \tau|\mathbf{D}_S)d\theta_S \times \pi(\gamma, \tau|\mathbf{D}_S)\end{aligned} \tag{19}$$

We note that :

$$\begin{aligned}\pi(\gamma, \tau|\mathbf{D}_S) &= \pi(\gamma|\mathbf{D}_S, \tau)\pi(\tau|\mathbf{D}_S) \\ &= \pi(\gamma|\tau)\pi(\tau),\end{aligned} \tag{20}$$

and:

$$\pi(\theta_T|\gamma, \tau, \theta_S, \mathbf{D}_S) = \pi(\theta_T|\tau, \theta_S).$$

Therefore:

$$\pi(\theta_T, \gamma, \tau|\mathbf{D}_S) = \int \pi(\theta_T|\theta_S, \tau)\pi(\theta_S|\gamma, \mathbf{D}_S)d\theta_S \times p(\gamma|\tau)p(\tau) \tag{21}$$

The term $\pi(\theta_S|\gamma, \mathbf{D}_S)$ corresponds to a normalized power prior, so that :

$$\pi(\theta_T, \gamma, \tau|\mathbf{D}_S) = \int \pi(\theta_T|\theta_S, \tau)\frac{\mathcal{L}(\theta_S|\mathbf{D}_S)^\gamma \pi_0(\theta_S)}{\int \mathcal{L}(\theta_S|\mathbf{D}_S)^\gamma \pi_0(\theta_S)d\theta_S}d\theta_S \times p(\gamma|\tau)p(\tau) \tag{22}$$

Note that in equation (4) of Hobbs et al. (2011), the prior $\pi_0(\theta_S)$ is omitted, and $d\theta_S$ is misplaced.

In the Gaussian likelihood case, the "location commensurate power prior" is given by (Hobbs et al. 2011) :

$$p\left(\theta_T, \gamma, \tau \mid \mathbf{D}_S\right) \propto \mathcal{N}\left(\theta_T \mid \hat{\theta}_S, \frac{1}{\tau} + \frac{\hat{\sigma}_S^2}{\gamma N_S}\right) \times \text{Beta}\left(\gamma \mid g(\tau), 1\right) \times \pi(\tau) \quad ,$$

where $\hat{\sigma}_S$ is the standard deviation observed in the source study. The posterior is therefore:

$$p\left(\theta_T \mid \mathbf{D}_S, \mathbf{D}_T, \gamma, \tau\right) \propto \mathcal{N}\left(\theta_T \,\middle|\, \frac{\gamma N_S \tau \sigma_T^2 \hat{\theta}_S + N_T u \hat{\theta}_T}{\gamma N_S \tau \sigma_T^2 + N_T u}, \frac{u\sigma^2}{\gamma N_S \tau \sigma_T^2 + N_T u}\right),$$

where σ_T is the standard deviation in the target study, which is assumed known, and $u = \gamma N_S + \hat{\sigma}_S^2 \tau$. Moreover:

$$\begin{aligned}p\left(\gamma, \tau \mid \mathbf{D}_S, \mathbf{D}_T, \sigma^2\right) \propto &\mathcal{N}\left(\theta_T - \hat{\theta}_S \mid 0, \frac{\sigma^2}{N_T} + \frac{1}{\tau} + \frac{\hat{\sigma}_S^2}{\gamma N_S}\right) \\ &\times \text{Beta}\left(\gamma \mid g(\tau), 1\right) \times \pi(\tau).\end{aligned}$$

5.7.2 Parameters to be varied

Hobbs et al. (2011) considered the case of Gaussian likelihoods. They chose $g(\log(\tau)) = \max(\log(\tau), 1)$ and put a flat tails Cauchy(0, 30) prior on $\log(\tau)$. For the choice of priors on τ, see Table 5. So the study investigated how the prior on τ influences inference on γ and θ_T in the various scenarios.

5.7.3 Implementation

We used a custom implementation in Stan. Generalizing the Normalized Power Prior to borrow treatment effect in the Aprepitant case study is not straightforward. Therefore, in this case, we assumed a normal likelihood.

5.8 Robust Mixture Prior

5.8.1 Method description

Schmidli et al. (2014), followed by Röver et al. (2019), and based on earlier work by Greenhouse and Waserman (1995), proposed the use of a mixture prior in order to adapt the amount of borrowing while making the analysis more robust to prior-data conflict:

$$\begin{aligned}\pi(\theta_T|\mathbf{D}_S = d_S) = &w\pi(\theta_T|M_{\text{source}}, \mathbf{D}_S = d_S) \\ &+ (1-w)\pi(\theta_T|M_{\text{weak}}, \mathbf{D}_S = d_S),\end{aligned} \tag{23}$$

where M_{source} is a model corresponding to either consistency, subject-level exchangeability, or study-level exchangeability. The weight w corresponds to $\Pr(M_{\text{source}}|\mathbf{D}_S)$, the prior belief corresponding to this model. By contrast, M_{weak} is an alternative model corresponding to unrelated treatment effects in the source and target studies. Each component in the mixture corresponds to a different assumption about the relationship

between studies: $\pi(\theta_T|M_{\text{source}}, \mathbf{D}_S)$ corresponds to an informative component based on the assumption that studies are related, whereas $\pi(\theta_T|M_{\text{weak}}, \mathbf{D}_S)$ is typically a vague component. The posterior distribution from the source study was used as the informative component $\pi(\theta_T|M_{\text{source}}, \mathbf{D}_S)$.

The posterior distribution of the target study treatment effect θ_T is a weighted average of the posterior distributions under each model, weighted by their respective posterior model probabilities:

$$\begin{aligned}\pi(\theta_T \mid \mathbf{D}_T = d_T, \mathbf{D}_S = d_S) = \tilde{w}\pi\left(\theta_T \mid M_{\text{source}}, \mathbf{D}_T = d_T, \mathbf{D}_S = d_S\right) \\ + (1-\tilde{w})\pi\left(\theta_T \mid M_{\text{weak}}, \mathbf{D}_T = d_T, \mathbf{D}_S = d_S\right),\end{aligned} \tag{24}$$

where the updated weight $\tilde{w}$ corresponds to the posterior $Pr\left(M_{\text{source}} \mid \mathbf{D}_T = d_T, \mathbf{D}_S = d_S\right)$.

So the mixture introduces robustness by allowing the vague prior to dominate if the heterogeneity between source and target trials is large compared to within-trial variance.

As recommended by Schmidli et al. (2014), we selected the variance of the vague component so that it corresponds to a unit-information prior. More precisely, the variance of the vague component is such that corresponds to the information brought by one subject per arm in the target study. Note that, given that the variance of the outcome in the target study is assumed equal to the empirical variance, setting this vague component in the prior corresponds to a form of empirical Bayes. For normally distributed treatment effects, the vague component variance is set to $N_T\sigma^2_{\theta_T}$, where σ_{θ_T} is the standard error on the treatment effect obtained from the target data alone.

5.8.2 Parameters to be varied

The parameter that determines the amount of borrowing is the prior mixture weight w. A grid of values ranging from 0 to 1 in steps of 0.1 would be considered.

5.8.3 Implementation

In the case of a normal likelihood, we used the RBesT package. In the Aprepitant case, we relied on a custom implementation using Stan.

6 Prior Effective Sample Size

The amount of borrowing is most easily measured using the concept of prior effective sample size (ESS). Prior ESS corresponds to the number of pseudo-observations required to update a vague conjugate prior to the prior of interest (viewed as the posterior from previous analysis). For instance, in a beta-binomial model, the parameters of the $Beta(a, b)$ prior can be interpreted as the posterior obtained after observing a successes and b failures, starting from a vague Beta prior (with a and b arbitrarily small). Similarly, a normal prior with variance σ^2/n corresponds to a prior ESS of n, starting from a normal prior with variance σ^2. However, the prior ESS is not clearly defined for non-conjugate priors.

Neuenschwander et al. (2020) introduced an information-based ESS (in the context of one-dimensional parameters), the expected local-information-ratio (ELIR), which has the property of being "predictively consistent", meaning that the expected posterior predictive ESS for a sample of size N_T is equal to the sum of the prior ESS and N_T. The ELIR is defined as follows:

$$ELIR = \mathbb{E}_\theta\left[\frac{\mathcal{I}_\pi(\theta)}{\mathcal{I}_1(\theta)}\right] \tag{25}$$

where $\mathcal{I}_1(\theta)$ is the expected Fisher information for one information unit, given by:

$$\mathcal{I}_1(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathscr{L}(\theta|\mathbf{D}_1)}{\partial\theta^2}\middle|\theta\right], \tag{26}$$

and $\mathbf{D}_1$ denotes a dataset with one subject per arm. The Fisher prior information $\mathcal{I}_\pi(\theta)$ is given by:

$$\mathcal{I}_\pi(\theta) = -\frac{\partial^2 \log p(\theta)}{\partial\theta^2}. \tag{27}$$

Importantly, the ELIR is predictively consistent and thus correctly quantifies the amount of information as an equivalent number of observations.

7 Bayesian operating characteristics

The motivations for using Bayesian equivalents of type 1 error and power are well summarized and discussed in Best et al. (2023). The main argument for considering these operating characteristics is that Bayesian metrics should be used for Bayesian designs, with a willingness to adopt the Bayesian approach also for assessing the risk of, e.g., declaring a treatment as effective which in reality is ineffective.

7.1 Background on Bayesian equivalents of type 1 error rate and power

All metrics related to type 1 error rate and power are a special case of the success decision criterion rate (Psioda and Ibrahim 2019; Best et al. 2023):

$$\begin{aligned} r(\theta_T|d_S) &= \mathbb{E}_{p_d(\theta_T)}[\varphi_B(\mathbf{D}_T|\mathbf{D}_S = d_S)|\theta_T] \\ &= \int \Pr(\text{ Study success } \mid \theta_T, \mathbf{D}_S = d_S) p_d(\theta_T) d\theta_T, \end{aligned} \tag{28}$$

where φ is the decision function, which equals 1 when the null hypothesis is rejected, and 0 otherwise; and where the conditional power is defined as:

$$\begin{aligned} \text{CP}(\theta_T|d_S) &= \Pr(\text{Study success } \mid \theta_T, \mathbf{D}_S = d_S) \\ &= \int \mathbb{I}\left\{\Pr\left(\theta_T > \theta_0 \mid \mathbf{D}_T, \mathbf{D}_S = d_S\right) \geq \eta\right\} p\left(\mathbf{D}_T \mid \theta_T\right) d\mathbf{D}_T, \end{aligned} \tag{29}$$

and $p_d(\theta_T)$ is the so-called design (or sampling) prior. Note that this prior typically depends on the source study data. Indeed, the prior on the treatment effect used when computing the conditional power in 29 is the prior that is used at the analysis stage, which is why it is called the analysis prior (also referred to as the fitting prior). By contrast, the design prior, used when computing the expected conditional power in 28, can be distinct.

7.2 Computation of Bayesian Operating Characteristics

To compute the average type 1 error rate, the pre-posterior probability of false (or true) positive, the prior probability of study success and the average power, we considered two alternatives. A first possibility is to rely on nested Monte Carlo integration, by sampling N_θ values of θ_T from the design prior, and estimate $\Pr(\text{Study success}|\theta_T)$ based on N_R replicates, and finally compute the average probability of success. However, this approach bears a high computational cost, due to nested integration.

Another approach is to estimate the probability of success evenly on a given range, then compute the Bayesian Operating Characteristics by integrating $\Pr(\text{Study success}|\theta_T)p_d(\theta_T)$ using Simpson's rule. This approach is computationally beneficial, as it does not require recomputing the probability of success for each design prior and, allows reusing results obtained from the computation of frequentist OCs. In pilot comparisons between an MC-based estimation of Bayesian OCs and a method based on deterministic integration, we observed that deterministic integration led to an accurate estimate of Bayesian OCs, despite the small number of values of θ_T considered (from 25 to 50). This is probably due to the smoothness of the integrands. Note that this method implies a limited integration range, and therefore implicitly assumes that the integrand goes to zero for extreme values of the chosen range. The integration range corresponded to the range of treatment effects considered in the definition of the scenarios.

Note that it is not possible to compute Bayesian Operating Characteristics with an analysis design prior for methods that use empirical Bayes. Indeed, in this case, it is not possible to sample from the prior before having access to the data. However, these data are themselves sampled from the prior. Therefore, for several methods, we only reported Bayesian Operating Characteristics obtained with a unit-information design prior or with the source posterior as design prior. However, Bayesian OCs obtained with such priors usually provide bounds for those obtained with an analysis prior.

7.3 Bayesian type 1 error rate and power

Ibrahim et al. (2012), Chen et al. (2014), and Psioda and Ibrahim (2019) suggested an approach in which the usual frequentist type 1 error is integrated with respect to a null design prior distribution for the treatment effect. They define the Bayesian type 1 error rate (or average type 1 error rate) as:

$$\mathbb{E}_{p_{\text{null}}(\theta_T)}\left[r(\theta_T|d_S)\right], \tag{30}$$

where $p_{\text{null}}(\theta_T)$ denotes the null design prior.

Metric	**Design prior**		
Average TIE	Truncated analysis prior	Truncated UI prior	Truncated source posterior
Prior proba. of no treatment benefit	Analysis prior	UI prior	Source posterior
Pre-posterior proba. of FP			
Upper bound on the proba. of FP			

Table 6: Summary of design priors used to compute Bayesian OCs related to type I error.

Metric	**Design prior**		
Average power	Truncated analysis prior	Truncated UI prior	Truncated source posterior
Prior probability of study success	Analysis prior	UI prior	Source posterior
Pre-posterior proba. of FP			

Table 7: Summary of design priors used to compute Bayesian OCs related to power.

Metric	**Definition**
Average TIE	$\int Pr(\text{Study success}\|\theta_T)\frac{\pi(\theta_T\|\mathbf{D}_S=d_S)\mathbb{I}\{\theta_T\leq\theta_0\}}{Pr(\theta_T\leq\theta_0)}d\theta_T$
Prior proba. of no treatment benefit	$Pr(\theta_T \leq \theta_0)$
Pre-posterior proba. of false positive	$Pr(\text{Study success}, \theta_T \leq \theta_0) = \int_{\theta_T\leq\theta_0} Pr(\text{Study success}\|\theta_T)p_d(\theta_T)d\theta_T$
Upper bound on the proba. of false positive	$Pr(\text{Study success}\|\theta_T = \theta_0) \times Pr(\theta_T \leq \theta_0)$

Table 8: Summary of Bayesian OCs related to type I error.

Metric	**Definition**
Average power	$\int Pr(\text{Study success}\|\theta_T)\frac{\pi(\theta_T\|\mathbf{D}_S=d_S)\mathbb{I}\{\theta_T>\theta_0\}}{Pr(\theta_T>\theta_0)}d\theta_T$
Prior probability of study success	$Pr(\text{Study success}) = \int Pr(\text{Study success}\|\theta_T)p_d(\theta_T)d\theta_T$
Pre-posterior probability of true positive	$Pr(\text{Study success}, \theta_T > \theta_0) = \int_{\theta_T>\theta_0} Pr(\text{Study success}\|\theta_T)p_d(\theta_T)d\theta_T$

Table 9: Summary of Bayesian OCs related to power.

Similarly, the Bayesian power (or average power) is defined by Psioda and Ibrahim (2019) as :

$$\mathbb{E}_{p_{\text{alt}}(\theta_T)}\left[r(\theta_T|d_S)\right], \tag{31}$$

where $p_{\text{alt}}(\theta_T)$ denotes the alternative design prior. Note that this is related to the concept of probability of success (POS) (also called assurance, O'Hagan et al. (2005) and Chuang-Stein and Kirby (2017)), but POS is computed with equal analysis and design priors.

When computing the average type 1 error rate (resp. power), Psioda and Ibrahim (2019) suggest that a logical choice for the design prior is the normalized analysis prior truncated on the range of values for the treatment effect that are consistent with the null (resp. the alternative). They define the default design priors as :

$$p_d^i(\theta_T) = \pi(\theta_T|\mathbf{D}_S = d_S, \theta_T \in \Theta_i), i \in \{0, 1\}$$

For example, the null design prior $p_{\text{null}}(\theta_T)$ corresponds to the analysis prior distribution of θ_T, $\pi(\theta_T|\mathbf{D}_S = d_S)$, assuming H_0:

$$p_{\text{null}}(\theta_T) = \frac{\pi(\theta_T|\mathbf{D}_S = d_S)\mathbb{I}\{\theta_T \leq \theta_0\}}{\Pr(\theta_T \leq \theta_0)},$$

Thus, the sampling priors under the null and alternative hypotheses arise by truncation of the prior elicited from the source study and subsequent normalization. We followed this approach when choosing the design priors.

Additionally, we used the two following design priors:

- "Truncated source posterior", chosen to be the normalized truncated lower tail ($\leq \theta_0$) of the posterior from the source studies under an initial improper prior.
- "Truncated UI prior", chosen to be the normalized truncated lower tail of a unit-information prior based on source data, centered on θ_0. Concretely, if we assume a normal likelihood, the UI design prior is $\mathcal{N}(\hat{\theta}_S, s_S^2)$, where s_S^2 is the sample variance in the source study. To define a unit-information prior in the Aprepitant case study. We used the following approach : we performed a separate analysis of data similar to the source data, but with a treatment effect estimate of 0. We then approximated the corresponding posterior using a mixture of Beta distributions. We computed the corresponding ESS using RBesT's moments matching method and scaled the parameters of the Beta components by dividing them by the ESS.

These two design priors gave us an estimate of the range of values that the Bayesian OCs can take for a given analysis prior, from a skeptical design prior (the truncated UI prior) to an optimistic prior (the truncated source posterior).

In the Aprepitant case study, the UI prior is defined using the following approach: we perform inference on the target data using a separate analysis. We then approximate the posterior using a mixture of beta distributions thanks to RBesT. We compute the moment-based ESS of this mixture approximation and scale the coefficients a and b of each comment by dividing them by the ESS. This ensures that the resulting ESS of the mixture is 1. In the Aprepitant case, the source posterior, used as a design prior, is defined using a Beta Binomial model starting from a flat prior on each rate. That is, by noting $N_S^{(a)}$ the number of participants in arm a of the source study, and $n_S^{(a)}$ the number of successes in arm a of the source study:

$$\pi(\theta_T) = \int_{\max(-\theta_T,0)}^{\min(1,1-\theta_T)} Beta\left(p_T^{(c)} + \theta_T|1 + n_S^{(t)}, 1 + N_S^{(c)} N_{failures}^{(t)}\right) Beta\left(p_T^{(c)}|1 + N_{successes}^{(c)}, 1 + N_{failures}^{(c)}\right) dp_T^{(c)} \tag{32}$$

7.4 Pre-posterior probability of a false positive result

Best et al. (2023) suggested using an alternative metric to address "the inconsistency between the prior information and the null treatment effect by explicitly accounting for the probability that the treatment effect

is null or harmful under a suitably-chosen design prior":

$$\begin{aligned}\widetilde{m}\left(CP(\theta_T), \Pr\left(\theta_T \leq \theta_0\right), p(\theta_T)\right) &= \underbrace{m\left(CP(\theta_T), p_{\text{null}}\left(\theta_T\right)\right)}_{\text{Average type 1 error}} \times \underbrace{\Pr\left(\theta_T \leq \theta_0\right)}_{\text{Prob treatment effect is null/harmful}} \\ &= \int \Pr(\text{ Study success } \mid \theta_T) \frac{\pi(\theta_T|\mathbf{D}_S = d_S)\mathbb{I}\left\{\theta_T \leq \theta_0\right\}}{\Pr\left(\theta_T \leq \theta_0\right)} d\theta_T \times \Pr\left(\theta_T \leq \theta_0\right) \\ &= \int \Pr(\text{ Study success } \mid \theta_T)\pi(\theta_T|\mathbf{D}_S = d_S)\mathbb{I}\left\{\theta_T \leq \theta_0\right\} d\theta_T \\ &= \int_{\theta_T \leq \theta_0} \Pr(\text{ Study success } \mid \theta_T)\pi(\theta_T|\mathbf{D}_S = d_S) d\theta_T\end{aligned}$$

This is the average type 1 error rate (with respect to the null design prior) multiplied by the prior probability (under the corresponding untruncated version of the design prior) of the treatment effect being null or harmful. It is equivalent to the joint probability of the true treatment effect being null or harmful and the study being declared a success It is sometimes referred to as "type III error" of actually drawing a false positive conclusion (Spiegelhalter and Freedman 1986), or pre-posterior probability of a false positive result. We reported this metric, as well as the corresponding pre-posterior probability of a true positive (see definition in Table 9, and corresponding design priors in Table 7).

Other metrics important at the design stage are the prior probability of no treatment benefit, the prior probability of study success (which, from a Bayesian point of view, can be seen as a prior predictive probability), and upper bound on the pre-posterior probability of FP (see definitions in Table 8 and corresponding design priors in Table 6).

8 Major deviations from the protocol

8.1 Aprepitant case study

Initially, the approach envisioned for the Aprepitant case study (in which the treatment effect is a difference in proportions), was to put beta priors on the proportion of responders in each arm, and define the treatment effect as the difference in these proportions. However, with this approach, we do not directly define a prior on the treatment effect, but indirectly through the response rates in each arm. We instead followed an approach initially described in Jin and Yin (2021), in which a prior is put on the target study control rate (such as a beta prior or a uniform prior in the [0,1] range), and a prior is put on the target study treatment effect (such as a truncated normal). The target study response rate in the treatment arm is the difference between the treatment effect and the rate in the control arm.

8.2 Prior Effective Sample Size

To introduce the concept of prior Effective Sample Size (ESS), consider a normal prior on the parameter of interest, $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. The likelihood is $p(\mathbf{D}|\mu, \sigma) = \prod_{i=1}^{n} \phi(x_i|\mu, \sigma^2)$, where σ is known, and $\phi(x|\mu, \sigma^2)$ is the normal probability density function with mean μ and variance σ^2. We denote σ the sampling standard deviation (it is denoted as the "reference scale" in the RBesT package). The posterior after observing n data points is $p(\mu|\mathbf{D}, \sigma) = \mathcal{N}(\mu|\mu_n, \sigma_n^2)$, with $\sigma_n^2 = (n/\sigma^2 + 1/\sigma_0^2)^{-1}$, and: $\mu_n = n\frac{\sigma_n^2}{\sigma^2}\overline{x} + \frac{\sigma_n^2}{\sigma_0^2}\mu_0$. The prior ESS, is a measure of the informativeness of the prior distribution in terms of number of samples. If we start with a noninformative prior ($\sigma_0 \to +\infty$), the effective sample size of the posterior distribution is n, and the effective sample size of the prior is 0. The variance of the posterior will be σ^2/n in this case. Therefore, in the more general case where the prior has variance σ_0^2, the prior ESS is σ^2/σ_0^2.

So, if we have some distribution $\mathcal{N}(\mu, \tau^2)$, by assuming this distribution corresponds to the posterior derived from an uninformative prior updated after observing m data points sampled from a normal distribution with known standard deviation σ (the reference scale), we have that : $\tau^2 = \sigma^2/m$. So, we have an ESS : $m = \frac{\sigma^2}{\tau^2}$. What should be the value of σ? In our simulation study, when inferring the treatment effect from data, we assumed that the sampling standard deviation σ for the target study data is known and corresponds to the target study data sample standard deviation. Therefore, we set $\sigma = s_T$, where s_T is the target study sample standard deviation.

In our study, we estimated the prior ESS by computing the ESS of the posterior distribution, and subtracting the target study sample size per arm. To compute the ESS of the posterior distribution, we used the RBesT package in a three steps procedure: first, we sampled 1000 samples from the posterior distribution,

then we approximated the posterior based on these samples using a mixture of normal distributions (or a mixture of beta distributions, in the Aprepitant case study), then, we computed the ESS of the corresponding mixture approximation. The goal of using a mixture approximation is solely to be able to use RBesT.

Moments-based posterior ESS We used RBesT functionalities to compute the moment-based ESS of the mixture approximation. The moment-based matching method used in RBesT is the following:

1. Compute the moments of the distribution of interest.
2. Define a distribution from a family for which computing the ESS is trivial (such as normal, beta, or gamma) with the same moments.
3. Compute the corresponding ESS, which is an approximation to the ESS of the distribution of interest.

For example, in the case of a Gaussian posterior with posterior variance $\sigma^2_{(p),i}$ at iteration i, and sample standard deviation σ_i , the moment-based ESS is $\sigma_i^2/\sigma_{(p),i} - N_{T/2}$. In the Aprepitant case study, before approximating the distribution with a mixture, we linearly transformed the samples so that they fit in the $[0, 1]$ range instead of the $[-1, 1]$ range: we transformed each sample x into $(x + 1)/2$. Indeed, there is no standard distribution with support $[-1, 1]$.

Precision-based posterior ESS The precision-based matching method, inspired from the moment-based matching method, proceeds as follows:

1. Compute the mean of the distribution of interest, and the half-width of the 95% credible interval (CrI).
2. Define a distribution from a family for which computing the ESS is trivial (such as normal, beta, or gamma) with the same precision. Note that this may not be sufficient to uniquely define a matching distribution (for example, in the Gaussian case, any translation of this distribution would have the same precision). Therefore, it may also be required to match the mean for matching distributions with two degrees of freedom (which is the case of the normal, beta, and gamma distributions).
3. Compute the corresponding ESS, which is an approximation to the ESS of the distribution of interest.

When the summary measure is assumed normally distributed, it makes sense to match the posterior distribution of the treatment effect with a Gaussian distribution. Consider that the variance of the posterior distribution over the treatment effect is τ^2, and denote ρ the half-width of the 95% CrI. For a Gaussian with variance τ^2, the half-width of the 95% CrI is given by $\Phi^{-1}(1 - \alpha/2)\tau$, with $\alpha = 0.05$. Therefore, if we match any distribution with a Gaussian with the same mean and half-width of the 95% CrI, the matching distribution will have standard deviation $\tau = \frac{\rho}{\Phi^{-1}(1-\alpha/2)}$. Therefore, we can simply reuse the moment-based matching code in RBesT by replacing the standard deviation of the matching distribution with $\tau = \frac{\rho}{\Phi^{-1}(1-\alpha/2)}$.

When the summary measure likelihood is the model structure in Figure 2 (Aprepitant case), we could match the posterior distribution over the treatment effect with a linearly transformed gamma or beta distribution, however, there is no analytical formula for the precision in the case of a beta distribution or a gamma distribution. Therefore, we could match the precision and mean of the distribution of interest and the transformed gamma/beta by numerically solving a minimization problem (note that this is a 1D minimization problem, as we can easily match the mean). We opted for a simpler approx, by matching the posterior distribution of the treatment effect with a Gaussian. This method is suitable in cases where the posterior distribution is approximately Gaussian and the 95% high density interval of the matching gaussian is within the $[-1, 1]$ range.

8.3 Commensurate Power Prior

The commensurate power prior was implemented for a variety of priors on the heterogeneity parameter. However, preliminary tests showed that a Cauchy prior on log(heterogeneity) could lead to divergence issues. Reducing the scale parameter from 30 to 10 led to relatively similar priors with less divergences.

8.4 Maximum type 1 error rate

The tender required to report the "maximum type 1 error in a pre-specified [drift] range". Given that there is only one value for the type 1 error rate (for $\delta = \theta_0 - \hat{\theta}_S$), we do not report the "maximum type 1 error rate".

8.5 PDCCPP

In order to apply the PDCCPP method, we reused the code provided in Nikolakopoulos et al. (2018). This required minor adaptation to work for treatment effect extrapolation, instead of control arm borrowing. In many scenarios, we found that the method encountered convergence issues when inferring the calibration parameter c. It was not possible to determine why this convergence issue occurred, and in which range of scenarios. Therefore, we did not include this method in the final simulaton study.

8.6 Elastic prior

The elastic prior implies running a simulation study for a specific method, and then calibrating the amount of borrowing based on the results of the simulation. This calibration aims at maximizing some objective function that implements the trade-off between an increase in power in the "congruent" case and an increase in type 1 error in the incongruent case. This calibration procedure is not specific of the method and could be applied to all partial extrapolation methods that include a parameter determining the borrowing strength (Jiang et al. 2021). However, the code to maximize this objective function is not provided in the original paper's supplementary material. Rather, the maxima are hard-coded for each example. This makes it impractical to adapt the code to our specific use cases. Moreover, the calibration step implies a significant computational burden, making it impractical to investigate our method in our simulation study, which involves a very large number of scenarios and replicates. Therefore, we finally did not include this method in our simulation study.

8.7 Configurations

The configuration originally specified in the protocol, with 10,000 replicates per simulation, implied huge computational costs. We adapted the number of drift values considered, the number of replicates for two NPP case studies (Mepolizumab and Teriflunomide), and the number of scenarios (sample size factors, target-to-source standard deviation ratio), depending on the case study and the method. We made sure that the reduced configuration still allowed reliable interpretation of the results, in particular regarding Monte Carlo uncertainty associated with the estimated operating characteristics. Concretely, we first used a light configuration to cover a large number of drift values (see Table 10). Then, for the three main treatment effect values (no effect, half effect, and same effect as in the source population), we additionally considered a configuration with more replicates (see Table 11).

The original configuration is provided, for transparency, in table 12

9 Simulation study implementation

9.1 Code availability

The code developed in this study is packaged and can be used for running and analyzing simulation studies and for analyzing data using partial extrapolation. It is available as a GitHub repository at `https://github.com/quinten-health-os/BayesianExtrapolationSimulation`. The exact version that was used for running the simulation study is v0.0.2, whereas the version that was used for the analysis of the results, results quality checks, and for producing tables and figures is v0.0.3 (no changes were introduced in the simulation part between v0.0.2 and v0.0.3). The code is precisely documented and version-controlled. In particular, the main components of the code are illustrated in a series of vignettes. The `README.md` file provides instructions for accessing the documentation, including functions references and vignettes, via a web browser. Additionally, a reference manual is provided in pdf format. Finally, should the code be reused in future projects, a `TODO.md` file contains a list of suggested improvements for overall quality and reusability.

9.2 Rationale of the design

The core of the code implements the inference logic: a source data class represents source data, which are observed. Target data are represented in a different class. They different as we can sample target data replicates. When initializing a model, source data are provided to it, as well as methods parameters and MCMC configuration (if applicable), so as to define a prior. Inference is performed by using the inference method on target data, which updates properties of the model object with moments of the treatment effect posterior distribution and, if applicable, computes the posterior distribution of borrowing parameters. Note that inference proceeds in two steps: first, if the method uses empirical Bayes, prior parameters are updated based on the target data sample, second Bayes" rule is applied.

Method	**Case**	**Likelihood**	**# replicates**	**# drift values**	N_S/N_T	**Denom. change factor**	σ_T/σ_S
EBPP	Botox/Dapagliflozin	Normal	5000	33	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	5000	33	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Normal	5000	33	1, 2, 4, 6	NA	NA
NPP	Botox/Dapagliflozin	Normal	1000	23	2, 4	NA	1
	Belimumab/Mepolizumab/Teriflunomide	Normal	1000	23	2, 4	1	NA
	Aprepitant	Normal	1000	23	2, 4	NA	NA
Comm. PP	Botox/Dapagliflozin	Normal	1000	23	2, 4	NA	1
	Belimumab/Mepolizumab/Teriflunomide	Normal	1000	23	2, 4	1	NA
	Aprepitant	Normal	1000	23	2, 4	NA	NA
Others	Botox/Dapagliflozin	Normal	5000	33	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	5000	33	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Binomials	1000	23	2, 4	NA	NA

Table 10: Light configuration used in the simulation study. Other methods include separate analysis, pooling, RMP, CPP, and Test-then-Pool (equivalence test or difference test).

Method	Case	Likelihood	# replicates	# drift values	N_S/N_T	Denom. change factor	σ_T/σ_S
EBPP	Botox/Dapagliflozin	Normal	10000	3	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	10000	3	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Normal	10000	33	1, 2, 4, 6	NA	NA
NPP	Botox/Dapagliflozin	Normal	10000	3	2, 4	NA	1
	Belimumab	Normal	10000	3	2, 4	1	NA
	Mepolizumab/Teriflunomide	Normal	8000	3	2, 4	1	NA
	Aprepitant	Normal	10000	3	2, 4	NA	NA
Comm. PP	Botox/Dapagliflozin	Normal	10000	3	4	NA	1
	Belimumab/Mepolizumab/Teriflunomide	Normal	10000	3	4	1	NA
	Aprepitant	Normal	10000	3	4	NA	NA
Others	Botox/Dapagliflozin	Normal	10000	3	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	10000	3	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Binomials	10000	3	4	NA	NA

Table 11: Compute-intensive configuration used in the simulation study for the three main treatment effect values. Other methods include separate analysis, pooling, RMP, CPP, and Test-then-Pool (equivalence test or difference test).

Method	Case	Likelihood	# replicates	# drift values	N_S/N_T	Denom. change factor	σ_T/σ_S
EBPP	Botox/Dapagliflozin	Normal	10,000	33	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	10,000	33	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Normal	10,000	33	1, 2, 4, 6	NA	NA
NPP	Botox/Dapagliflozin	Normal	10,000	33	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	10,000	33	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Normal	10,000	33	1, 2, 4, 6	NA	NA
Comm. PP	Botox/Dapagliflozin	Normal	10,000	33	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	10,000	33	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Normal	10,000	33	1, 2, 4, 6	NA	NA
Others	Botox/Dapagliflozin	Normal	10,000	33	1, 2, 4, 6	NA	1, 2
	Belimumab/Mepolizumab/Teriflunomide	Normal	10,000	33	1, 2, 4, 6	1/2, 1, 3/2	NA
	Aprepitant	Binomials	10,000	30	1, 2, 4, 6	NA	NA

Table 12: Configurations planned in the simulation study protocol. Other methods include separate analysis, pooling, RMP, CPP, PDCCPP, the Elastic Prior, and Test-then-Pool (equivalence test or difference test).

When launching a simulation study, all scenarios are generated based on the provided configuration files. These scenarios are then sequentially treated: source data and target data are initiated, and a model is defined. Target data replicates are then generated depending on the target data characteristics. The model is then fitted on each data replicate, and the corresponding inference metrics and frequentist metrics are computed. The data from the simulation are then analyzed: the sweet spot is computed for each metric of interest, and the Bayesian operating characteristics are estimated.

9.3 Use of existing code and packages

We tried, as far as possible, to reuse existing methods implementations. Not only did we need to perform inference with the method of interest, but, to compute the prior ESS using the different approaches we included (ELIR, difference between the moment-based or precision-based ESS of the posterior and the target study sample size per arm), we needed to sample from the prior and the posterior distribution. We identified several packages and code repositories that could potentially be used. The *RBesT* package (`https://opensource.nibr.com/RBesT/`) implements inference with conjugate mixture priors and Bayesian meta-analysis. It also allows for approximating distributions using mixture distributions, based on samples, and to compute ELIR and moment-based ESS. We therefore used this package for ESS computation. For Gaussian endpoints, we used *RBesT* for Separate Bayesian analysis and Pooling, as well as for the use of Gaussian Robust Mixture Priors.

The *NPP* package (`https://cran.r-project.org/web/packages/NPP/index.html`) contains an implementation of the Normalized Power Prior for normally distributed endpoints. However, it uses a custom MCMC implementation, whereas an analytical posterior is available in this case (see Section 5.4). Moreover, the *NPP* package requires individual-level data as input. These two elements prompted us to write a custom implementation for computational efficiency.

The *historicalborrow* package (`https://wlandau.github.io/historicalborrow/index.html`) is focused on control group borrowing. It includes hierarchical models such as the MAC model, as well as pooled and separate analysis. It also implements simulation routines, but with limited flexibility: it would not allow us to simulate time-to-event or recurrent event data. Therefore, we did not use this package.

The *PowerPriorVari* repository (`https://github.com/lxt3/PowerPriorVari`), which implements variations of the power prior to borrow from a single source study (Thompson et al. 2021) does not contain reusable code.

The *ESS* repository (`https://github.com/DKFZ-biostats/ESS`), contains implementations of different methods to estimate different prior ESS measures for a wide variety of models and is well documented. However, to use a limited number of packages, we only relied on *RBesT* for ESS computation.

psborrow2 (`https://genentech.github.io/psborrow2/index.html`) is an R package for conducting Bayesian dynamic borrowing analyses and simulation studies. However, it focuses on the borrowing of an external control arm and only implements the hierarchical commensurate prior. It was therefore of limited use for our study.

The *StudyPrior* mostly focuses on the case of binomial likelihood. For Gaussian likelihood, the package implements the Empirical Bayes Power Prior and the Normalized Power Prior. However, for the Empirical Bayes Power Prior, we used the implementation provided in Nikolakopoulos et al. (2018), and for the Normalized Power Prior, we used a custom implementation of the analytical posterior to avoid relying on computationally expensive approximations.

The *hdbayes* package was released in April 2024 (`https://github.com/ethan-alt/hdbayes`), well after the start of the implementation phase of the project. It implements a variety of Bayesian borrowing methods for generalized linear models. However, it always uses Stan for inference, whereas we were able to use analytical posteriors in several cases, which provided significant computational gains.

To summarize, we implemented the Gaussian Conditional Power Prior, which is straightforward, and validated the results by comparing them with Pooling and Separate analysis for a power parameter $\gamma = 1$. For PDCCPP, Test-then-Pool, and Empirical Bayes method, we used the code in Nikolakopoulos et al. (2018). We only renamed variables for consistency across the code base and split it into different functions. For the two variants of test-then-pool (with a difference or an equivalence test), depending on the result of the test, the Pooling or the Separate method (based on RBesT) was used. We also implemented the p-value based Power Prior, which inherits from the Gaussian Conditional Power Prior. For the Commensurate Power Prior, we also used a custom implementation in Stan. This was also motivated by the fact that to adapt existing models to the case where a binomial likelihood is used for each arm data, it would be much easier to adapt a working custom implementation. Indeed, to handle the model structure described in Figure 2, we had to rely on custom implementations in Stan.

9.4 Implementation and quality checks

We separately considered two aspects of quality control for the study: code quality and statistical accuracy. The code incorporates integration tests, which aim to ensure that code functionality is preserved across changes in the codebase. The version specifically used for running the simulation study was tagged (v0.0.2).

Moreover, the code was reviewed by a team member who did not directly participate in the implementation. The following check-list was provided to the reviewer :

- Simulation logic
  - Simulation scenarios
  - Definition of the source study parameters
  - Definition of the target study parameters
  - Data generation logic for continuous, binary, time-to-event, recurrent event endpoints
- Frequentist Operating Characteristics
  - Test decision
  - TIE, Power, 95% coverage, MSE, bias, precision
  - Associated MC error
- Inference metrics
  - Computation of inference metrics (posterior mean and credible interval)
- Bayesian Operating Characteristics
  - Definition of the design priors
  - Computation of the different Bayesian OCs
- Methods
  - Hierarchical dependencies between models classes
  - Priors definitions
  - Inference (posterior mean and variance)
- Analysis
  - Computation of power at equivalent type 1 error rate

Statistical accuracy of the results was validated based on manual checks, involving a comparison of the figures produced with relevant published figures. We initially intended to implement automatic tests comparing the results with relevant published tables. However, this proved to be difficult, due to differences in the settings considered (in particular, borrowing of control arm instead of treatment effect, which would require modification of the code to handle borrowing of control arm only), or lack of reproducibility of published results (which was confirmed using other packages such as RBesT).

To ensure the reproducibility of the study, the seed of the random number generator was set once at the same value when each scenario was simulated. The state of the random number generator was stored at the beginning and the end of each new simulation.

9.5 Deployment

Given the huge number of scenarios considered, and the large number of replicates, we required access to the French national supercomputer Adastra (National Computer Center for Higher Education). However, due to unknown reasons related to Stan, the post-processing of Markov chains was much longer than normal on this machine (of the tens of seconds). A similar issue arose on our local machines. We therefore attempted to use the containerization framework Singularity, as there exists Singularity container images that include *cmdstanr*. However, there were several issues with this solution. First, the available images run on the latest version of Ubuntu, which implies security issues which do not conform to the strict security requirements of high performance computing facilitites. Second, even within the Singularity container, we were not able to run inference with *cmdstanr*. However, cases that do not require *cmdstanr* can run within the container, and we provided the recipe file for building the Singularity container with the rest of the code, which provides very high reproducibility for our study. However, we managed to get Stan working on an AWS EC2 instance within a Docker container, although at a slower speed than what would have been achievable with Adastra. Therefore, all runs that required MCMC inference were performed on the AWS EC2 c5.12xlarge instance, while the remaining runs were launched on Adastra. Unfortunately, due to the difference between these machines, it would not be possible to make sensible comparisons between computation times. Although

these computation times are recorded in the results, we therefore will not present a comparison of the different methods' speed. Another issue we faced is that, when running in a Docker container on the AWS EC2 instance, the code would sometimes unexpectedly stop running without any error message being recorded in the logs. Despite our best efforts we were not able to identify the root cause of this issue. This incurred considerable time loss, and required automated checks to ensure simulations completeness and accurate concatenation of the different results files. While searching for a tradeoff between accuracy, computational time and time spent managing disparate results files, we decided to discard a Method/Case study result file as long as one scenario was missing and to rerun the corresponding Method/Case study with all scenarios. Moreover, an additional issue we encountered was the fact that tasks were limited to 24 hours on Adastra, which forced us to reduce some configurations so that run would be completed within this time window.

10 Application to real cases

The following sections give additional context on the selected studies for each type of endpoint. When a Bayesian analysis was published for a given study, we summarized the approach that was used.

10.1 Continuous endpoint

Botox

Case study: Introduced in Wang et al. (2022)

Endpoint: Score related to the disease

Summary measure: Difference in mean scores between the two arms (normally distributed)

Context: Published phase III placebo-controlled randomized paediatric clinical trial to evaluate the safety and efficacy of a single treatment of two doses (4 U/kg and 8 U/kg) of Botox with standardized physical therapy (PT) in paediatric patients with lower limb spasticity. Paediatric approval was based on this study.

The same product was previously approved in adults on the basis of a single-phase III placebo-controlled study in a similar indication. In the paediatric trial, 412 subjects 2 to 16 years and 11 months of age were randomized in a 1:1:1 ratio to the Botox 8 U/kg group, Botox 4 U/kg group, or control group. The full label information is available at `https://www.fda.gov/media/131444/download`. The original analyses for both the adult and paediatric trials were frequentist approaches.

Bayesian analysis: Wang et al. (2022) re-analyzed the primary efficacy endpoints using a Bayesian model. The approximate 95% CI for the treatment difference between Botox 4 U/kg group and control is (-0.10, 0.30) in the paediatric trial, which contains zero, i.e., not enough evidence to declare treatment superiority to control. Therefore, Wang et al. (2022) aimed at proposing an innovative Bayesian adaptive design to achieve treatment efficacy while maintaining good trial property.

For the case study, Wang et al. (2022) focused on the Bayesian analysis on two arms, the Botox 4 U/kg group and control group as the Botox 4 U/kg group was less efficacious. They studied a Bayesian adaptive design based on an informative prior (derived from adults data).

Dapagliflozin

Case study: None

Endpoint: Decrease in HbA(1c) from baseline (%).

Summary measure: Difference in mean decrease in HbAc between the two arms (%), from baseline to week 24/26 (normally distributed).

Context: Correction of hyperglycaemia and prevention of glucotoxicity are important objectives in the management of type 2 diabetes. Metformin is the regulatory-approved treatment of choice for most youth with type 2 diabetes early in the disease. However, metformin does not always provide adequate glycemic control, thereby necessitating add-on treatment.

However, the fact that metformin provides adequate control for many patients (up to 50% of patients in the Treatment Options for Type 2 Diabetes in Adolescents and Youth (TODAY) study), sizably diminishes the patient pool. In addition, insulin use has traditionally been an exclusion criterion, eliminating approximately

half of the paediatric patient population. Additional inclusion/exclusion criteria (e.g., required HbA1c range, major medical conditions, concomitant meds, and prior diabetes medication use) further shrink the available patient pool.

Dapagliflozin, a selective sodium-glucose cotransporter-2 inhibitor, reduces renal glucose reabsorption in an insulin-independent manner.

Shehadeh et al. (2023) report a 26-week, phase 3 trial with a 26-week extension among young patients (10 to 17 years of age) with uncontrolled type 2 diabetes (HbA(1c) 6.5 to 10.5%) receiving metformin, insulin, or both. Participants were randomly assigned 1:1:1 to 5 mg of dapagliflozin (N=81), 2.5 mg of saxagliptin (N=88), or placebo (N=76). Patients in active treatment groups with HbA(1c) $\geq$7% at week 12 were further randomly assigned 1:1 at week 14 to continue the dose or up-titrate to a higher dose (10 mg of dapagliflozin or 5 mg of saxagliptin). The primary endpoint was change in HbA(1c) at week 26. Analysis of the data demonstrated effectiveness of dapagliflozin.

Bailey et al. (2010) describe a phase 3, multicentre, double-blind, parallel-group, placebo-controlled trial, including 546 adults with type 2 diabetes who were receiving daily metformin ($\geq$1500 mg per day) and had inadequate glycaemic control (NCT00528879). They were randomly assigned to receive one of three doses of dapagliflozin (2.5 mg, n=137; 5 mg, n=137; or 10 mg, n=135) or placebo (n=137) orally once daily. The primary outcome was change from baseline in HbA(1c) at 24 weeks. For correspondence between the study in adults and paediatrics, we focused on the arms receiving 5mg daily dapagliflozin. 267 patients were included in analysis of the primary endpoint (dapagliflozin 5 mg, n=133; placebo, n=134). At week 24, mean HbA(1c) had decreased by -0.30% (95% CI -0.44 to -0.16) in the placebo group, compared with -0.70% (-0.85 to -0.56, $p < 0.0001$) in the dapagliflozin 5 mg group. We use the aggregate data from Bailey et al. (2013): the treatment effect is 0.36 (95% CI 0.16 to 0.56).

10.2 Binary endpoint

Belimumab

Case study: From Best et al. (2023) and Psioda and Xue (2020) (study how adult data could have been prospectively used in the design of the paediatric trial). The results of the PLUTO trial in paediatrics are reported in Brunner et al. (2020). Brunner et al. (2021) performed a comparison of studies, including PLUTO and trials in adults (BLISS).

Endpoint: SLE Responder Index

Summary measure: Log odds ratio for Benlysta compared to placebo (normal approximation)

Context: FDA approval of belimumab (Benlysta) IV formulation for use in paediatrics aged 5-17 years with active, seropositive lupus erythematosus (SLE).

Benlysta was approved by the FDA for adult patients with SLE in 2011. A paediatric post-marketing study was required and the applicant undertook to conduct a randomized, double-blind, placebo-controlled trial targeting to enroll 100 paediatric subjects 5 to 17 years of age with active systemic SLE. The paediatric study was not fully powered by design, efficacy was planned to be descriptive and no formal statistical hypothesis testing was proposed. The study was completed in 2018, with a total of 92 subjects.

Bayesian analysis: To facilitate the review of Benlysta, the FDA requested a post-hoc Bayesian analysis to further evaluate the efficacy of Benlysta in paediatric SLE patients by utilizing relevant information from the adult studies. The rationale was to provide more reliable efficacy estimates in the paediatric study in a setting where the clinical review team believed that the disease and patient response to treatment are likely to be similar between adults and paediatrics (see FDA's review https://www.fda.gov/media/127912/download for details). The analysis was based on the use of a robust mixture prior.

Evidence of efficacy has been established in adults in two independent pivotal Phase 3 trials, which are pooled and considered to be one single source of historical data. The primary endpoint was response at week 52 on the SLE responder index (SRI), and the summary measure of treatment effect was the odds ratio for Benlysta compared to placebo. The pooled odds ratio based on a total of $N_S = 1125$ subjects from these studies was 1.62 (95% CI 1.27 - 2.05), which on the log odds ratio scale corresponds to a point estimate of $y_S = 0.48$ with standard error of $s_S = 0.121$.

Aprepitant

Case study: From Jin et al. (2021)

Endpoint: No vomiting and no use of rescue therapy 0–24 h after surgery.

Summary measure: Difference in response rates between aprepitant and placebo.

Context: Aprepitant for the prevention of postoperative nausea and vomiting in paediatric subjects.

A multicenter, randomized, partially-blinded phase IIb study (Salman et al. 2019) evaluated the pharmacokinetics (PK)/pharmacodynamics, safety, and tolerability of aprepitant in paediatric subjects for the prevention of postoperative nausea and vomiting (PONV). Subjects aged birth to 17 years scheduled to undergo surgery and receive general anesthesia with $\geq$1 risk factor for PONV were randomly assigned to 1 of 3 aprepitant dose regimens (a single oral dose of aprepitant equivalent to adult doses of 10 mg, 40 mg, or 125 mg), or a control regimen of ondansetron before anesthesia. Assessments included PK, safety, and exploratory efficacy (complete response [CR; no emesis, retching, or dry heaves and no rescue therapy within 0-24 h following surgery] and no vomiting [NV; no emesis, retching, or dry heaves within 0-24 h following surgery]).

The difference in response rates in the treatment group and control group is 3.4% and the lower bound of the 95% CI is -11.2% . The study did not meet the non-inferiority criterion with margin -10% . This could be due to the fact that the study is not adequately powered (the post hoc power is 43%). An adult trial with sample size 293 and 280 in the treatment and control groups was completed before (Diemunsch et al. 2007), and the response was 63.0% in the treatment group, and 55.0% in the control group.

10.3 Time-to-event endpoint

Teriflunomide

Case study: Bovis et al. (2022)

Endpoint: Time to first relapse

Summary measure: Log hazard ratio for active treatment compared to placebo (normal approximation)

Context Multiple Sclerosis (MS) is rare in paediatrics. The Safety and Efficacy of Teriflunomide vs Placebo in paediatric Multiple Sclerosis (TERIKIDS) study (Chitnis et al. 2021) was a negative trial assessing teriflunomide in paediatrics (57 placebo vs 109 teriflunomide). The 34% reduction in the incidence of relapses observed in the teriflunomide treatment group of the TERIKIDS trial failed to achieve statistical significance. However, no compelling biological or clinical reasons indicate that evidence obtained in adults should be ignored when deciding treatment strategies for paediatric MS.

Bovis et al. (2022) applied a Bayesian approach for estimating the effect of teriflunomide in paediatrics in the TERIKIDS study , by integrating the available knowledge on teriflunomide in adults. As source studies, they used published data from 2 randomized clinical trials testing teriflunomide (14 mg) in adult patients with MS (TEMSO3: 363 placebo vs 359 teriflunomide, O'Connor et al. (2011); TOWER4: 389 placebo vs 372 teriflunomide, Confavreux et al. (2014)).

Bayesian analysis: They pooled hazard ratios (HRs) and 95% CIs on time-to-first relapse (log scale) by inverse of variance weighting. To account for differences between the adult and the paediatric populations and between some details of the study designs, the prior distributions were down-weighted by 50% or 75%. The log(HR) values were assumed to be normally distributed.

The observed HRs of teriflunomide on time-to-first relapse in TEMSO, TOWER, and in TERIKIDS were 0.72 (95% CI, 0.58-0.90), 0.63 (95% CI, 0.50-0.79), and 0.66 (95% CI, 0.39-1.11), respectively. The prior distribution obtained by pooling the results of the 2 trials in adults was centered at HR 0.68 (95% CI, 0.58-0.79).

10.4 Recurrent event endpoint

Mepolizumab Ortega et al. (2014) contains data for the placebo, Mepolizumab SC and Mepolizumab IV group (adults and adolescents are pooled). The adolescent group included 9 control patients (see page 95 of the EPAR), 16 received Mepolizumab IV or SC, (see page 95 of the EPAR), for a total of 25 patients. The total number of adult patients was 551, with 182 in the control group and 369 receiving Mepolizumab. According to Best et al. (2021), the log(RR) in adolescents is -0.40 with standard error 0.703, whereas the log(RR) in adults is -0.69, with standard error 0.13.

To determine the rate in the adult control group, we used the data from Ortega et al. (2014) we assumed that the effect of the paediatric subgroup in the overall rate computation is negligible, and therefore set the adult control rate equal to the overall control rate, 1.74. We then computed the rate in the treatment group so as to be consistent with the control rate and the log(RR) reported in Best et al. (2021), that is 0.87.

Case study: Described in detail in Best et al. (2021), based on a post hoc analysis of the MENSA trial of mepolizumab in severe asthma (Ortega et al. 2014) by Keene et al. (2020).

Endpoint: Rate of clinically significant exacerbations, analyzed with a negative binomial generalized linear model with a log link function.

Summary measure: Log event rate ratio obtained from negative binomial regression of the observed exacerbation counts (normal approximation) for active treatment compared to placebo

Context: MENSA was a randomised, placebo-controlled, double-blind, parallel group trial comparing mepolizumab 100 mg subcutaneous (SC) (n = 194) and mepolizumab 75 mg intravenous (IV) (n = 191) with placebo (n = 191), given every 4 weeks for 32 weeks in patients with severe asthma with an eosinophilic phenotype who had a history of at least two asthma exacerbations in the previous year while receiving treatment with high dose inhaled steroids and at least 3 months of treatment with an additional controller. The trial was funded by GlaxoSmithKline (ClinicalTrials.gov number: NCT01691521). The primary endpoint was the rate of clinically significant exacerbations per year, which were defined as worsening of asthma such that the treating physician elected to administer systemic steroids for at least 3 days or the patient visited an emergency department or was hospitalised. The trial included 25 adolescent (ages 12-17) and 551 adult subjects (aged $\geq$ 18). In the overall population, the trial showed strong evidence of a reduction in the rate of exacerbations.

Analysis was performed using a negative binomial generalised linear model with a log link function. The model included a categorical covariate for age group (12-17 years old, $\geq$ 18 years old) and the interaction of age group with treatment group, with additional adjustment for baseline covariates (oral corticosteroid use, region, exacerbations in the previous year and baseline % predicted FEV1).

The yearly rate of exacerbations was reduced by 47% (95% CI: 28-60) among patients receiving 75 mg IV mepolizumab and by 53% (95% CI: 36-65) among those receiving 100 mg SC mepolizumab, as compared with those receiving placebo. The two active treatment arms provided similar reductions in exacerbation rate compared to placebo and were therefore combined for the evaluation of subgroups; overall the reduction with the two active treatments combined was 50% (95% CI: 35-61).

Bayesian analysis: There was interest in assessing the treatment effect in adolescents but due to the low incidence of severe asthma with an eosinophilic phenotype in adolescents, the conduct of a separate study was considered impractical and there were insufficient adolescent subjects in the MENSA study to show statistical significance when this subgroup was analysed separately. A Bayesian dynamic borrowing approach based on a Robust Mixture Prior allowed assessment of the degree of belief needed in the relevance of the adult data to conclude that there was evidence of efficacy in the adolescent subgroup. Indeed, based on knowledge of the disease pathology in adults and adolescents and the mechanism of action of mepolizumab, there is a strong rationale to believe that efficacy in adolescents should be consistent with that in adults. To assess the sensitivity to the strength of prior belief in the consistency assumption, a tipping point analysis was carried out to identify how much prior weight needed to be placed on the adult prior component of the robust mixture prior in order for the posterior estimate of efficacy for adolescents to show evidence of treatment benefit.

11 Results

The raw results files are available as supplementary material. The tables containing raw results for frequentist OCs, Bayesian OCs derived from Simpson integration, and sweet spots, can be found in the Zenodo record 14780493. The numerous figures and tables produced are available in `Figures/` and `Tables/`. Here, we only present a small selection so as to illustrate our results.

In the figures legends, a "Consistent" effect means that the true treatment effect in the target study is equal to the observed treatment effect in the source study, that is, there is no drift. A "partially consistent" effect means that the true treatment effect in the target study is equal to the half of the observed treatment effect in the source study.

11.1 Inference

11.1.1 Convergence issues

We encountered severe convergence issues when applying the Commensurate Power Prior with a Cauchy hyperprior on the heterogeneity parameter, although decreasing the scale of the Cauchy distribution from 30 to 10 tends to reduce the fraction of cases in which this problem occurs. However, using a half-normal or an inverse gamma distribution led to successful convergence. With the Commensurate Power Prior, we found it useful to adaptively increase chain length, as in several cases the MCMC ESS was much smaller than the total number of draws.

11.1.2 Impact of the drift on the posterior distribution

Figure 3 shows, in the Botox case study, the mean of the posterior distribution of the target study treatment for the three main treatment effects considered, and for the different methods. Error bars correspond to the 95% Credible Interval. The posterior distribution strongly depends on the amount of borrowing, with more borrowing leading to a posterior mean closer to the estimate of the source study treatment effect (that is, potentially increased bias), and narrower CrI (reduced variance). In most cases, the CrIs are similar or narrower when the treatment effect in the target study is more consistent with the treatment effect estimate in the source study. This implies that even fixed borrowing methods tend to reject the null hypothesis less often in case of inconsistencies between the source and the target studies.

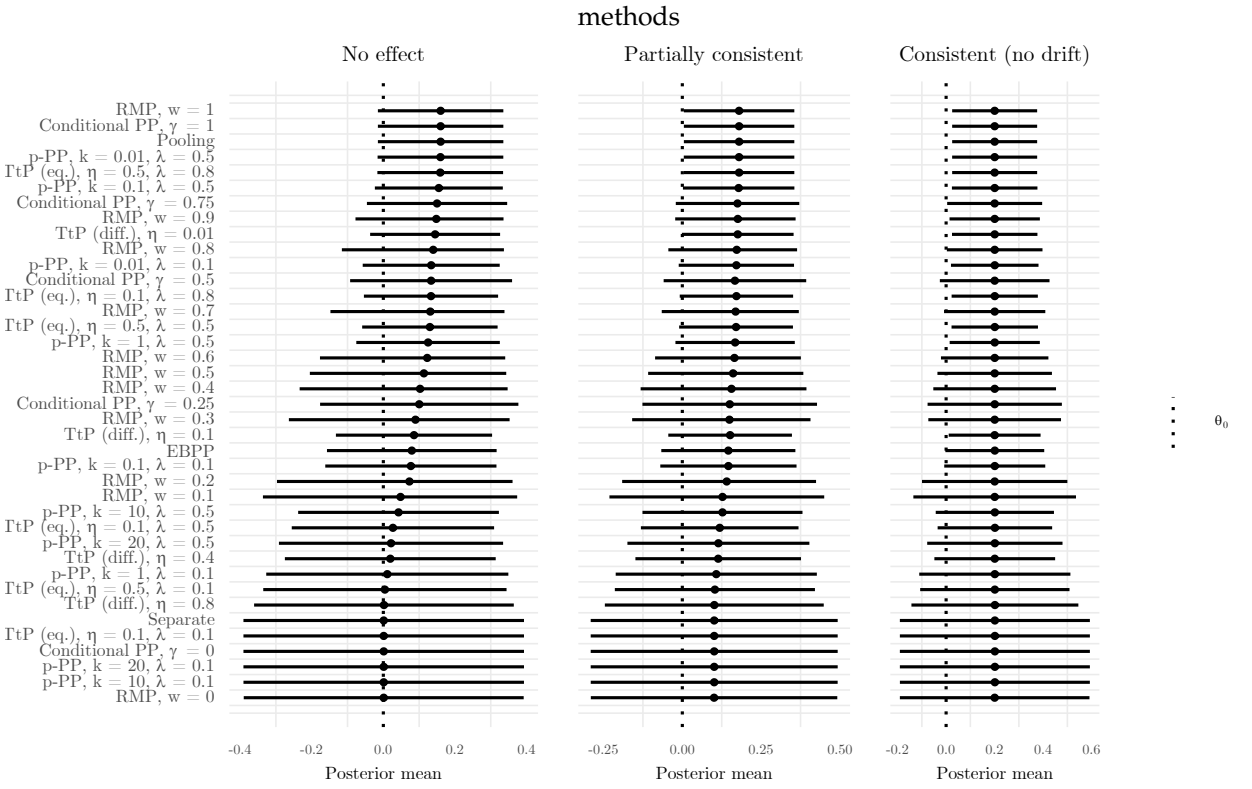

Figure 3: Posterior mean of the treatment effect in the target trial averaged across all simulation replicates, and associated 95% CrI (also averaged over all replicates), for the tree main treatment effects considered, in the Botox case study ($\hat{\theta}_S = 0.2$) with 234 patients per arm in the target trial.

11.2 Impact of borrowing on the probability of success

11.2.1 Type I error inflation

Type I error inflation, that is, a type 1 error increase above the value α that would be obtained for a Bayesian separate analysis with a critical value $\eta = 1 - \alpha$, is the main concern when using partial extrapolation in the context of clinical trials. We observed type 1 error rate inflation in the vast majority of scenarios, irrespective of the method used and its parameterization (supplementary file `noninflated_tie_cases.csv`). The only case where inflation was not observed are listed in table 13. The rare cases where type 1 error rate inflation was not observed in the botox case study occurred when the ratio between the target and source standard deviation was two, with the conditional power prior with $\gamma = 0.25$, and small sample sizes in the target trial ($N_T/2 = 58$ or 39). In the Teriflunomide case, the absence of TIE inflation occurred when the denominator of the source study summary measure was halved. We systematically observed TIE inflation due to borrowing in the Aprepitant, Mepolizumab, and Dapagliflozin case studies.

Figure 4 and 8 (left panel) illustrate type 1 error rate inflation across the different methods, showing that this behavior is systematic as long as information is borrowed, for all treatment effects considered.

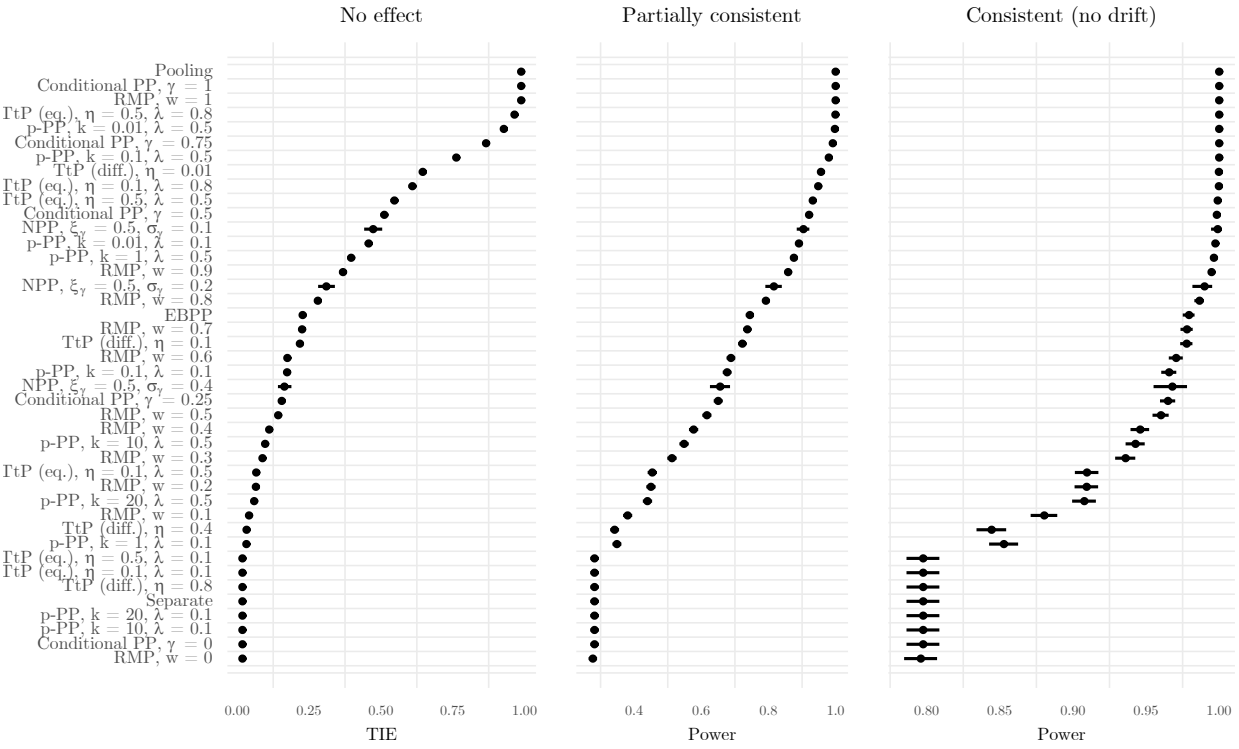

Figure 4: Comparison of the Probability of Success of the different methods in the Belimumab case study, with a target sample size per arm of 281

11.2.2 Power gains under type 1 error control

A major question regarding the use of partial extrapolation methods is whether power gains can be obtained by leveraging external data sources, with a controlled increase in type 1 error rate. If an increase in type 1 error rate is inevitable, then, would the power of the method be higher than the power of a frequentist test (without borrowing) at an equivalent TIE? Intuitively, the use of external information that was used in demonstrating the effectiveness of a drug should increase power, and this may also increase the risk of a type 1 error. In the absence of drift, one may presume that the type 1 error rate will not increase.

The estimated type 1 error rate of the test with borrowing α_B, and the estimated power for $\theta_T > \theta_0$ with borrowing $1 - \beta_B(\theta_T)$ ($\beta_B(\theta_T)$ denotes the type II error rate), are obtained using the following Monte Carlo

Table 13: Cases in which no inflation of the TIE was observed

Method	Case study	$N_T/2$	Source denom. change factor	σ_T/σ_S
Conditional PP, $\gamma = 0.25$	botox	39	1.0	2
	botox	58	1.0	2
RMP, $w = 0$	mepolizumab	68	0.5	NA
	teriflunomide	370	1.0	NA
Conditional PP, $\gamma = 0$	teriflunomide	741	0.5	NA
p-PP, $k = 1$, $\lambda = 0.1$	teriflunomide	741	0.5	NA
p-PP, $k = 10$, $\lambda = 0.1$	teriflunomide	741	0.5	NA
p-PP, $k = 20$, $\lambda = 0.1$	teriflunomide	741	0.5	NA
RMP, $w = 0$	teriflunomide	741	0.5	NA
Separate	teriflunomide	741	0.5	NA
TtP (diff.), $\eta = 0.1$	teriflunomide	741	0.5	NA
TtP (diff.), $\eta = 0.4$	teriflunomide	741	0.5	NA
TtP (diff.), $\eta = 0.8$	teriflunomide	741	0.5	NA
TtP (eq.), $\eta = 0.1$, $\lambda = 0.1$	teriflunomide	741	0.5	NA
TtP (eq.), $\eta = 0.5$, $\lambda = 0.1$	teriflunomide	741	0.5	NA

approximation :

$$\begin{aligned} \alpha_B &= \frac{1}{N_{\text{sims}}} \sum_{i=1}^{N_{\text{sims}}} \varphi_B(d_T^{(i)}|d_S), d_T^{(i)} \sim p(\mathbf{D}_T|\theta_T = \theta_0) \\ \beta_B(\theta_T) &= \frac{1}{N_{\text{sims}}} \sum_{i=1}^{N_{\text{sims}}} \varphi_B(d_T^{(i)}|d_S), d_T^{(i)} \sim p(\mathbf{D}_T|\theta_T) \end{aligned} \tag{33}$$

where N_{sims} is the number of samples drawn from $p(\mathbf{D}_T|\theta_T)$, and $\varphi_B(d_T^{(i)}|d_S)$ is an indicator of meeting the success criterion with borrowing for dataset $d_T^{(i)}$.

To allow for a fair comparison of the power of the test with and without borrowing, Kopp-Schneider et al. (2023) suggest evaluating the TIE rate of the test with borrowing, α_B, and to compare the power with and without borrowing ($\beta_B(\theta_T)$ and $\beta(\theta_T)$ respectively) at a TIE of α_B. Therefore, we also analytically evaluated the power of the frequentist test at level α_B and not only the standard 2.5%. This should allow us to determine whether the Bayesian method has any additional power beyond that gained from simply using a method that increases the traditional frequentist type 1 error rate. Put another way, the goal here is to ascertain whether the improved power is simply bought at the expense of type 1 error control, and if so, which, if any of the models outperform frequentist approaches with an explicitly greater type 1 error rate.

We systematically screened the simulation outputs in order to determine in which cases a power gain was observed at comparable type 1 error rate. The typical pattern we observed is depicted in Figure 5, when plotting the probability of success as a function of drift: the curve representing the probability of success of the borrowing method as a function of drift does not significantly differ from the corresponding curve of the frequentist t-test at equivalent type 1 error rate. Note that the uncertainty related to the frequentist power at equivalent type 1 error rate comes from the uncertainty regarding the equivalent type 1 error rate for the borrowing method.

In some cases a power gain was observed, which however only occurred in case studies other than Botox and Dapagliflozin. We observed, that in several cases, the method of interest was a separate Bayesian analysis (for example, a conjugate power prior with a power parameter of zero). We discuss this paradoxical result in detail in the discussion section.

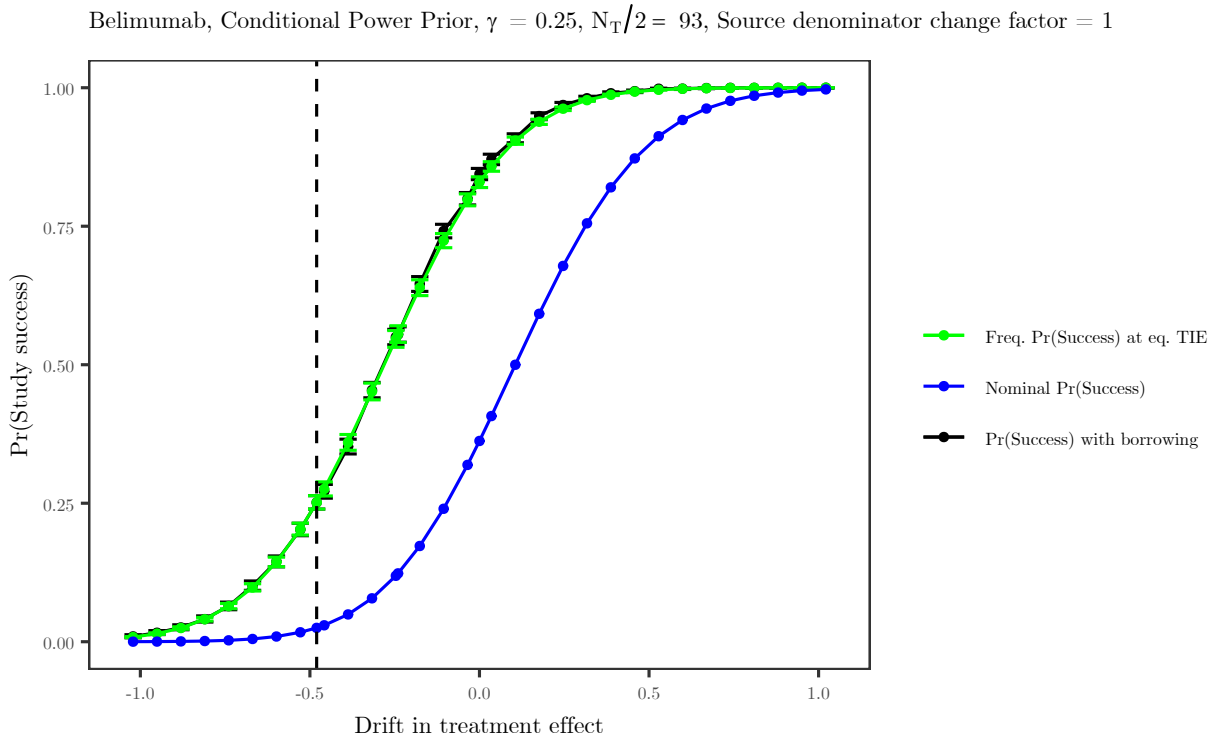

Figure 5: Probability of success of the Conditional Power Prior with $\gamma = 0.25$ as a function of the drift in treatment effect (black) in the Belimumab case study at a sample size per arm of 93, without change introduced in the denominator of the source study summary measure. The probability of success of the t-test at a nominal type 1 error rate of 0.025 as a function of drift is displayed in blue. The probability of success of the t-test at a type 1 error rate equal to the Conditional PP type 1 error rate is displayed in green. Borrowing of external data that favors the null hypothesis also implies that the probability of success of the borrowing method is always larger, in the alternative hypothesis space, than the probability of success of the frequentist method at the nominal type 1 error rate of 0.025. The power curves at equivalent type 1 error rate are identical. $\theta_T = \theta_0$ is indicated by a dashed line. Error bars correspond to the 95% Confidence Interval of the metric.

11.2.3 Power loss due to borrowing

We noticed a behavior of adaptive borrowing methods, in which the probability of success of the borrowing method is lower than the probability of success of the frequentist method at equivalent type 1 error rate in the alternative space (Figure 7). Again, we systematically screened the results for such cases (supplementary file `power_loss_cases.csv`), the vast majority of which occurred at inflated type 1 error rate (supplementary file `power_loss_inflated_tie_cases.csv`). This phenomenon occurred in all case studies, but we focused our analysis on the Botox and Dapagliflozin case studies when investigating this behavior, so as to exclude an effect due to a discrepancy between the data-generating process and assumptions in the t-test (see 12.3). This phenomenon mostly occurs with adaptive borrowing methods, and for most of them. Importantly, it depends on the target study sample size: most occurrences appeared for small target study sample sizes. Moreover, a ratio between the source and target standard deviation of 2 (instead of 1) also increased the sensitivity of methods to this phenomenon. Indeed, in this setting, the test-then-pool variants and the p-value-based PP incurred power loss in the Botox case study for a sample size per arm as large as 234, and the EBPP, which seems more robust overall, was subject to power loss only in this case of higher standard deviation (for a sample size per arm up to 58).

The test-then-pool method based on a test for difference seems especially prone to this issue for a large value of the significance threshold, as it was heavily affected for a target sample size as large as 234 patients per arm in the Botox case study, for a significance threshold of 0.8.

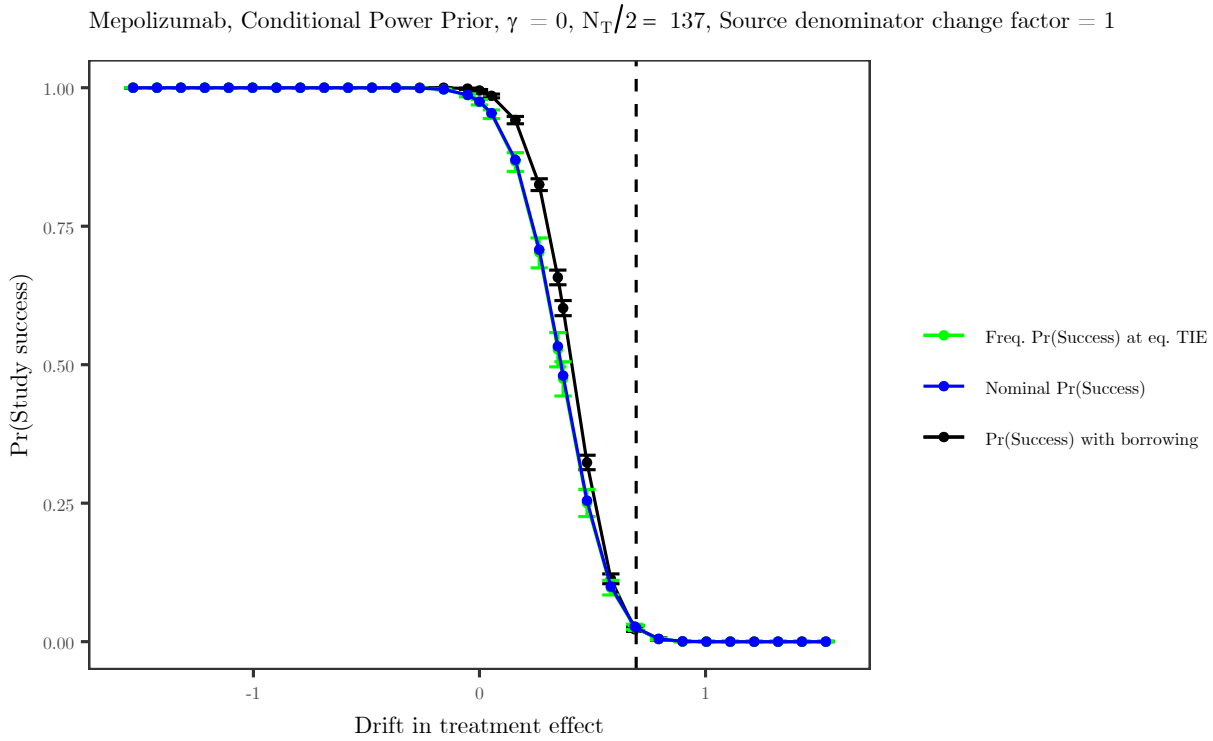

Figure 6: Probability of success of the Conditional Power Prior with $\gamma = 0$ (no borrowing) as a function of the drift in treatment effect (black) in the Mepolizumab case study at a sample size per arm of 137, without change introduced in the denominator of the source study summary measure. The probability of success of the t-test at a nominal type 1 error rate of 0.025 as a function of drift is displayed in blue. The probability of success of the t-test at a type 1 error rate equal to the Conditonal PP type 1 error rate is displayed in green. In this example, an apparent power gain is observed despite the Bayesian analysis being a separate analysis. Error bars correspond to the 95% Confidence Interval of the probability of success.

11.2.4 Impact of the drift on the probability of success.

Figure 8 and Figure 4 show, in the Botox and Belimumab case studies, the probability of success across all simulation replicates and associated 95% CrI. Note that the ordering of methods is made with respect to type 1 error rate (absence of effect, corresponding to a large drift), and it is overall preserved in case of moderate or even absence of drift (partially consistent and consistent effects). We notice that the ordering of methods with respect to the Probability of Success is almost invariant with respect to drift. We observed a similar pattern across target study sample sample sizes and case studies. This implies that power gains are at the expense of increased type 1 error, with some methods sometimes incurring greater type 1 error rate inflation than others at equivalent power gains.

To get a more precise comparison of methods, we compared their power as a function of type 1 error. We observed that no method seemed to consistently outperform the others across scenarios and endpoints: most methods aligned a similar power vs type 1 error rate curve. However, the test-then-pool variants tended to show decreased power at equivalent type 1 error rate compared to other methods (Figure 10 and 9)

Figure 11 shows the probability of success as a function of drift, in the Belimumab case study, across methods. The vertical dashed lines mark the three values of interest for the drift indicating no effect, partially consistent effect and consistent effect compared to the source study (corresponding to figure 8). This figure gives a more detailed view of the variation of the probability of success across methods. The ranking of the methods remains consistent across all drift scenarios. Conditional PP consistently emerges as the top-performing method, exhibiting the highest success probabilities regardless of the drift level. Separate and RMP methods consistently rank lower, with reduced success probabilities. This stability in ranking suggests that the relative effectiveness of these methods is independent of the drift magnitude.

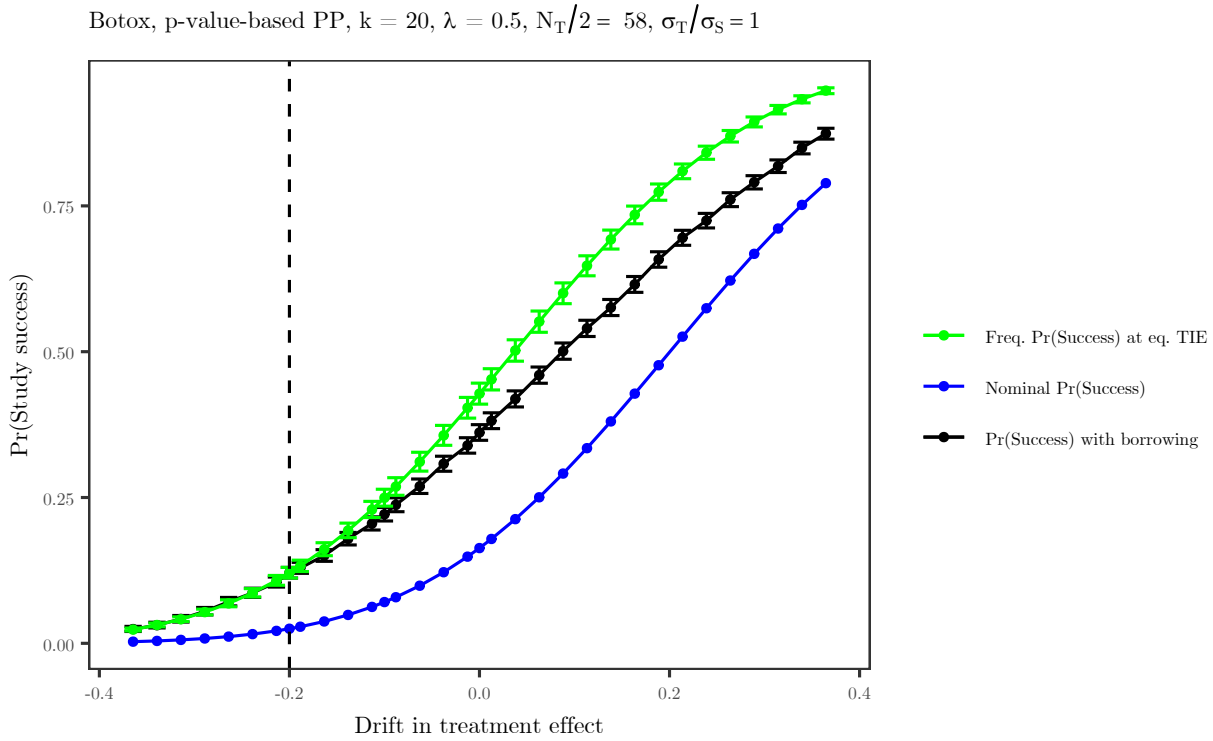

Figure 7: Probability of success of the p-value-based Power Prior with parameters $k = 20$ and $\lambda = 20$ as a function of the drift in treatment effect (black) in the Botox case study at a sample size per arm of 58, with a sampling standard deviation equal between the source and target study. The probability of success of the t-test at a nominal type 1 error rate of 0.025 as a function of drift is displayed in blue. The probability of success of the t-test at a type 1 error rate equal to the p-value based PP type 1 error rate is displayed in green. $\theta_T = \theta_0$ is indicated by a dashed line. In this example, the power of the p-value based power prior is lower than the power of the frequentist t-test at equivalent type 1 error rate in the whole alternative hypothesis space. Error bars correspond to the 95% Confidence Interval of the probability of success.

Drift range for which the success probability is lower than the nominal type 1 error rate Figure 12 presents the drift range for which the success probability is lower than the nominal Type I Error rate in the Botox case study (with a sample size of 234 subjects per arm in the target trial). The figure highlights the fact for very small drift values, the probability of success gets smaller than the nominal type 1 error rate.

Drift range for which the success probability is larger than the nominal power Figure 13 illustrates the drift range in treatment effect for which the success probability exceeds the nominal statistical power (that is, the power of a separate analysis) in the Botox case study, using a sample size of 58 subjects per arm in the target trial. We observe that the power of the borrowing method is uniformly larger than the nominal power in the whole alternative hypothesis space, while the type 1 error rate is inflated. This is consistent with our previous results showing that borrowing tend to increase power at the expense of TIE inflation.

11.2.5 Impact of the target study sample size on the probability of success.

Figure 14 shows the probability of success as a function of the study sample size per arm in the Belimumab case study with a partially consistent effect. The probability of success increases with the study sample size for all the methods, except for the EBPP and Test and Pool with $\eta = 0.1$, for which it is decreasing. In this setting, pooling the source and target data tends to increase the power. However as the target study sample size increases, it becomes clearer that there is an inconsistency between the source and target study, so dynamic borrowing methods will tend to reject the source study data more. This rejection of source study data can compensate for the benefit of increased sample size, hence resulting in decreasing power with increased target study sample size.

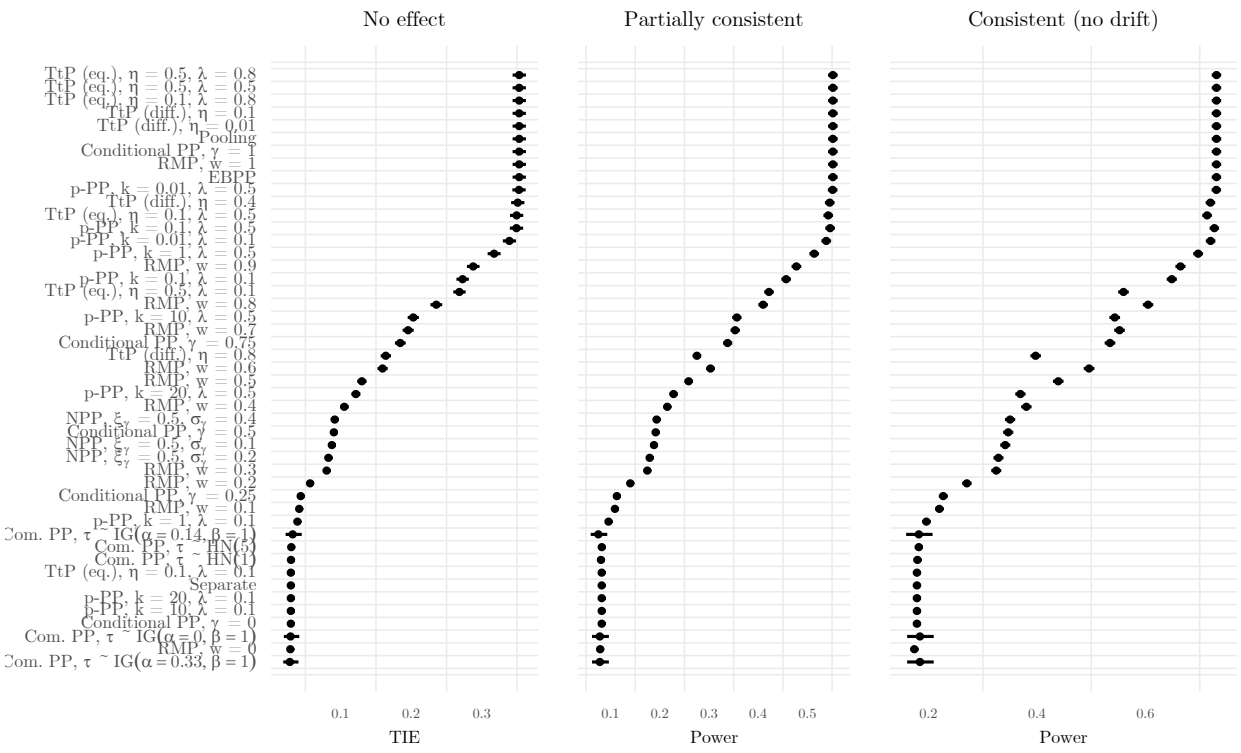

Figure 8: Probability of success across all simulation replicates and associated 95% CI, for the three main treatment effects considered, in the Botox case study with 58 samples per arm in the target trial.

11.2.6 Impact of changes in the denominator of ratio summary measures in the source study on the probability of success.

We studied the impact of change in the denominator of the source study summary measure on inference in the Belimumab, Mepolizumab, and Teriflunomide case studies. However, a study in terms of prior ESS is made difficult when changing the source standard deviation for the reasons explained above (Figure 60). We observed that changes in the denominator of source studies summary measure had little impact on the probability of success, but a larger denominator led to a higher probability of success, in a manner that did not widely differ between methods (Figure 15).

11.2.7 Probability of success as a function of type 1 error rate across methods.

Figure 9 compares, in the Botox case study, the probability of success as a function of type 1 error rate across methods. The results are explicited in Table 14.

Table 14: Botox, partially consistent treatment effect, $N_T/2 = 58, \sigma_T/\sigma_S = 1$.

Method	Type 1 error rate	Power difference
Com.PP, $\tau \sim$ HN(1)	0.030 [0.026, 0.033]	0.0063 [0.0010, 0.0118]
Com.PP, $\tau \sim$ HN(5)	0.030 [0.027, 0.034]	0.0060 [0.0007, 0.0116]
Com.PP, $\tau \sim$ IG($\alpha = 0$, $\beta = 1$)	0.029 [0.019, 0.041]	0.0040 [-0.0119, 0.0224]
Com.PP, $\tau \sim$ IG($\alpha = 0.14$, $\beta = 1$)	0.032 [0.022, 0.045]	-0.0030 [-0.0186, 0.0151]
Com.PP, $\tau \sim$ IG($\alpha = 0.33$, $\beta = 1$)	0.028 [0.019, 0.040]	0.0060 [-0.0099, 0.0244]
Conditional Power Prior, $\gamma = 0$	0.029 [0.026, 0.033]	0.0078 [0.0025, 0.0133]
Conditional Power Prior, $\gamma = 0.25$	0.043 [0.040, 0.048]	0.0036 [-0.0025, 0.0100]

Conditional Power Prior, $\gamma = 0.5$	0.090 [0.085, 0.096]	-0.0035 [-0.0112, 0.0044]
Conditional Power Prior, $\gamma = 0.75$	0.184 [0.177, 0.192]	0.0037 [-0.0056, 0.0131]
Conditional Power Prior, $\gamma = 1$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
EBPP	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
NPP, $\xi_\gamma = 0.5$, $\sigma_\gamma = 0.1$	0.087 [0.082, 0.093]	-0.0040 [-0.0116, 0.0038]
NPP, $\xi_\gamma = 0.5$, $\sigma_\gamma = 0.2$	0.083 [0.077, 0.088]	-0.0067 [-0.0142, 0.0010]
NPP, $\xi_\gamma = 0.5$, $\sigma_\gamma = 0.4$	0.092 [0.086, 0.097]	-0.0016 [-0.0093, 0.0063]
Pooling	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
RMP, $w = 0$	0.029 [0.025, 0.032]	0.0047 [-0.0005, 0.0102]
RMP, $w = 0.1$	0.041 [0.037, 0.045]	0.0047 [-0.0013, 0.0110]
RMP, $w = 0.2$	0.057 [0.052, 0.061]	0.0019 [-0.0048, 0.0089]
RMP, $w = 0.3$	0.080 [0.075, 0.086]	-0.0095 [-0.0169, -0.0019]
RMP, $w = 0.4$	0.105 [0.099, 0.111]	-0.0029 [-0.0109, 0.0053]
RMP, $w = 0.5$	0.130 [0.123, 0.137]	-0.0025 [-0.0111, 0.0062]
RMP, $w = 0.6$	0.159 [0.152, 0.167]	0.0011 [-0.0101, 0.0080]
RMP, $w = 0.7$	0.196 [0.188, 0.203]	0.0040 [-0.0054, 0.0135]
RMP, $w = 0.8$	0.235 [0.227, 0.244]	0.0018 [-0.0079, 0.0115]
RMP, $w = 0.9$	0.288 [0.279, 0.297]	0.0085 [-0.0013, 0.0183]
RMP, $w = 1$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
Separate	0.029 [0.026, 0.033]	0.0078 [0.0025, 0.0133]
Test-then-pool (difference), $\eta = 0.01$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
Test-then-pool (difference), $\eta = 0.1$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
Test-then-pool (difference), $\eta = 0.4$	0.351 [0.341, 0.360]	-0.0118 [-0.0216, -0.0020]
Test-then-pool (difference), $\eta = 0.8$	0.164 [0.157, 0.171]	-0.0337 [-0.0424, -0.0248]
Test-then-pool (equivalence), $\eta = 0.1$, $\lambda = 0.1$	0.029 [0.026, 0.033]	0.0078 [0.0025, 0.0133]
Test-then-pool (equivalence), $\eta = 0.1$, $\lambda = 0.5$	0.349 [0.340, 0.359]	-0.0148 [-0.0246, -0.0050]
Test-then-pool (equivalence), $\eta = 0.1$, $\lambda = 0.8$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
Test-then-pool (equivalence), $\eta = 0.5$, $\lambda = 0.1$	0.268 [0.260, 0.277]	-0.0282 [-0.0379, -0.0184]
Test-then-pool (equivalence), $\eta = 0.5$, $\lambda = 0.5$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
Test-then-pool (equivalence), $\eta = 0.5$, $\lambda = 0.8$	0.353 [0.344, 0.362]	-0.0074 [-0.0172, 0.0024]
p-value-based PP, $k = 0.01$, $\lambda = 0.1$	0.339 [0.330, 0.348]	-0.0059 [-0.0157, 0.0039]
p-value-based PP, $k = 0.01$, $\lambda = 0.5$	0.353 [0.343, 0.362]	-0.0079 [-0.0177, 0.0019]
p-value-based PP, $k = 0.1$, $\lambda = 0.1$	0.273 [0.264, 0.281]	0.0026 [-0.0072, 0.0124]
p-value-based PP, $k = 0.1$, $\lambda = 0.5$	0.349 [0.340, 0.358]	-0.0110 [-0.0208, -0.0012]
p-value-based PP, $k = 1$, $\lambda = 0.1$	0.039 [0.035, 0.043]	-0.0033 [-0.0090, 0.0026]
p-value-based PP, $k = 1$, $\lambda = 0.5$	0.317 [0.308, 0.327]	0.0046 [-0.0052, 0.0144]
p-value-based PP, $k = 10$, $\lambda = 0.1$	0.029 [0.026, 0.033]	0.0078 [0.0025, 0.0133]
p-value-based PP, $k = 10$, $\lambda = 0.5$	0.203 [0.195, 0.211]	-0.0025 [-0.0119, 0.0070]
p-value-based PP, $k = 20$, $\lambda = 0.1$	0.029 [0.026, 0.033]	0.0078 [0.0025, 0.0133]
p-value-based PP, $k = 20$, $\lambda = 0.5$	0.121 [0.115, 0.128]	-0.0162 [-0.0244, -0.0079]

11.3 Impact of scenario parameters on the amount of borrowing.

11.3.1 Impact of the drift on the Prior Effective Sample Size.

Drift tends to reduce the amount of borrowing of adaptive borrowing methods, in a way that strongly depends on their parameters and the scenario (Figure 16). This behavior is what characterizes adaptation to drift of these methods. The parameters and methods are not equally conservative: some, like the Test-then-Pool variants considered, only start discarding external information for very large drift values, so that even in the absence of treatment effect in the target study, the prior ESS stays of the same order of magnitude as the target study sample size. By contrast, some adaptive borrowing methods like the EBPP and the RMP (for $w \neq 1$) seem to never fully borrow external information. This behavior is not problematic in itself, as it may happen that consistency between the target and the source treatment effects happened by mere chance.

11.3.2 Impact of the target study sample size on the Prior Effective Sample Size.

As the target study sample size increases, the confidence in the level of agreement between the source and target study data increases. Therefore, as the target study sample size increases, one may expect the prior ESS of adaptive borrowing methods to increase in case of consistent treatment effects, and to decrease in case of inconsistency. We see in the example in Figure 17, where the treatment effects are consistent, that this is indeed the case for at least some methods, in particular the RMP, the p-value-based power prior, and the test-then-pool variants. Methods like the Commensurate Power Prior or the EBPP did not show increase in moment-based ESS (which we also observed in other case studies as well). We observed that, even with a partially consistent effect, several methods (test-then-pool with an equivalence test, the RMP, and the p-value-based power prior), would still borrow more information despite an increase in N_T. This is even more the case when $\sigma_T/\sigma_S = 2$ (see Figure 18). Even in the absence of effect, methods like the p-value-based power prior and the test-then-pool with an equivalence test would still borrow more information from the source as N_T increases, which seems paradoxical (Figure 19).

11.3.3 Impact of changes in the denominator of ratio summary measures in the source study on the Prior Effective Sample Size.

We found that there was little to no effect of changes in the denominator of the source study summary measure on the prior ESS (Figure 20).

11.3.4 Impact of the standard deviation in the target study on the Prior Effective Sample Size.

For methods borrowing information from the source study, the moment-based ESS strongly depended on the ratio between the target and source studies standard deviation (Figure 21). This is explained by the fact that, as the standard deviation increases in the target study, the relative weight of the prior increases.

11.4 Impact of borrowing on MSE and bias.

11.4.1 Comparison of MSE and bias across methods

Figure 22 compares the MSE of the different methods in the Botox case study for the three main treatment effects considered. No differences were observed between partially consistent and consistent effect scenarios. Performance seems to be more driven by the parameters scenarios and level of borrowing than by the actual methods.

Figure 23 shows the bias of the different methods in the Botox case study for the three main treatment effects considered.

11.4.2 Comparison of MSE and bias versus type 1 error rate across methods

We plotted, for each method/parameters combination, the MSE against the type 1 error rate of the corresponding method in Figure 25. This provides a measure of the accuracy of the estimation, for a given type 1 error rate inflation. We observed, in the Botox case study, that the test-then-pool variants tend to incur much larger MSE than other methods at similar type 1 error rate. Similarly, the p-value based power prior displays

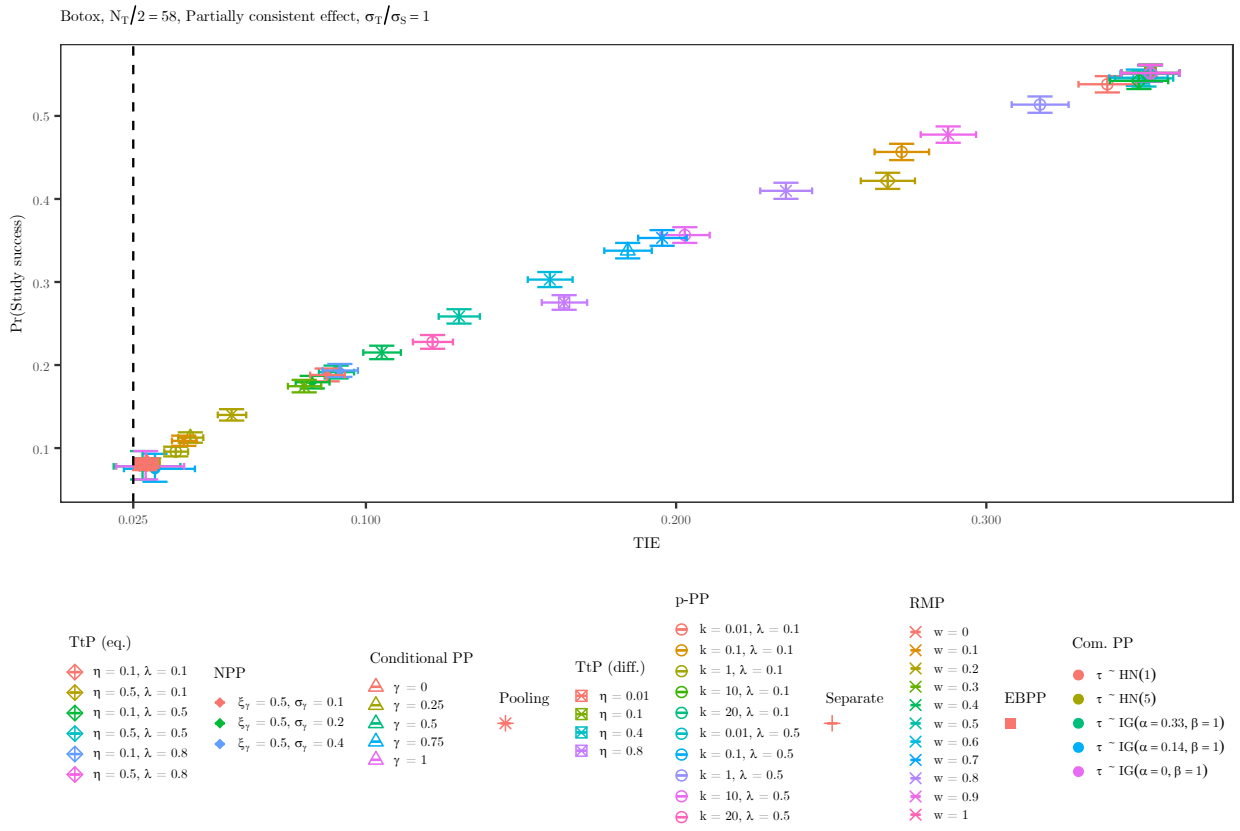

Figure 9: Probability of success as a function of TIE in the Botox case study with a sample size per arm of 58, across all the methods and parameters. The treatment effect is partially consistent, the target-to-source standard deviation ratio is 1. Error bars correspond to the 95% Confidence Interval of the Probability of Success and TIE. Dashed vertical line represents the nominal TIE of 0.025.

higher MSE than most other methods. Methods that provide the smaller MSE at comparable type 1 error rate are the EBPP and the conditional power prior. Importantly, we observe a similar relative behavior of the different methods across sample sizes and drift values. A similar behavior could be observed in other case studies. In the Belimumab case study, we observed a similar behavior, apart from the RMP that was the less performing method in the absence of effect in the target study, but similar to the conditional PP case of consistent effect.

11.4.3 Impact of the drift on the MSE and bias.

While fixed borrowing methods display a quadratic MSE as a function of drift (and a linear bias) (Figures 26 and 27), adaptive borrowing methods discard external information for large drift values (Figure 28). Therefore, for large drift values, the frequentist operating characteristics of adaptive methods are equivalent to those of frequentist methods.

For small drift values, the bias is small while the variance of the posterior distribution of the treatment effect is reduced compared to the case of a separate analysis. Therefore, the precision of the estimation is improved compared to a method that borrows less (Figure 28).

For each frequentist operating characteristic, we call the range of drift values in which the operating characteristic is improved compared to a separate analysis the "sweet spot". Given the limited number of drift values considered in the simulation, we used linear interpolation to get an estimate of the bounds of the sweet spots.

If we consider the RMP, it appears that the "discarding" behavior only occurs for extremely large drift values.

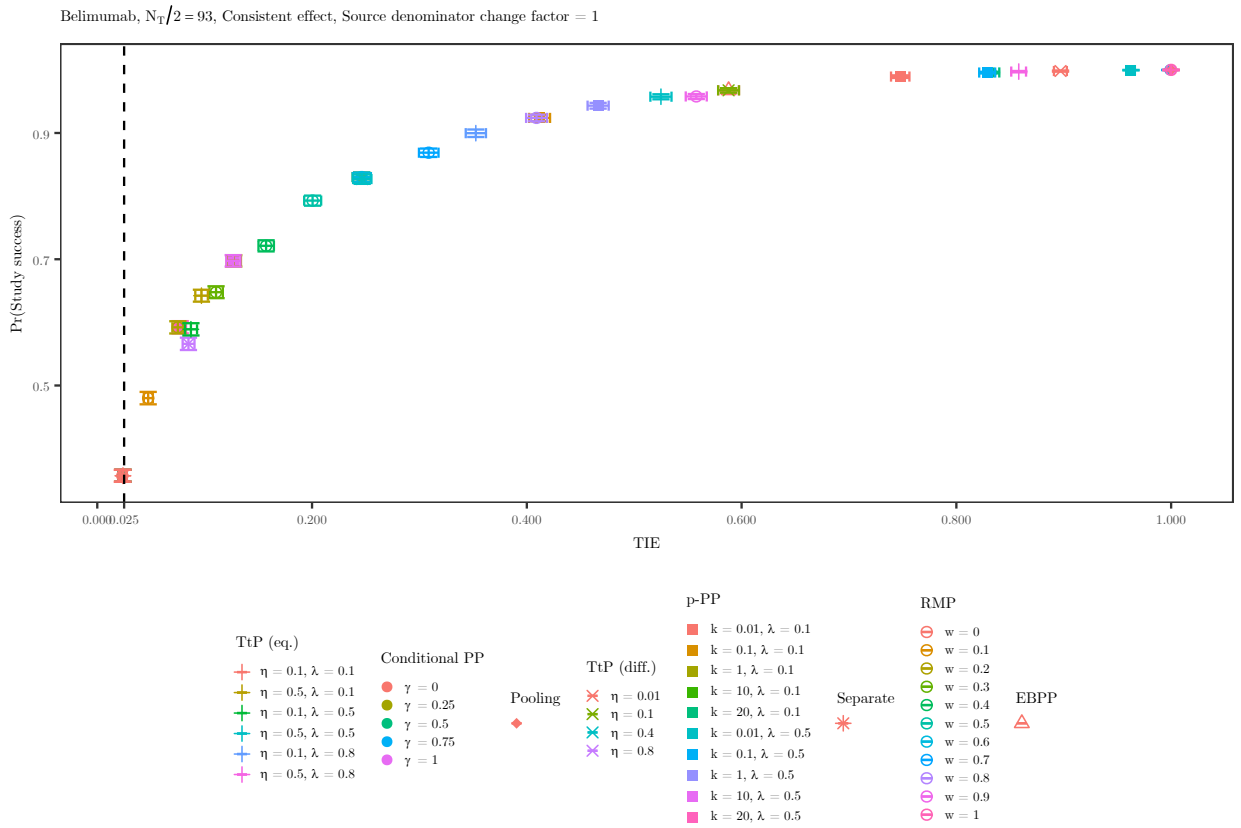

Figure 10: Probability of success as a function of TIE in the Belimumab case study with a sample size per arm of 93, across all the methods and parameters. The treatment effect is consistent, the target to source standard deviation ratio is 1. Error bars correspond to the 95% Confidence Interval of the Probability of Success and type 1 error rate. Dashed vertical line represents the nominal type 1 error rate of 0.025.

Sweet spot for the MSE. Figure 29 show the sweet spot for MSE, in the Botox case study, with $N_T/2 = 58$.

We see that all methods, except the test-then-pool variants, p-value based-power prior, and separate analysis, show wide sweet spots relative to MSE, that encompass the range of drift values from no treatment effect to consistent effect. This benefit of borrowing compared to separate analysis is however less pronounced when the sample size increases in the target study.

Impact of the target study sample size on the sweet spot for the MSE. As the sample size increases in the target study, the width of the sweet spot for MSE decreases (Figure 30), indicating less robustness of the benefit of borrowing to drift in treatment effect.

11.4.4 Impact of changes in the denominator of ratio summary measures in the source study on MSE

We observed that MSE tended to decrease with an increase in the denominator of source study summary measures, although this effect tended to be small. We did not observe systematic differences in the behavior of different methods in response to these changes. 31.

11.4.5 Impact of changes in the standard deviation in the target study on MSE

An increase in σ_T/σ_S resulted in larger MSE 32, which can be explained by an increased variance of the posterior distribution. However, this increase was mitigated by methods that borrowed a lot of information from the source study. Again, this can be explained by the limited impact a small data sample with high standard deviation would have on the variance of the posterior distribution when pooled with a large amount of source data.

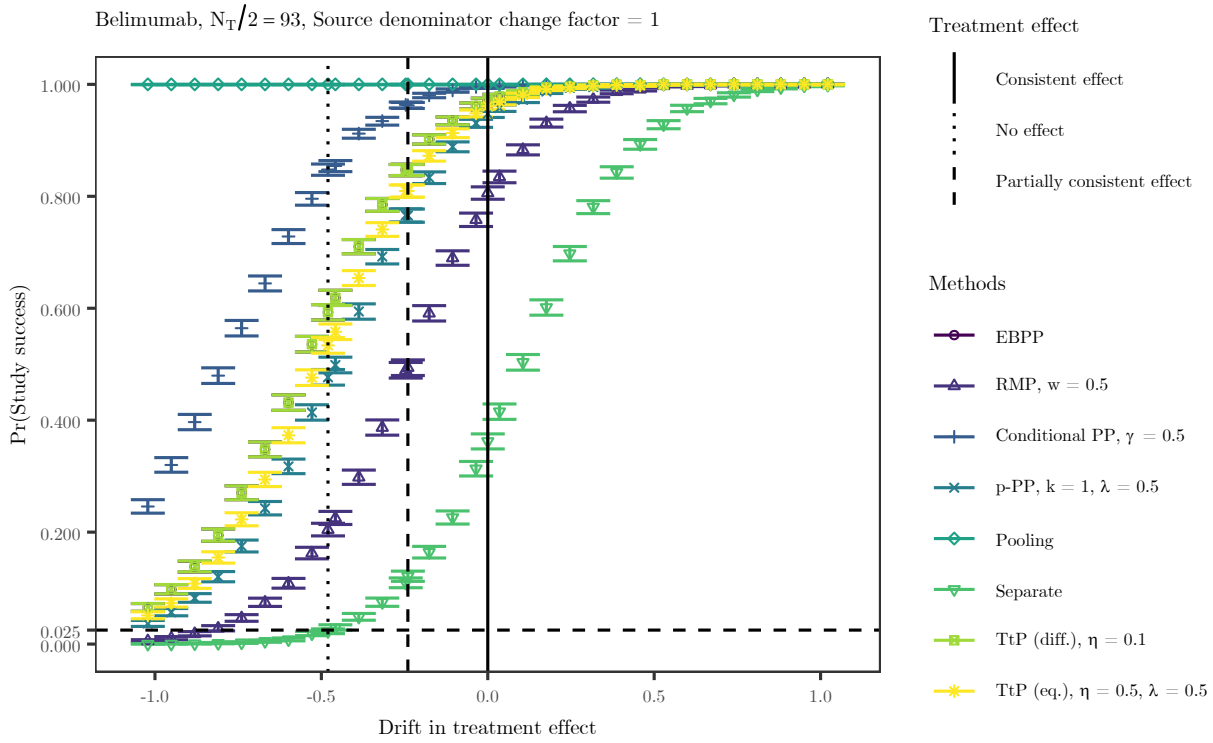

Figure 11: Probability of success across methods as a function of the drift in treatment effect in the Belimumab case study at a sample size per arm of 93, without change introduced in the denominator of the source study summary measure.

11.5 Impact of borrowing on precision.

We measured the precision of the borrowing methods as the mean half-width of the 95% Credible Interval. We observed on (Figure 33) that the precision is strongly driven by the amount of information borrowed. The half-width of the 95% Credible Interval is a measure of the strength of the belief represented by the posterior distribution. It corresponds to the variance component in the bias-variance tradeoff implied when extrapolating.

Figure 34 shows the typical pattern observed when considering the effect of drift on precision for adaptive borrowing methods: the precision decreases (i.e., the half-width of the 95% increases), as the drift in treatment increases. A symmetrical behavior occurs for negative drift values, whereby the precision decreases as the drift value goes away from zero. This can be understood by considering the behavior of the prior ESS of adaptive borrowing methods with drift: the prior ESS also decreases in a similar fashion when the drift value goes away from zero. This implies that the posterior will be less sharp, hence the wider 95% confidence interval.

As expected, for increasing sample size per arm, the precision also increases (Figure 35).

11.5.1 Precision as a function of type 1 error rate across methods.

The precision was observed to be decreasing with 1 error rate (see Figure 36).

Figure 37 illustrates, in the Botox case, the relationship between precision and type 1 error rate for the methods tested with a partially consistent treatment effect.

11.5.2 Impact of the drift on precision.

Figure 38 illustrates the precision of the different methods for three main treatment effects considered in the Botox case study with 39 samples per arm. Overall, the ordering of methods with respect to their precision tends to be preserved across drift values and mostly depends on the amount of borrowing. Drift seems to

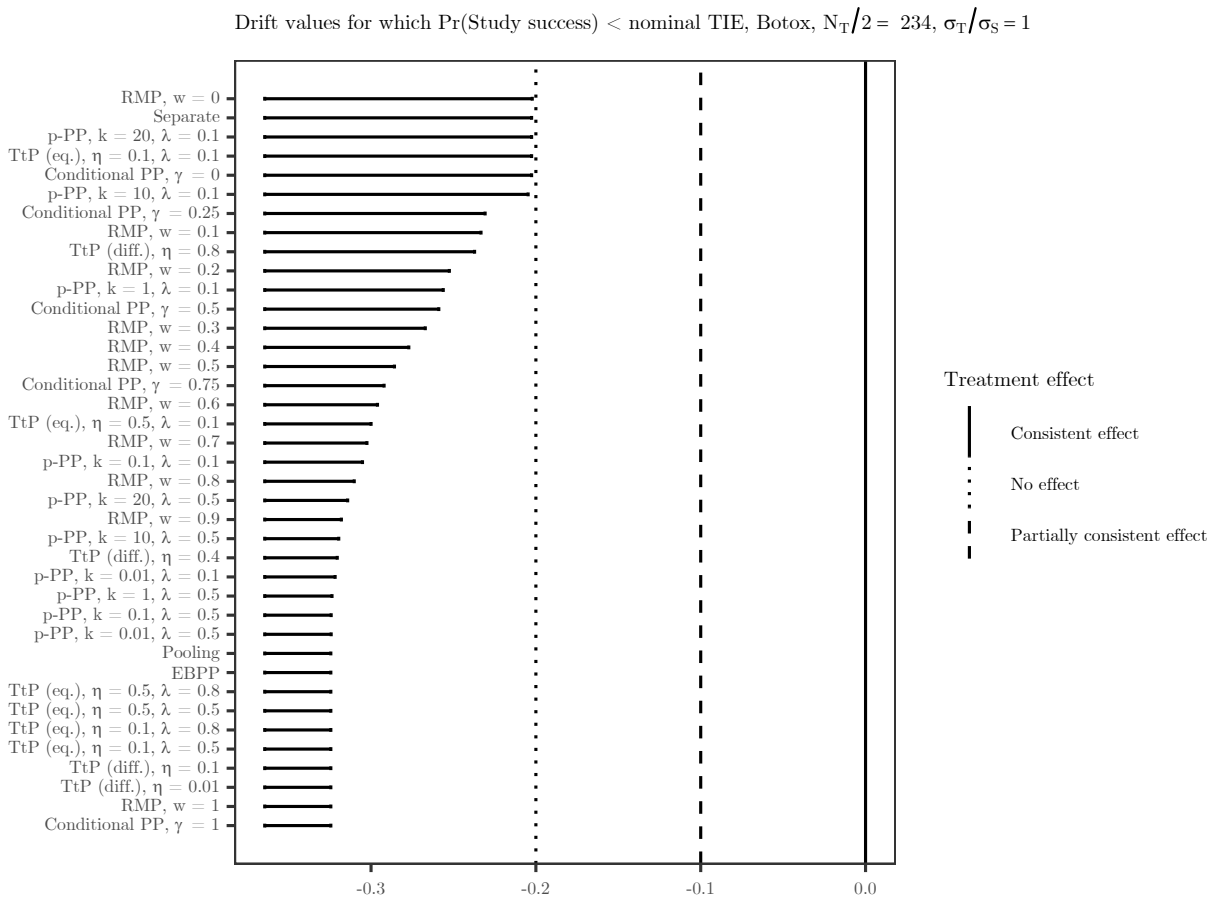

Figure 12: Sweet spot for the probability of success in the Botox case study with 234 patients per arm.

have a limited impact on precision, but because dynamic borrowing methods discard external information as drift increases, their precision also tends to decrease as drift increases.

We find that the extent of the sweet spot relative to precision (Figure 39) is strongly related to the amount of borrowing. Most methods displayed a sweet spot that encompasses

11.5.3 Prior probability of study success

Except in the absence of borrowing, when a UI design prior is more informative than a noninformative analysis prior, the prior probability of study success is maximal with the source posterior as design prior, and minimal with a UI design prior. As expected, the prior probability of study success monotonically increases with the target study sample size, unless a noninformative design prior is used (Figure 42). Indeed, given that the design priors favor a positive treatment effect, we expect larger sample sizes to increase the chances of meeting the success criterion. As expected, for all design priors, increased borrowing also leads to increased probability of study success, as the conditional probability of study success $Pr(\text{Study success}|\theta_T)$ increases with borrowing.

It is interesting to see that only a moderate increase in borrowing strength can dramatically increase the (Bayesian) probability of study success (Figure 43). This is related to the fact that the frequentist power can dramatically increase for moderate increase in the level α of the test.

Figure 44 shows the Bayesian type 1 error rate (or average type 1 error rate) for the three different design priors considered, in the Belimumab case study with a target sample size per arm of 93. Interestingly, the ordering of methods with respect to their type 1 error rate is the same across design priors, yet the use of the source posterior as design prior implies a much larger type 1 error rate compared to a UI design prior for methods that borrow from the source study. More borrowing systematically leads to higher TIE.

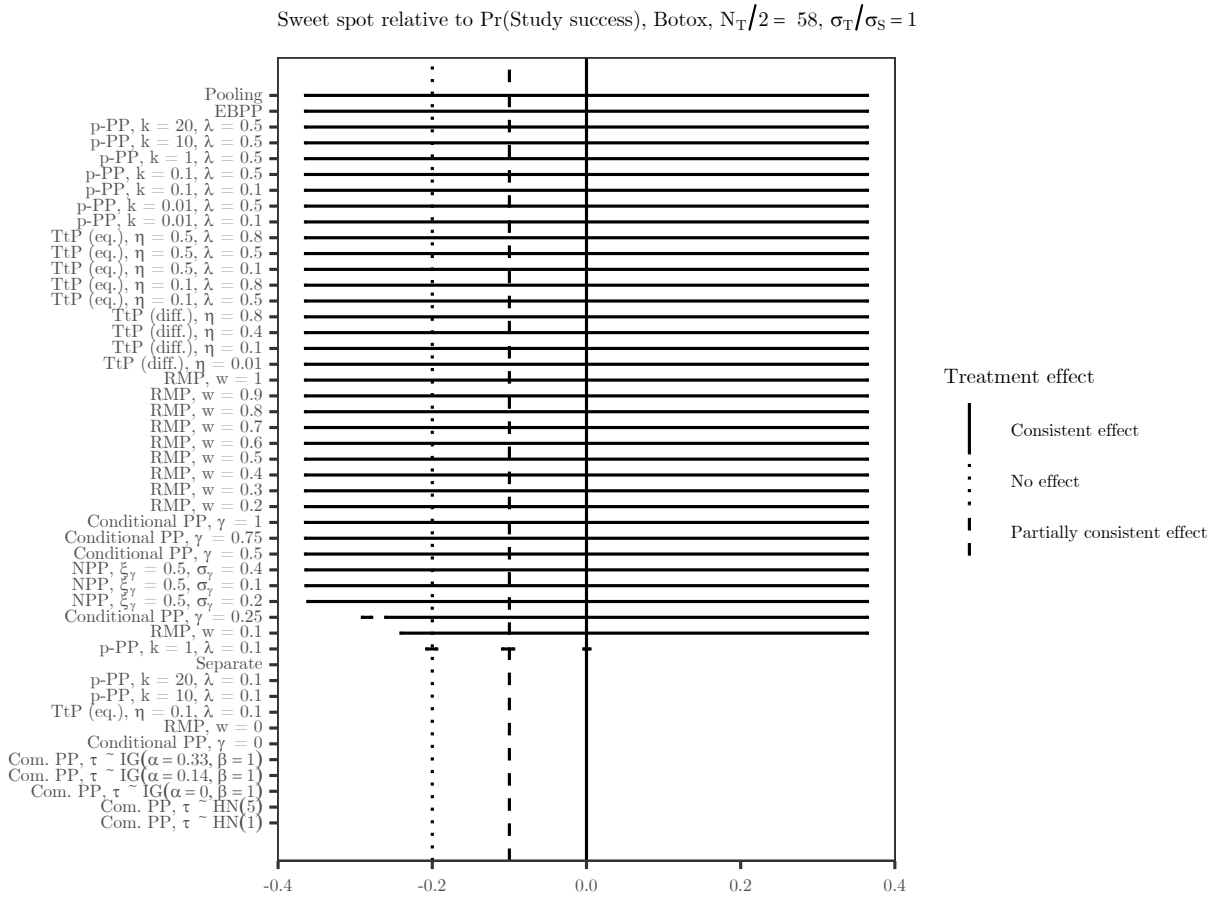

Figure 13: Sweet spot for the probability of success, in the Botox case study with 58 subjects per arm in the target trial.

Impact of the standard deviation in the target study on the Bayesian type 1 error. Figure 47 and Figure 48 show the average type 1 error rate and power, respectively, across different methods for the Belimumab case study with a sample size per arm in the target study of 281 patients. We observed a monotonic increase in average type 1 error rate and average power with increase in borrowing strength.

Note that a UI design prior does not lead to constant average type 1 error rate and power as a function of borrowing strength, as the mean of the prior is set to the mean of the source posterior. Therefore, we observed a very small increase in these Bayesian OCs with borrowing strength. As expected, the Bayesian OCs computed with UI prior and Source posterior bound these Bayesian operating characteristics, except in the absence of borrowing, where the source Analysis Prior is noninformative, and therefore less informative than a UI prior. As expected, the average type 1 error rate decreases while the average Power increases with target sample size per arm (Figure 45 and Figure 46).

Other metrics that can be considered at the design stage are the pre-posterior probability of a false positive, the prior probability of no treatment benefit, the prior probability of study success (which, from a Bayesian point of view, can be seen as a prior predictive probability), and upper bound on the pre-posterior probability of FP (see definitions in Table 8 and corresponding design priors in Table 6). We report these metrics in the supplementary material.

11.5.4 Bayesian power and type 1 error

Comparison of the Bayesian power across methods Figure 51 shows the Bayesian power (or average power) for the three different design priors considered, in the Belimumab case study with a target sample size per arm of 93. As for the Bayesian type 1 error rate, the ordering of methods with respect to their Bayesian power is the same across design priors, yet the use of the source posterior as design prior implies a

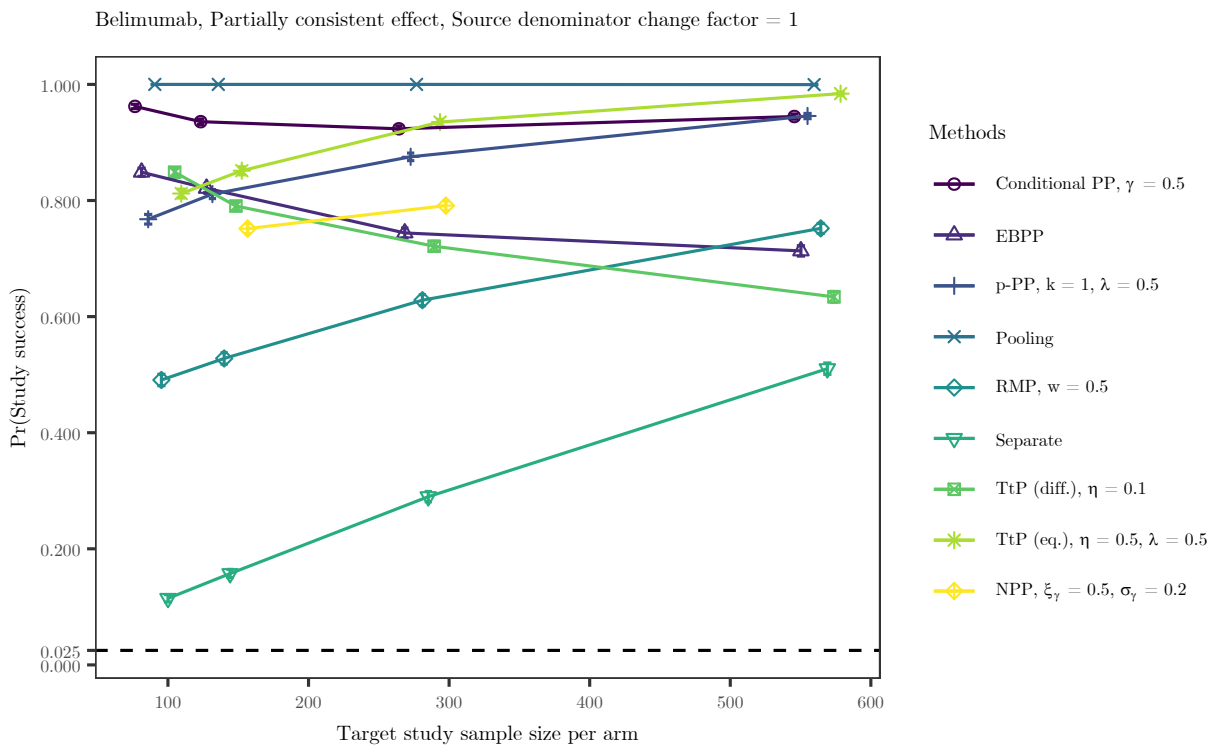

Figure 14: Probability of success as a function of the sample size per arm in the Belimumab case study, across methods, without change introduced in the denominator of the source study summary measure. Error bars correspond to the 95% Confidence Interval of the Probability of Success.

much larger Bayesian power compared to a UI design prior. More borrowing systematically leads to higher average power. Note that for the separate analysis, the average power is smaller with an analysis prior as designed prior compared to the UI design prior. This can be explained by the fact that the analysis prior, in this case, is uniform, and therefore even less informative than a UI design prior. We note that the ordering of methods with respect to their Bayesian power is the same as the ordering with respect to the Bayesian type 1 error rate (see Figure 44).

Impact of the target study sample size on the Bayesian power. The relationship between the Bayesian power and the target study sample size is helpful in determining the feasibility and risk of a clinical trial or program. The simulations show (see e.g. Figure 52) that the impact of the target study sample size is relatively limited and consistent across methods, except for the separate analysis for which a gain of up to +10 points on power can be observed by doubling the sample size.

Bayesian power as a function of Bayesian type 1 error rate across methods The simulation results illustrated by the Belimumab case below show that the average Bayesian power grows with average type 1 error rate, with a huge impact of the underlying design prior: source prior leading to a 3 times higher power vs. UI design prior. Figure 54.

Figure 55.

11.5.5 Pre-posterior probability of True Positives (TP)

The Figure 56 below illustrates on the Belimumab case how methods compare on the pre-posterior probability of true positives (TP) for various design priors. As illustrated below, this metric is highly dependent on the method, independently from the design prior option. Pre-posterior probabilities are more driven by the model parameters than by the methods themselves.

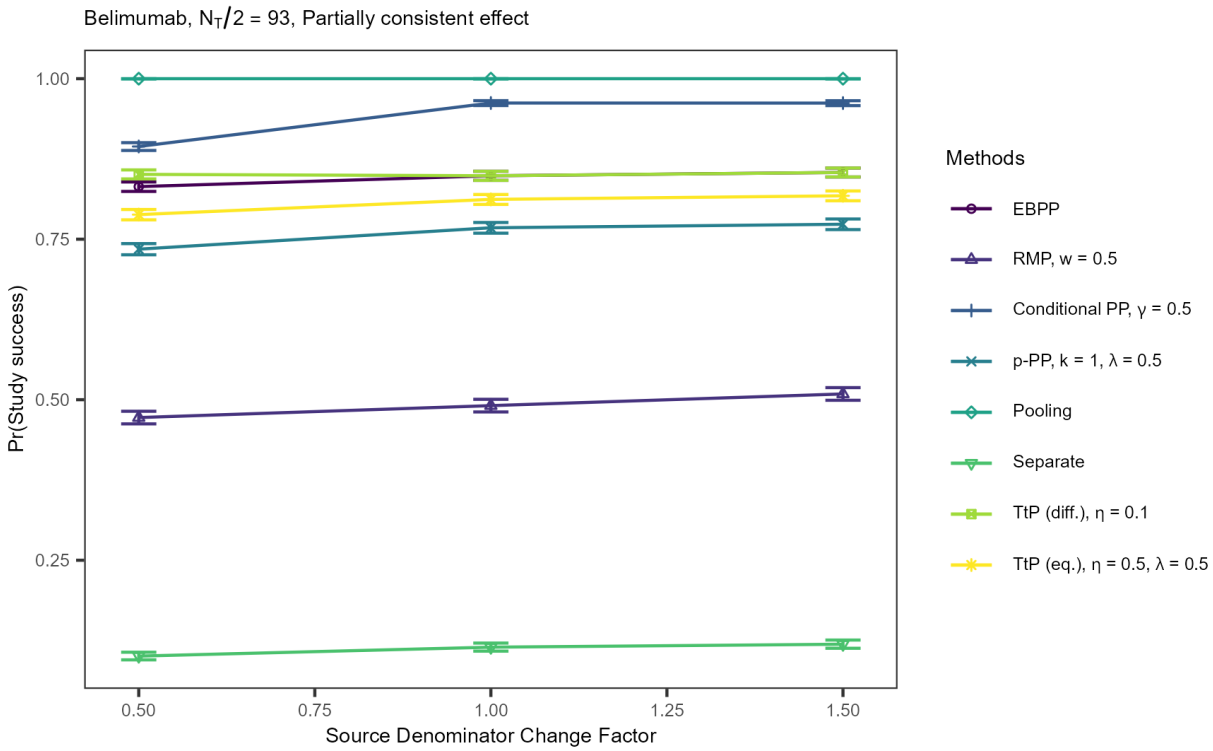

Figure 15: Success probability as a function of changes in the denominator of the summary measure from the source study, with a partially consistent treatment effect.

11.6 How methods' parameters and drift impact the amount of borrowing

In addition to ELIR, we computed the prior ESS as the difference between the effective sample size of the posterior distribution and the sample size of the target trial. The calculation of the effective sample size of the posterior distribution was, in turn, calculated using the "moment-based" matching method described in the RBesT package as well as a "precision-based" matching method. Briefly, the moment-based approach matches the mean and the variance of a given prior distribution to a posterior distribution from the conjugate family updated with data worth the effective sample size. Similarly, the "precision-based" approach matches on the half-width of the 95% credible interval. Differences between methods are multi-directional and can be further observed in the appended full results.

In the following sections, we will start by comparing the results obtained with the different prior ESS measures. Then, focusing on the moment-based ESS, we will describe, for each method, how the prior ESS varies with drift.

11.6.1 Consistency of the different prior ESS measures

Performing a systematic comparison of the different ESS measures, while possible based on our results, is out of the scope of this project. However, we observed very similar values between the mean precision-based ESS, the mean moment-based ESS, and the mean ELIR ESS when the posterior is Gaussian. Therefore, to avoid redundancy, we will focus our analysis on the moment-based ESS in these cases. Similar figures with the other ESS measures can be found in the supplementary material. Note that in some cases, such as for example with the NPP, the ESS ELIR shows erratic behavior for unknown reason. Moreover, in some cases discussed below, the ESS ELIR method shows widely different results compared to the two other ESS measures. Figures 57, 58 and 59 show the prior ESS of different methods, for the three main treatment effect values considered. We show that the ESS ELIR markedly differs from the other measures.

Note that in case studies with non-Gaussian endpoints, the standard error on the treatment effect not only depends on the sample size in the target study, but on the size of the treatment effect and the drift. However, the empirical standard deviation is used for computing the ESS. Therefore, care should be taken

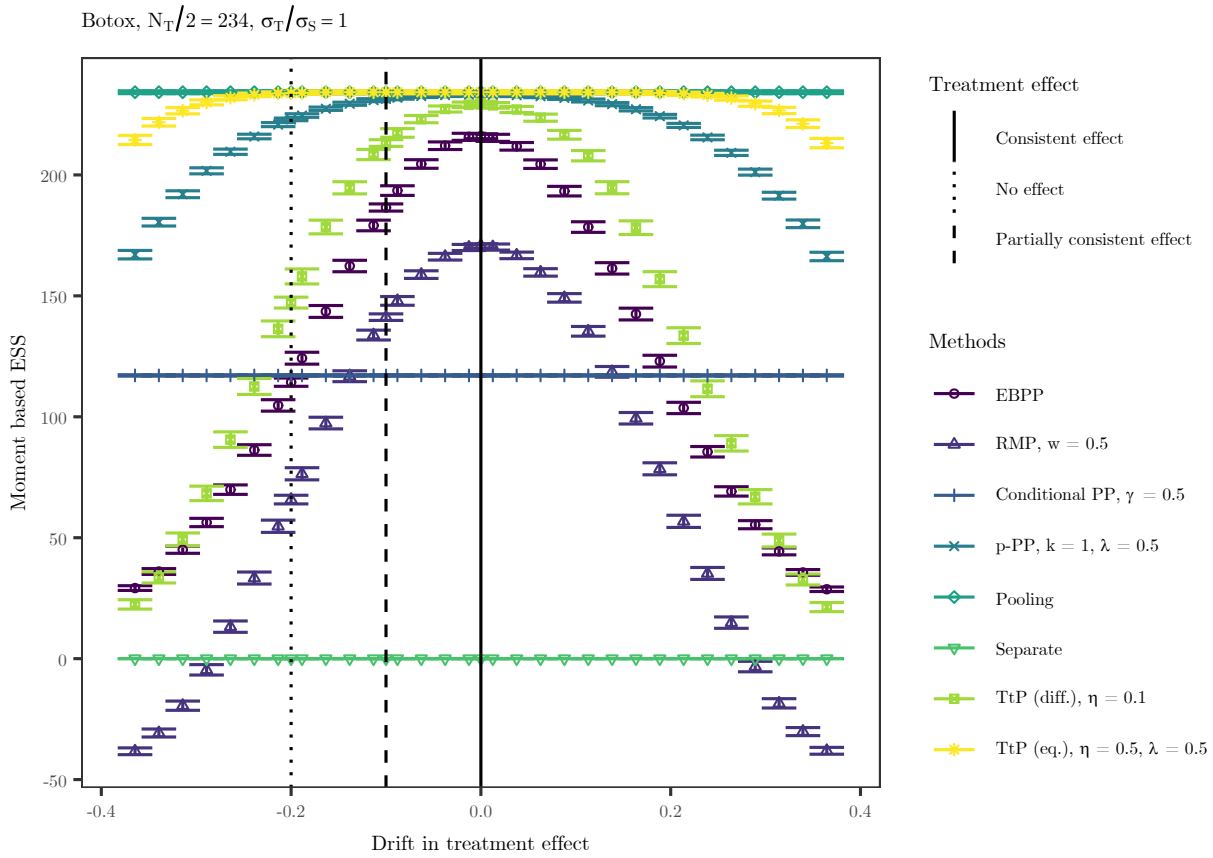

Figure 16: Moment-based ESS as a function of drift in the Botox case study with 234 patients per arm.

in these cases when interpreting the ESS. In theory, the ESS should not depend on the drift value when performing a separate or pooled analysis (see Figure 60).

11.6.2 Conditional Power Prior

The prior ESS of static borrowing methods such as the Conditional Power Prior is not affected by the characteristics of the target study (drift and target sample size per arm). The prior ESS is proportional to the borrowing strength (Figure 61), that is, in the case of the conditional power prior, proportional to the power parameter γ.

11.6.3 Normalized Power Prior

In the Normalized Power Prior, a Beta prior is put on the power parameter γ. We observed that an increase in the standard deviation of this Beta prior, σ_γ, affected differently the prior ESS (Figure 62):

- with a very small value for σ_γ, the posterior of the power parameter γ will very closely match the prior. Thefore, the drift will almost not affect the amount of borrowing. Since we set a mean $\xi_\gamma = 0.5$ for this prior, the prior ESS will be about half of the target study sample size per arm.
- With a null treatment effect, increase in σ_γ implies a decrease in the prior ESS. Indeed, this gives more flexibility for the posterior of γ to concentrate near zero, hence discarding source study data.
- With partially consistent or consistent treatment effect, increase in σ_γ implies an increase in the prior ESS. Indeed, this gives more flexibility for the posterior of γ to concentrate near one, hence borrowing more source study data.

In agreement with Pawel et al. (2023), we found that in all scenarios and all parameters considered, the Normalized Power Prior always discounts source data (with values of the upper bound of the 95% CI on the

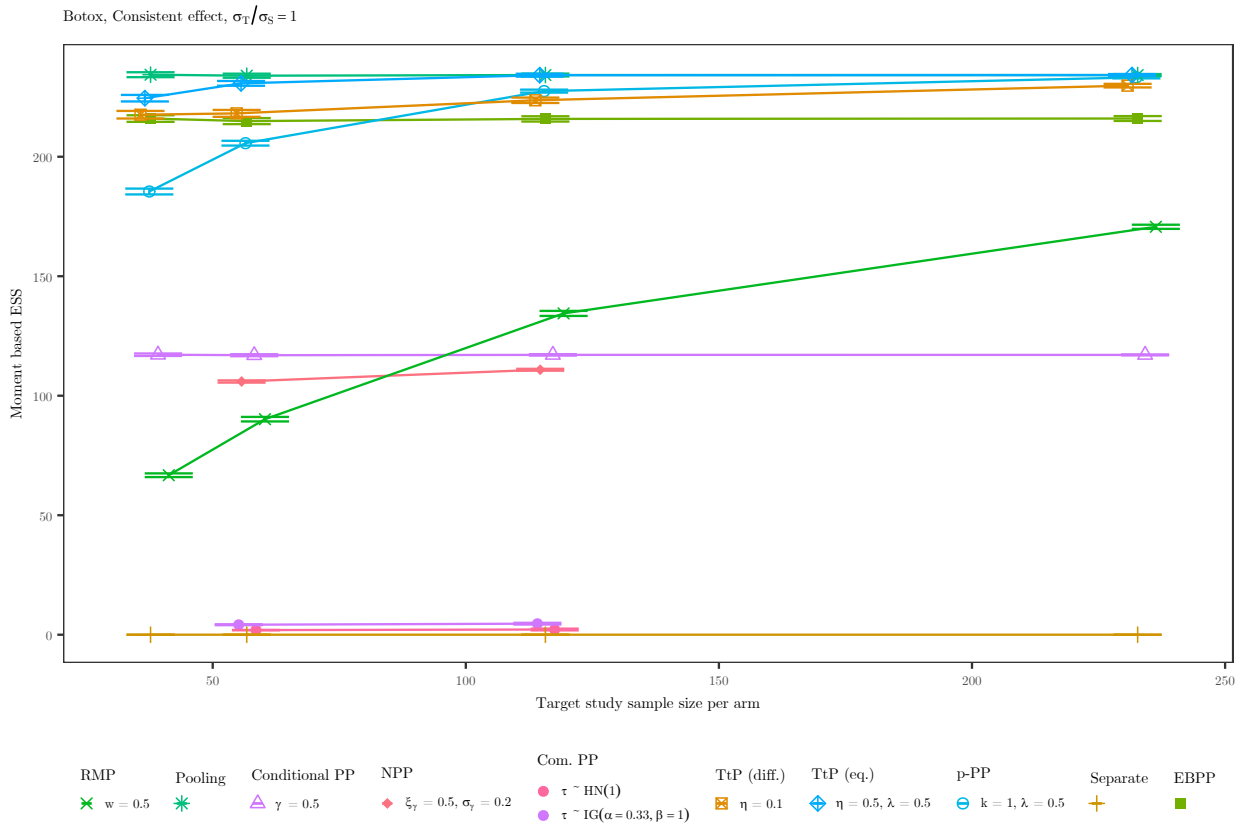

Figure 17: Moment-based ESS as a function of the sample size per arm in the Botox case study with a consistent treatment effect.

prior ESS way below the source study sample size per arm), even in the absence of drift. Indeed, to fully borrow information from the source study, the posterior on the power parameter γ must be concentrated near 1. However, with a beta prior on the power parameter that puts significant weight on values below 1, and given the limited amount of data, the posterior will also tend to put some significant weight on values of the power parameter gamma below 1 (Figures 63 and 64). The posterior distribution of the treatment effect implies integrating over gamma in the range [0,1]. Since significant mass of the distribution of gamma is away from 1, this results in a prior that never fully borrows information from the source.

In our case, with limited sample size of the target data. the power parameter puts significant weight on values away from 1.

Moreover, an increase in σ_γ also induced more variability in the prior ESS from one replicate to another, as can be seen by looking at the width of the 95% CI on the different prior ESS measures. Indeed, a larger standard deviation for the prior on the power parameter implies that the amount of borrowing will vary based on random fluctuation of the observed drift between source and target data.

These interpretations are confirmed by looking at the posterior mean of the power parameter distribution as a function of drift, for different values of the prior standard deviation σ_γ (Figure 63). The posterior mean of the distribution of the power parameter γ decreases as drift departs from zero. This translates into discarding external information. This discarding behavior is strongly dependent on the hyperprior standard deviation. The most conservative choice, corresponding to large values σ_γ, leads to a much stronger adaptive behavior as can be seen by a rapid decrease on the posterior mean of the power parameter distribution.

If we consider the posterior standard deviation of the power parameter distribution as a function of drift, for different values of the prior standard deviation σ_γ (Figure 64), we observe a more complex pattern: for drift values largely deviating from zero, the standard deviation of the power parameter distribution goes towards zero, meaning that the posterior of the power parameter concentrate near zero. This behavior is not noticeable, however, for very small prior σ_γ, translating into a lack of adaptibility. For large prior σ_γ, however, we see a slight decrease in the posterior value of σ_γ, the standard deviation of the posterior

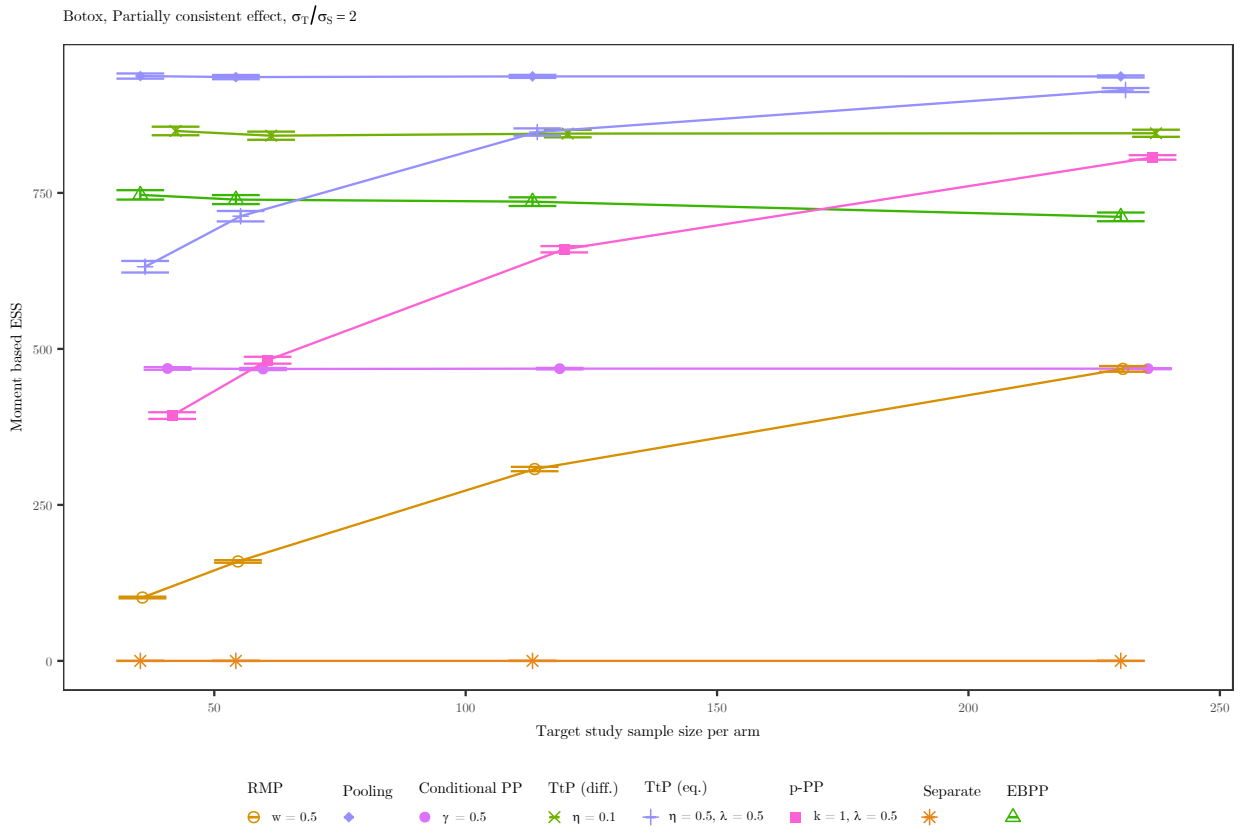

Figure 18: Moment-based ESS as a function of the sample size per arm in the Botox case study with a partially consistent treatment effect and $\sigma_T/\sigma_S = 2$.

distribution of γ. This implies that, despite a small drift, the posterior distribution of γ does not concentrate near larger values, but remains quite flat (considering, in particular, that the standard deviation of a Beta distribution is upper bounder by 0.5). Note also that the larger the stronger the sample size, the larger the variations for the posterior value of σ_γ with drift. However, the lack of concentration of the posterior distribution of γ is a phenomenon we observed across targe study sample size values, which emphasizes the conservative nature of this method.

The parameter σ_γ is therefore crucial for the adapting behavior of the method. The impact the standard deviation of the Beta hyperprior has on the resulting posterior can be visualized for the three main treatment effect values in Figure 64 and Figure 66. Figure 64 clearly shows that the posterior standard deviation of the distribution of γ only weakly varies with the drift, but increases approximately linearly with the prior value of σ_γ. Figure 66 demonstrates the increased adaptibility of the method with the prior value of σ_γ.

In the illustrative case of belimumab (Figure 65), the posterior SD of the power parameter in the NPP grows linearly with its prior SD, independently from the effect scenario considered, here with a slightly smaller slope when no effect.

11.6.4 Robust Mixture Prior

Figure 67 shows the mean Moment-based ESS of the RMP as a function of the prior weight w, in the Belimumab case study. The prior weight w of the RMP directly influences the amount of borrowing, as can be seen by the increase in the prior ESS with w for consistent or partially consistent treatment effects. The moment-based ESS as we measure it can be negative, which would suggest that in case of strong discrepancy between the source and target studies, the effect of the source study is equivalent to "removing patients" in the target study (a similar phenomenon is observed for the precision-based ESS as well). However, we see that this effect is not appropriately captured as the ESS does not monotonically vary with w in the absence of treatment effect : the ESS decreases for values of w in $[0, 0.5]$, and increases for values of w in $[0.5, 1]$. We

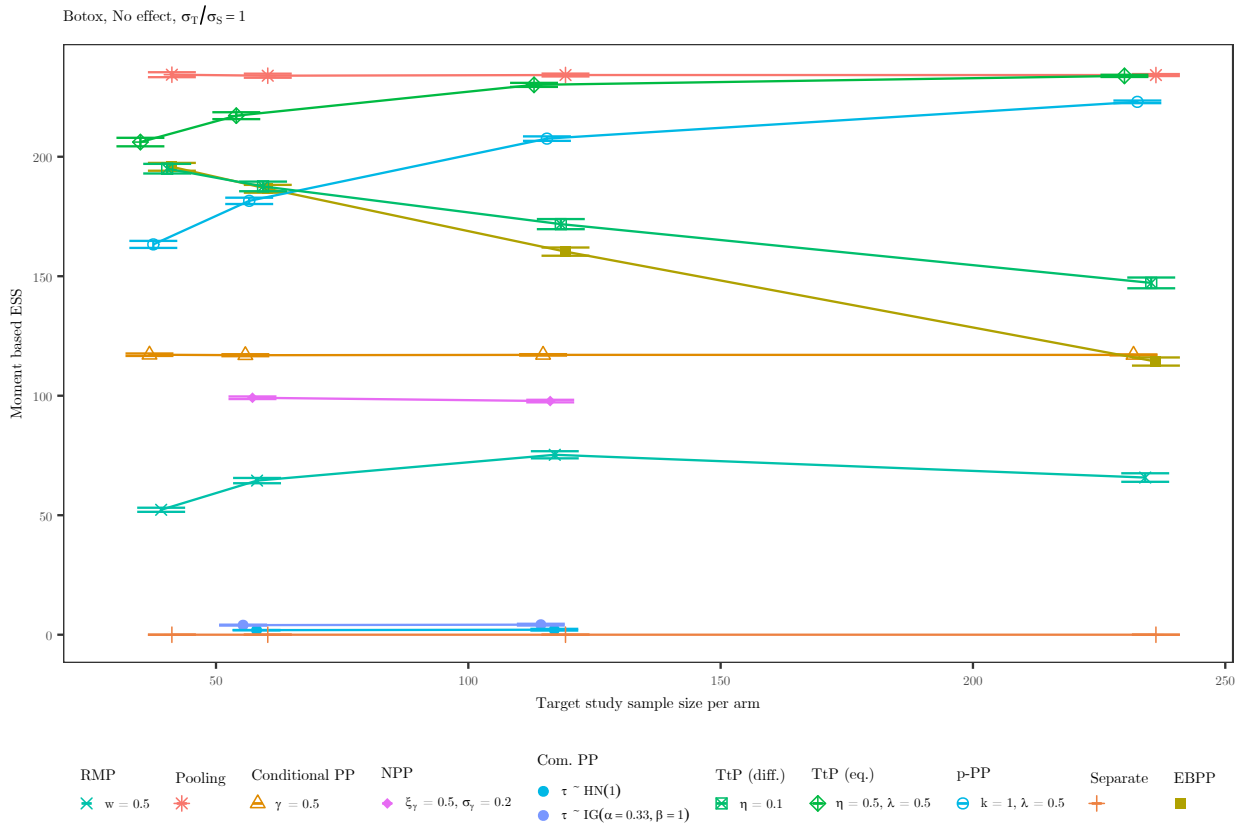

Figure 19: Moment-based ESS as a function of the sample size per arm in the Botox case study with no treatment effect.

notice that for $w = 1$, the ESS varies for different drift values, despite systematic pooling of the two studies. This is caused by the fact that the standard deviation of the generated data, used when computing the ESS, differs from one replicate to another)

We observed a strange and erratic behavior of ELIR ESS as a function of drift in treatment effect (Figure 68), with the ELIR of ESS of the RMP increasing with the drift. This may be related to the concerns we raised previously concerning the implementation of the ELIR ESS for mixture distributions in RBesT.

Figure 69 shows the mean moment-based ESS as a function of drift in treatment effect for various target sample size per arm, in the case of the RMP. Again, the plots highlight the discarding behavior as the drift increases. The speed at which external information is discarded with drift increases with the target study sample size. Similarly, the magnitude at which information is borrowed for small drift increases with he target study sample size, as more information is available that demonstrates consistency between the two studies. Note, however, an interesting phenomenon of negative moment-based ESS for moderate drift which only occurs for large sample sizes, which indicates a sample size/drift regime at which borrowing is strongly detrimental : the external information degrades inference accuracy but does not differ significantly enough to be completely discarded. Interestingly, we observed this phenomenon for all values of the prior weight w except 0 and 1.

11.6.5 Test-then-pool equivalence

Figure 70 shows the mean Moment-bases ESS of the test-then-pool method with an equivalence test as a function of drift, for various parameters. For clearer visualization, Figure 71 and 72 show, in the same scenario, the mean Moment-based ESS as a function of either the equivalence margin λ or the significance threshold η, respectively. We see that small values of the equivalence margin lead to systematically rejecting the source data, unless the significance threshold is largely increased. λ and η have similar effects of increasing the amplitude of borrowing while decreasing sensitivity to drift.

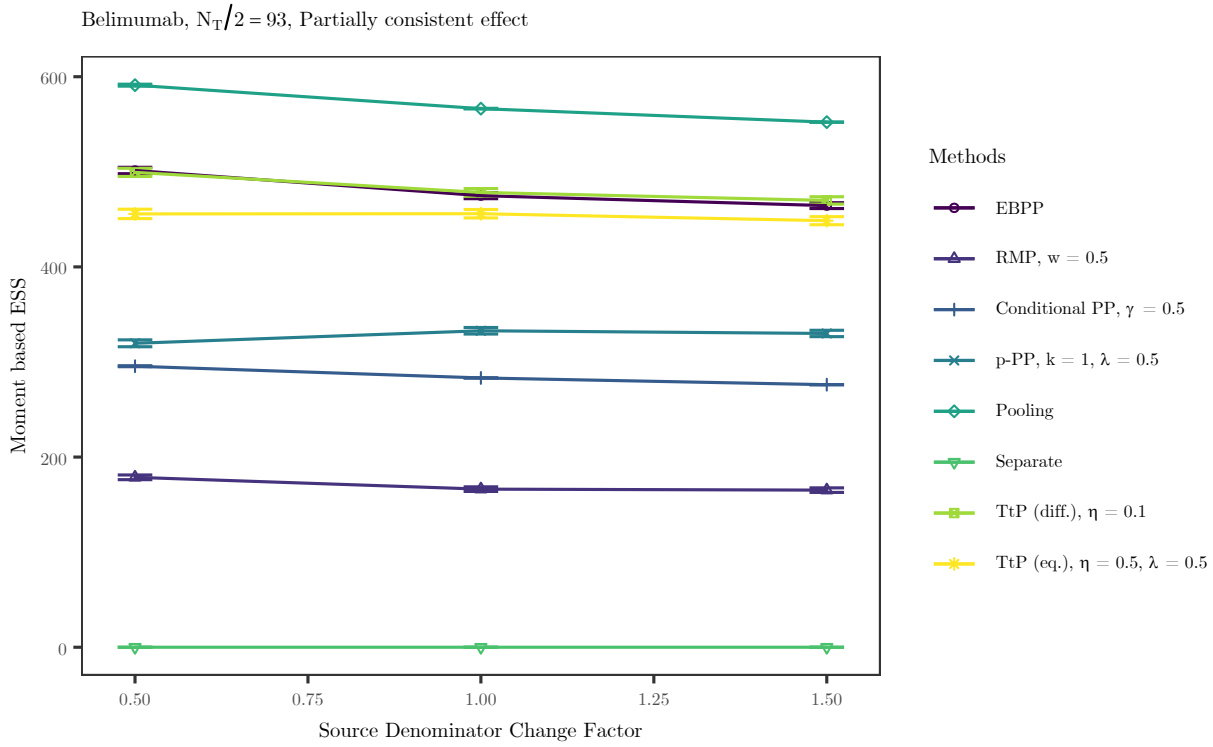

Figure 20: Moment-based ESS as a function of changes in the denominator of the summary measure in the source study in the Belimumab case study with 93 patients per arm.

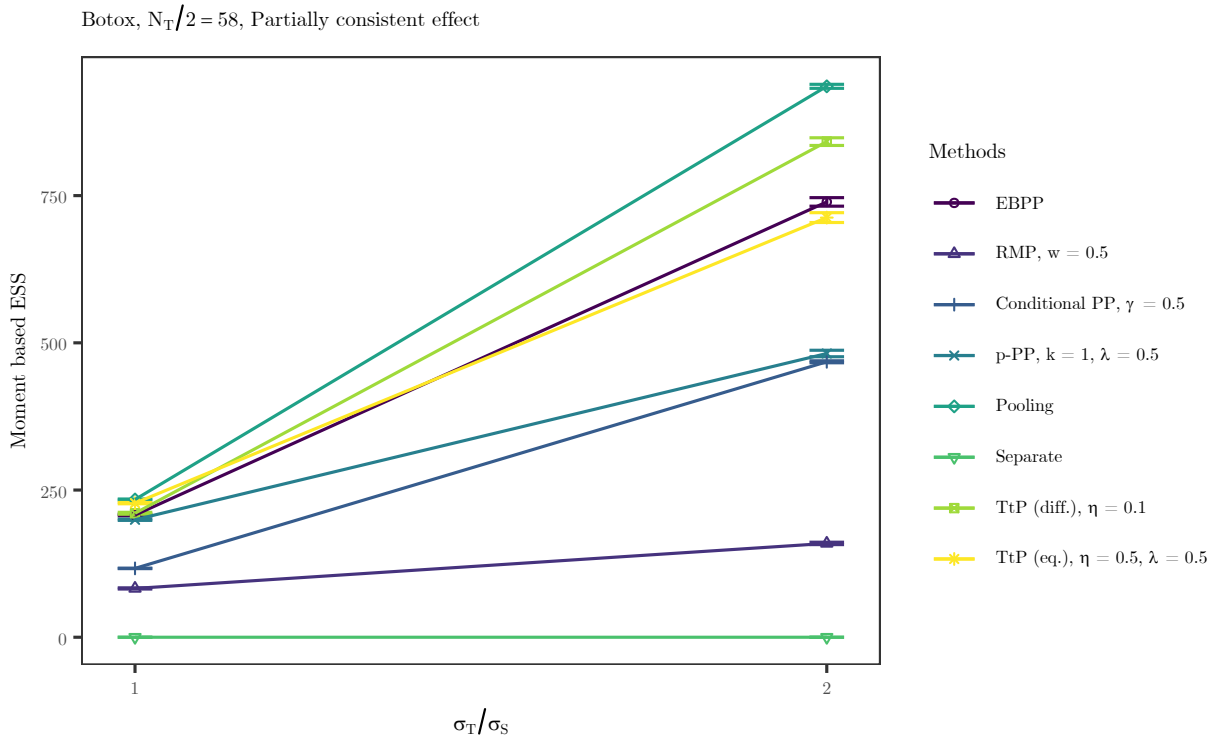

Figure 21: Moment-based ESS as a function of the ratio between the target and source studies standard deviation, with a partially consistent treatment effect.

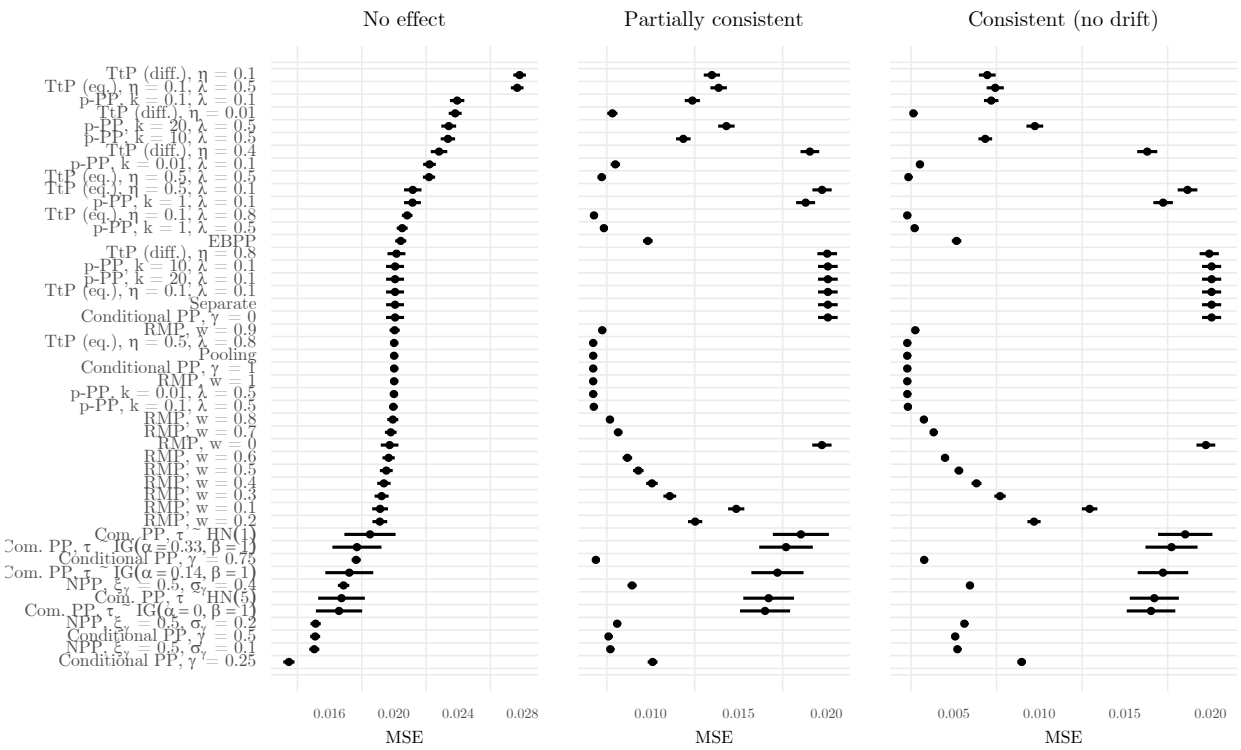

Figure 22: Comparison of the Mean Squared Error (MSE) of the different methods and associated 95% CI, for the three main treatment effects considered, in the Botox case study with 117 samples per arm in the target trial.

Note that even if the curves are smooth in these plots as they represent averages over simulation replicates, for individual replicates the source data are either pooled or not, which is not the case in other adaptive-borrowing methods.

11.6.6 Test-then-pool difference

Figure 73 shows the mean Moment-based ESS of the test-then-pool method with a test for the difference.

The behavior of the method is very similar to the one of the test-then-pool method with a test for equivalence: as the significance threshold increases, the null hypothesis of absence of difference between the source and target data tend to be more frequently rejected, resulting in less frequent borrowing. The motivation for using an equivalence test instead of a test for difference stems from the idea that a small sample size would lead to large p-values in the test for difference, hence leading to systematically pooling the data. However, we did not observe this phenomenon, as can be seen in Figure 74: over the quite large range of target treatment effect values considered, there was no systematic pooling whichever the value of η between 0.01 and 0.8.

11.6.7 p-value-based Power Prior

We found that similar to the test-then-pool method with a test for equivalence, an increase in the equivalence margin used in the equivalence test of the p-value based Power Prior tend increase the amount of borrowing. The shape parameter k modulates the amount of borrowing, but quite surpringly, seems to only have limited impact on the extent to which external information is discarded with drift (Figure 75). Although on average, the borrowing behavior of this method may look similar compared to the one of the test-then-pool method, information from the source can be only partially borrowed;

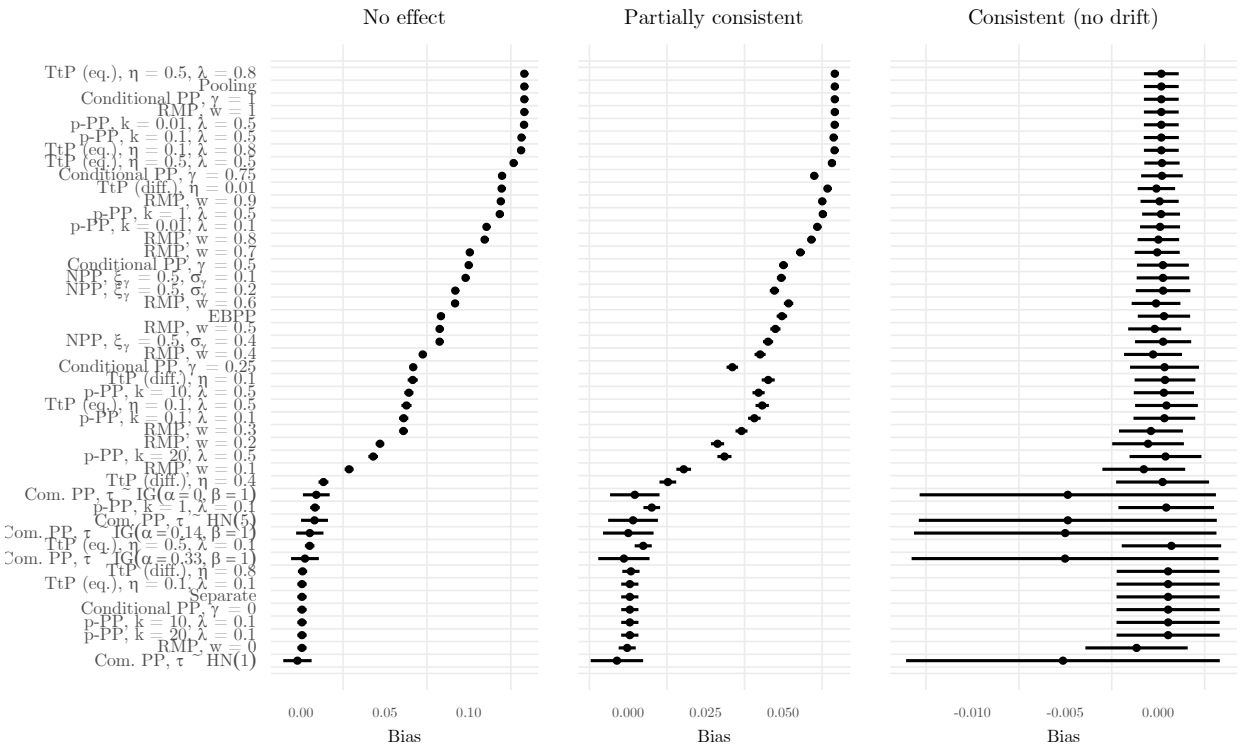

Figure 23: Comparison of the bias of the different methods and associated 95% CI, for the three main treatment effects considered, in the Botox case study with 117 samples per arm in the target trial.

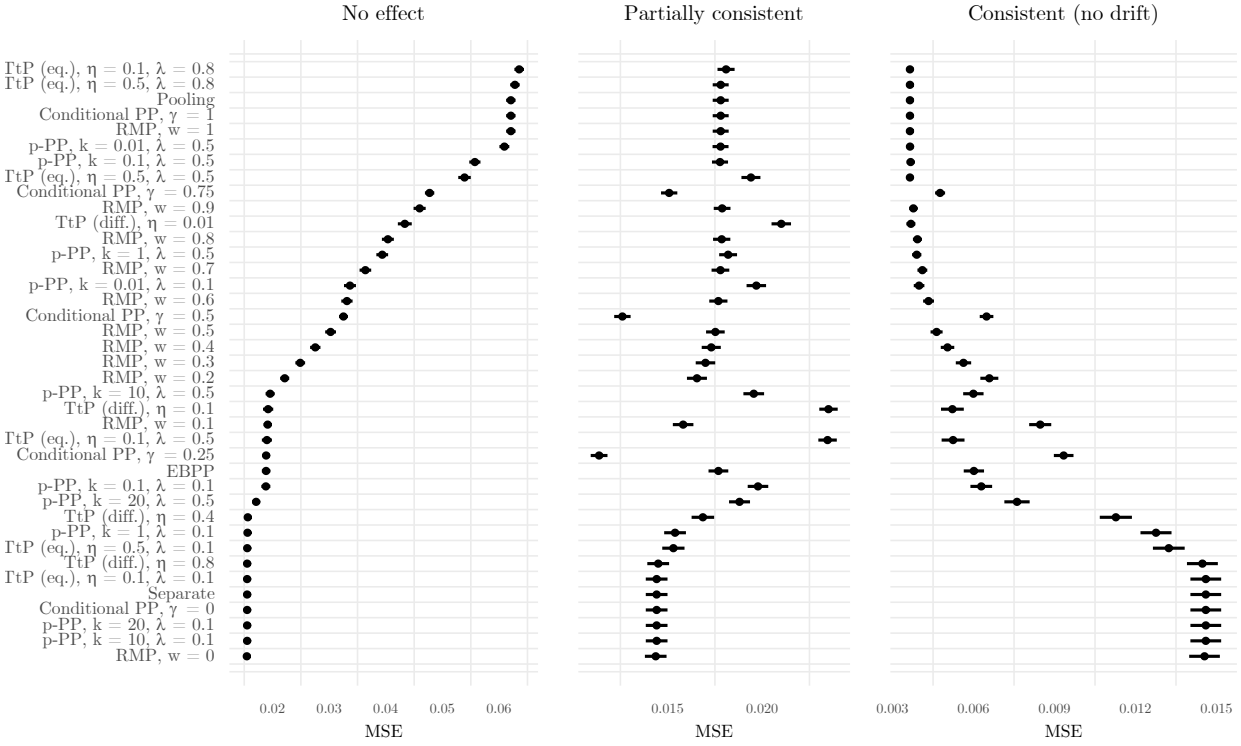

Figure 24: Comparison of the MSE of the different methods for the three main treatment effect values considered. Error bars correspond to the 95% Confidence Interval of the MSE.

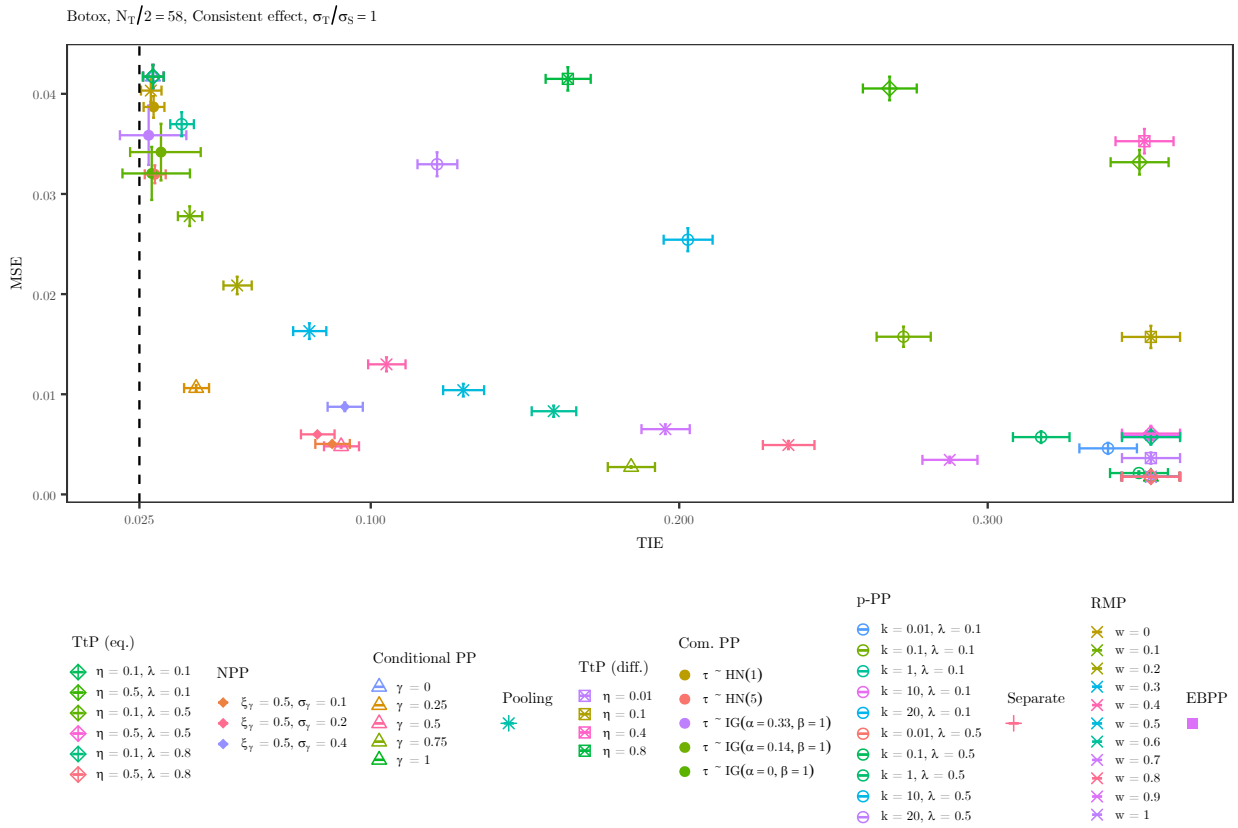

Figure 25: MSE as a function of type 1 error rate in the Botox case study with a sample size per arm of 58, across all the methods and parameters. The treatment effect is consistent, the target to source standard deviation ratio is 1. Error bars correspond to the 95% Confidence Interval of the MSE. Dashed vertical line represents the nominal type 1 error rate of 0.025.

11.6.8 Empirical Bayes Power Prior

The Empirical Bayes Power Prior does not include parameters that would allow modifying the propensity of the method to discard external information as drift increases. In Figure 76, we see that the method fully borrows information in the absence of drift, but tends not to discard external information for small target sample size arms as drift increases. This lack of conservatism can be attributed to the lack of evidence for difference between the studies for small sample sizes. However the method fully pool information in the absence of drift , irrespective of the sample size. This behavior should be compared to the one of other adaptive borrowing methods such as the RMP, which does not fully borrow information even in the absence of drift.

11.7 Impact of the use of a Gaussian approximation in the Aprepitant case study

Although we did not directly compare the OCs of a method with or without Gaussian approximation in the Aprepitant case study, we could perform an indirect comparison by comparing the relative behavior of EBPP to other methods. Indeed, in this case study, EBPP is the only method for which we used a normal likelihood. We did not observe a marked difference in terms or relative behavior compared to other case studies, suggesting that the Gaussian approximation would have limited impact overall.

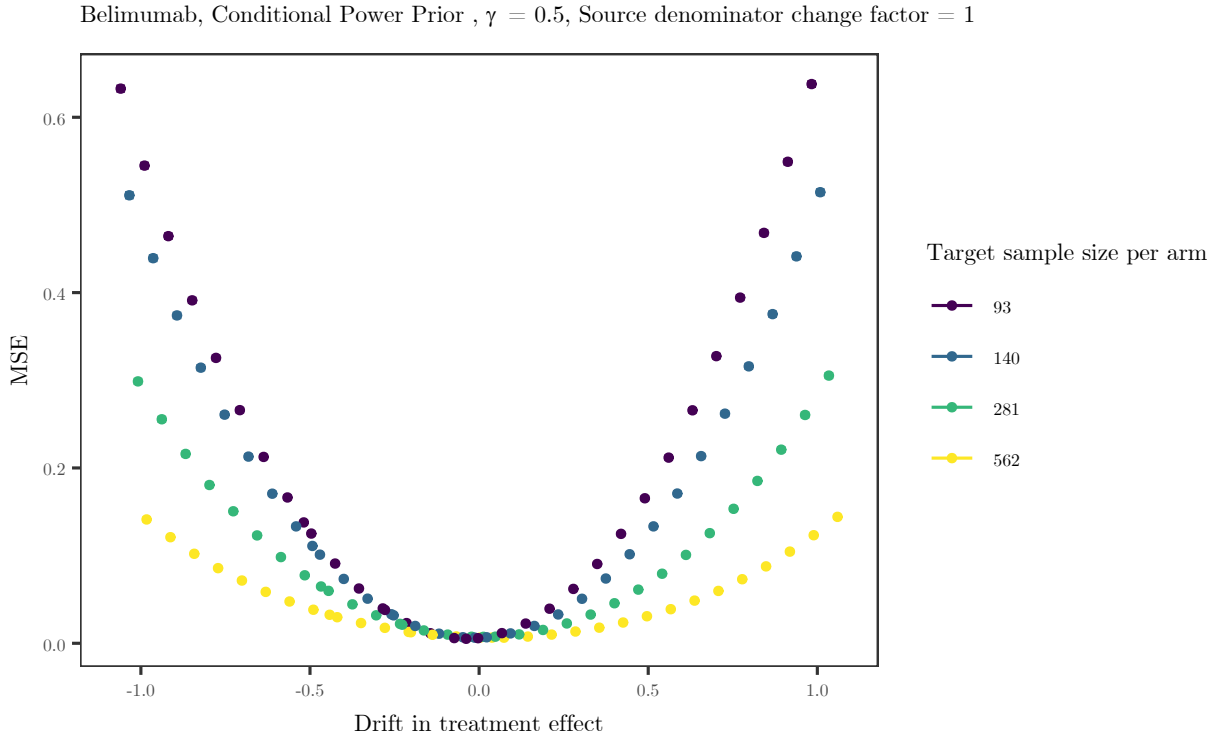

Figure 26: Mean Squared Error (MSE) of the Conditional Power Prior as a function of the drift in treatment effect, for various sample size per arm in the target study. The MSE of static borrowing methods is a quadratic function of the drift. Error bars correspond to the 95% Confidence Interval of the MSE.

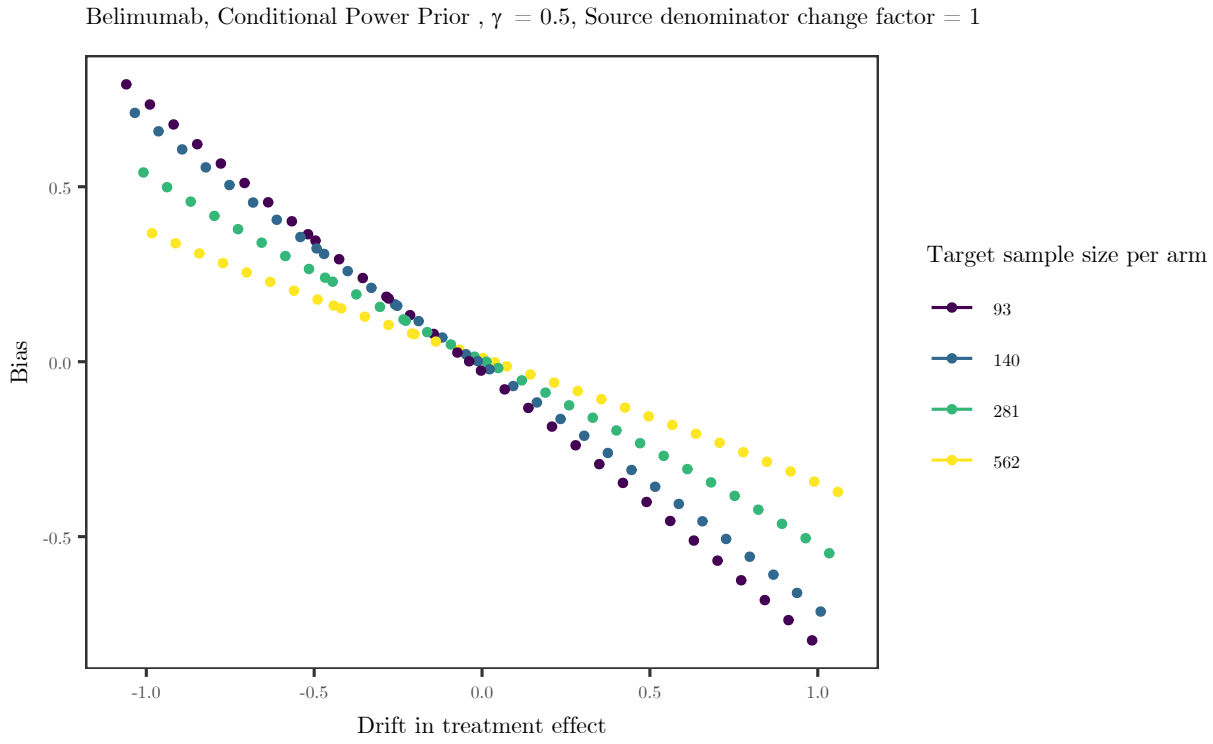

Figure 27: Bias of the Conditional Power Prior as a function of the drift in treatment effect, for various sample size per arm in the target study. The MSE of static borrowing methods is a linear function of the drift. Error bars correspond to the 95% Confidence Interval of the bias.

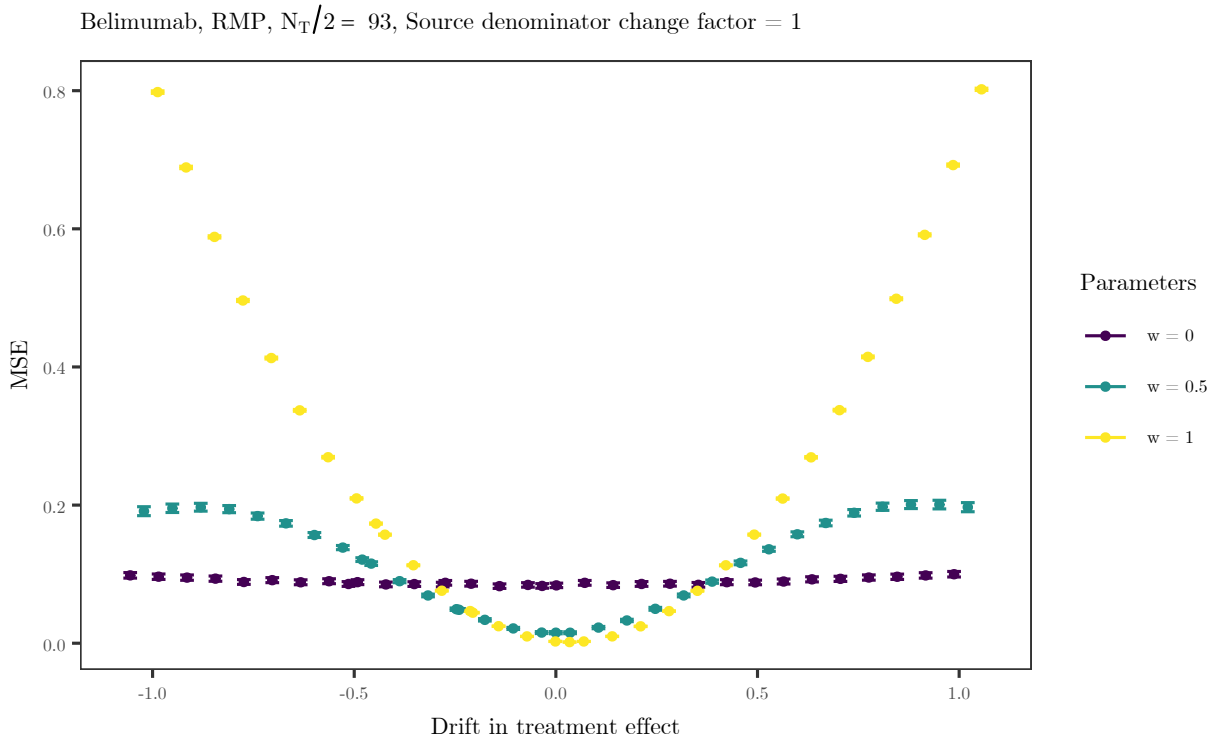

Figure 28: MSE as a function of drift for the RMP in the Belimumab case study, with a sample size per arm in the target study of 93 patients, for different values of the weight of the informative component w. Error bars correspond to the 95% Confidence Interval of the MSE.

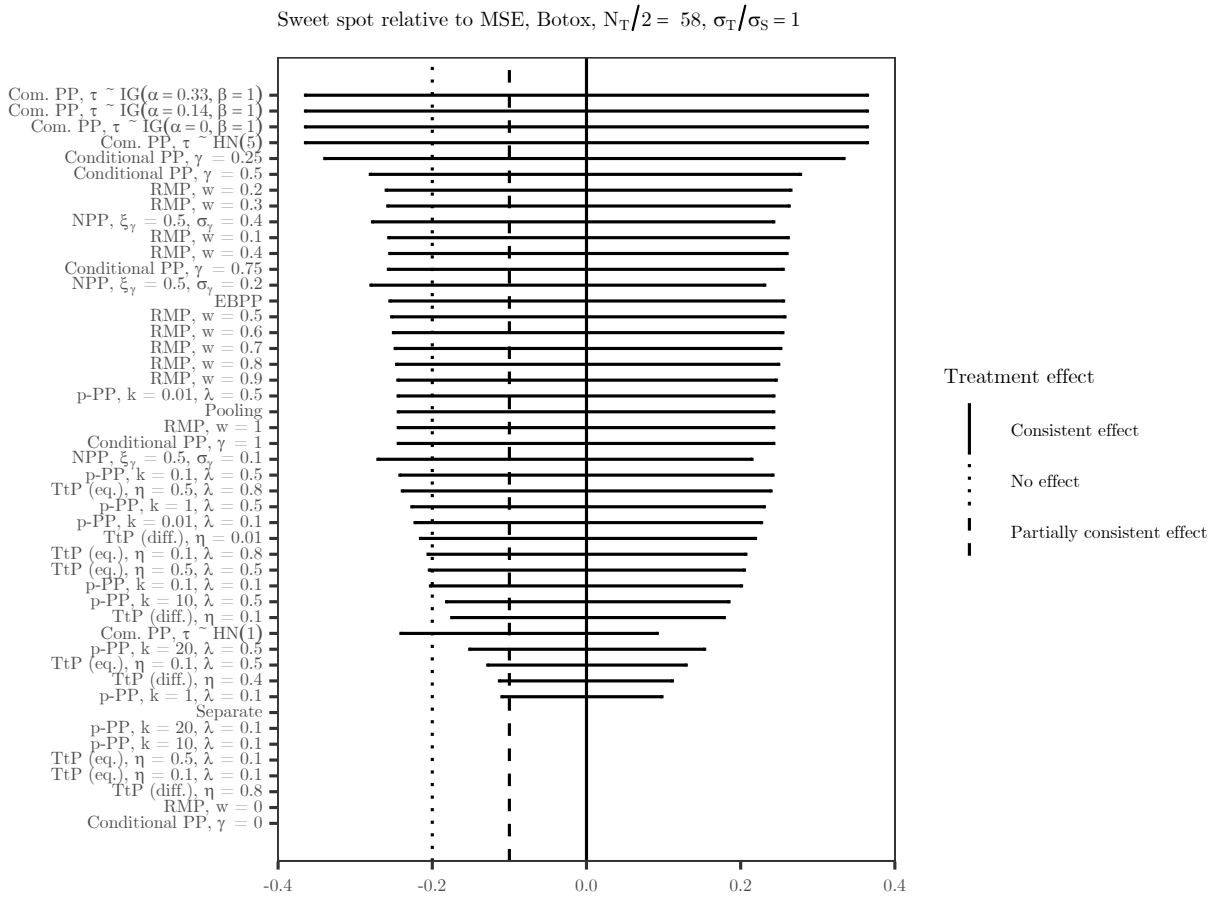

Figure 29: Sweet spot relative to MSE in the Botox case study, with a sample size per arm in the target study of 58 patients.

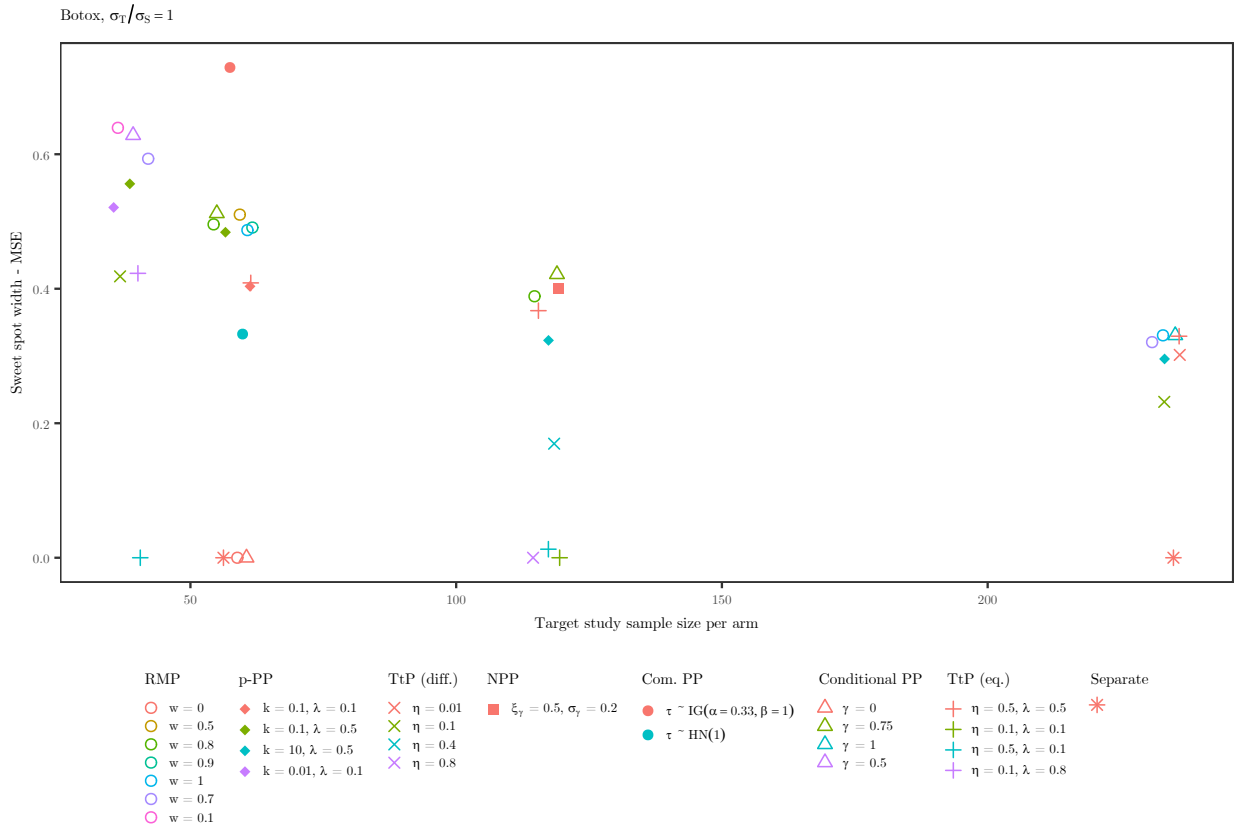

Figure 30: Sweet spot for the MSE as a function of the sample size per arm in the target study in the Botox case study.

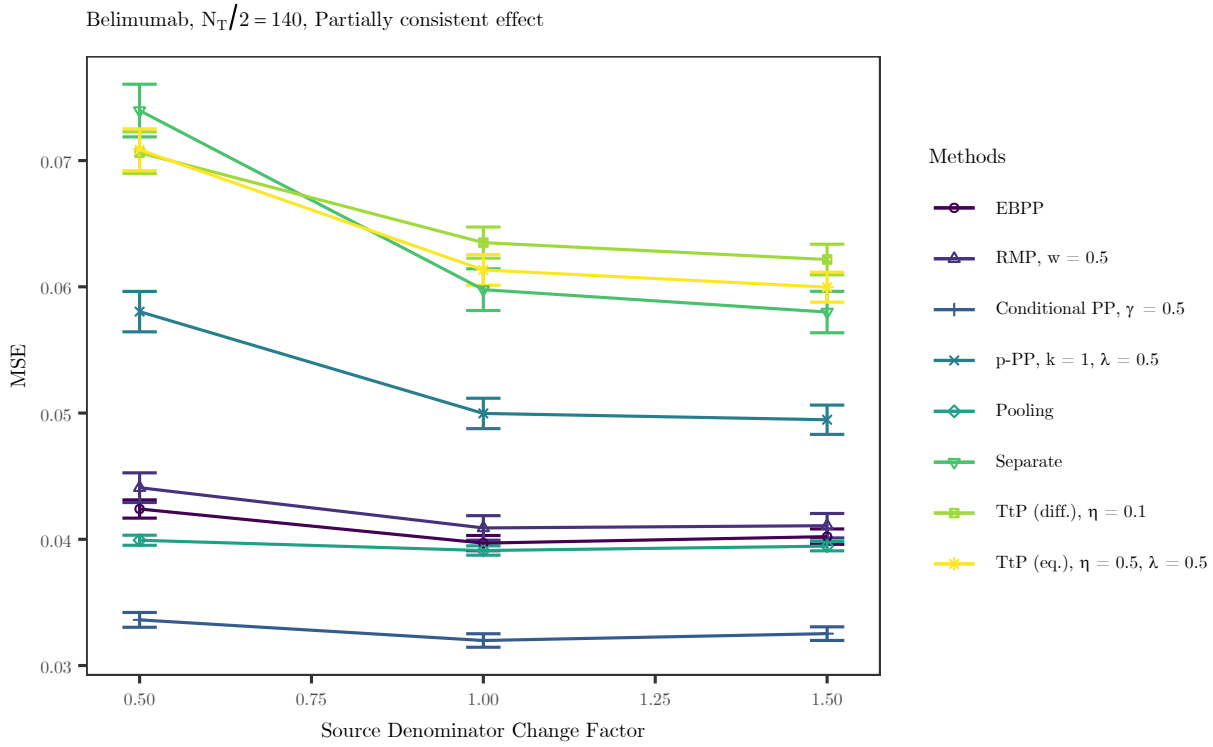

Figure 31: MSE as a function of the change factor applied to the denominator of the source study summary measure, in the Belimumab case study, with a partially consistent treatment effect.

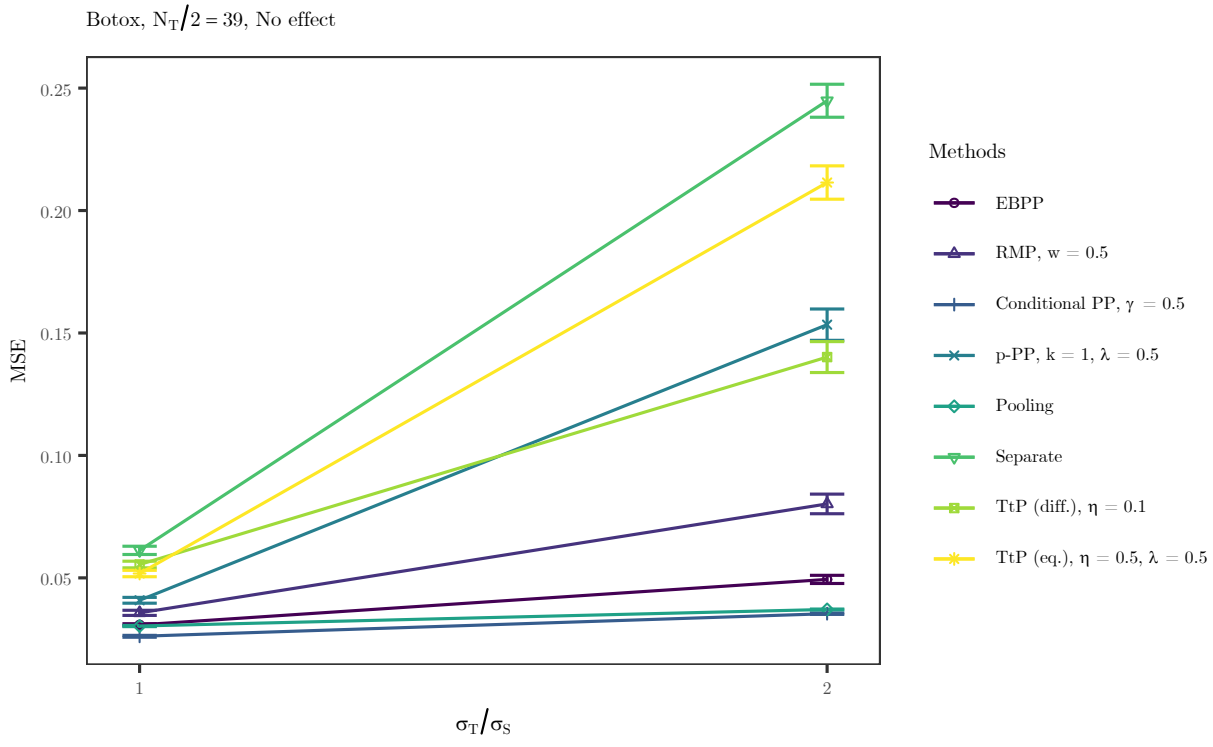

Figure 32: MSE as a function of the ratio between the target and source studies standard deviation, in the Botox case study, in the absence of treatment effect.

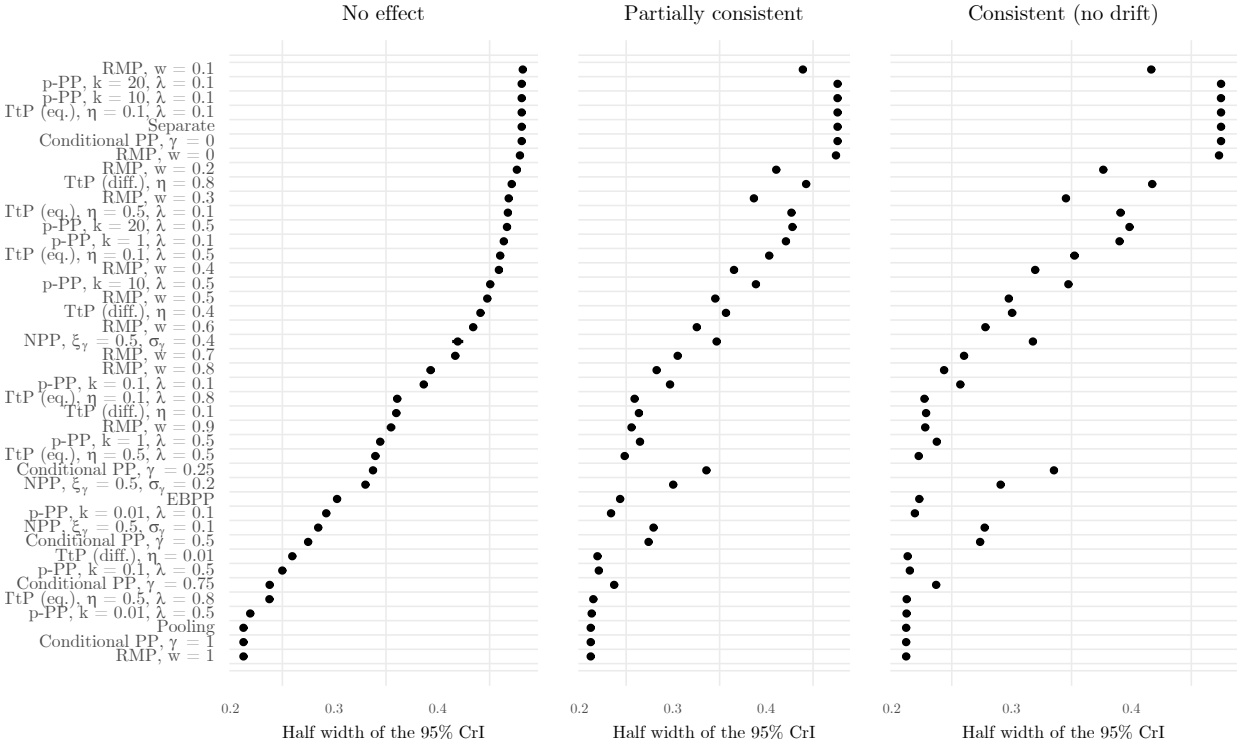

Figure 33: Comparison of the precision, measured as the mean half-width of the 95% Credible Interval, of the different methods for the three main treatment effect values considered in the Belimumab case study. Error bars correspond to the 95% Confidence Interval of the precision.

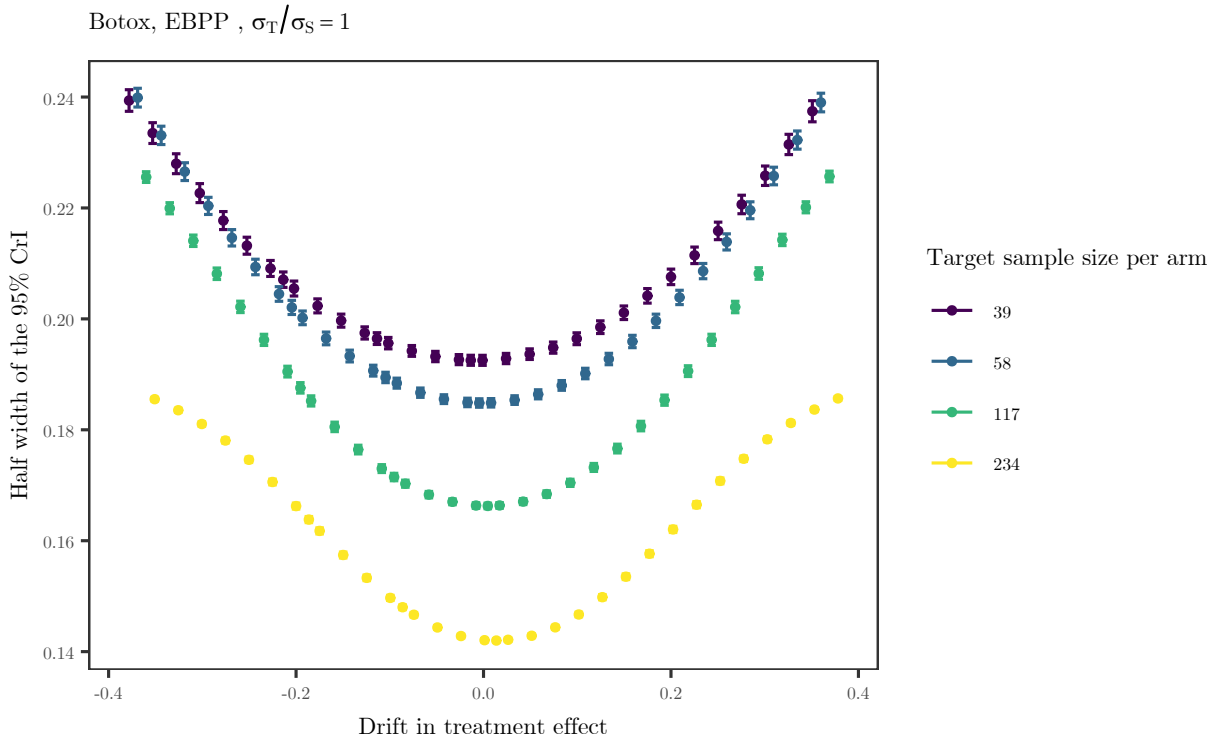

Figure 34: Precision of the EBPP, measured as the mean half-width of the 95% Credible Interval, as a function of the drift in treatment effect, for various sample sizes. Error bars correspond to the 95% Confidence Interval of the precision.

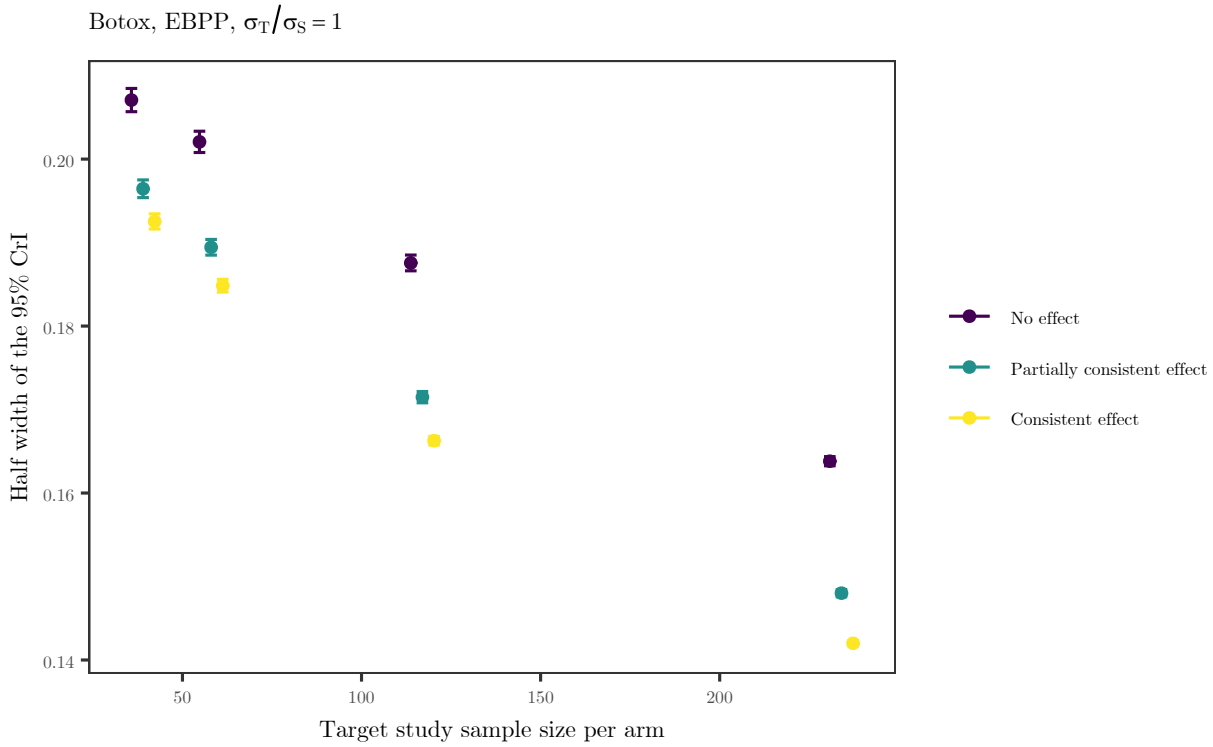

Figure 35: Precision of the EBPP, measured as the mean half-width of the 95% Credible Interval, as a function of the target study sample size per arm, for the three main treatment effect values considered. Error bars correspond to the 95% Confidence Interval of the precision.

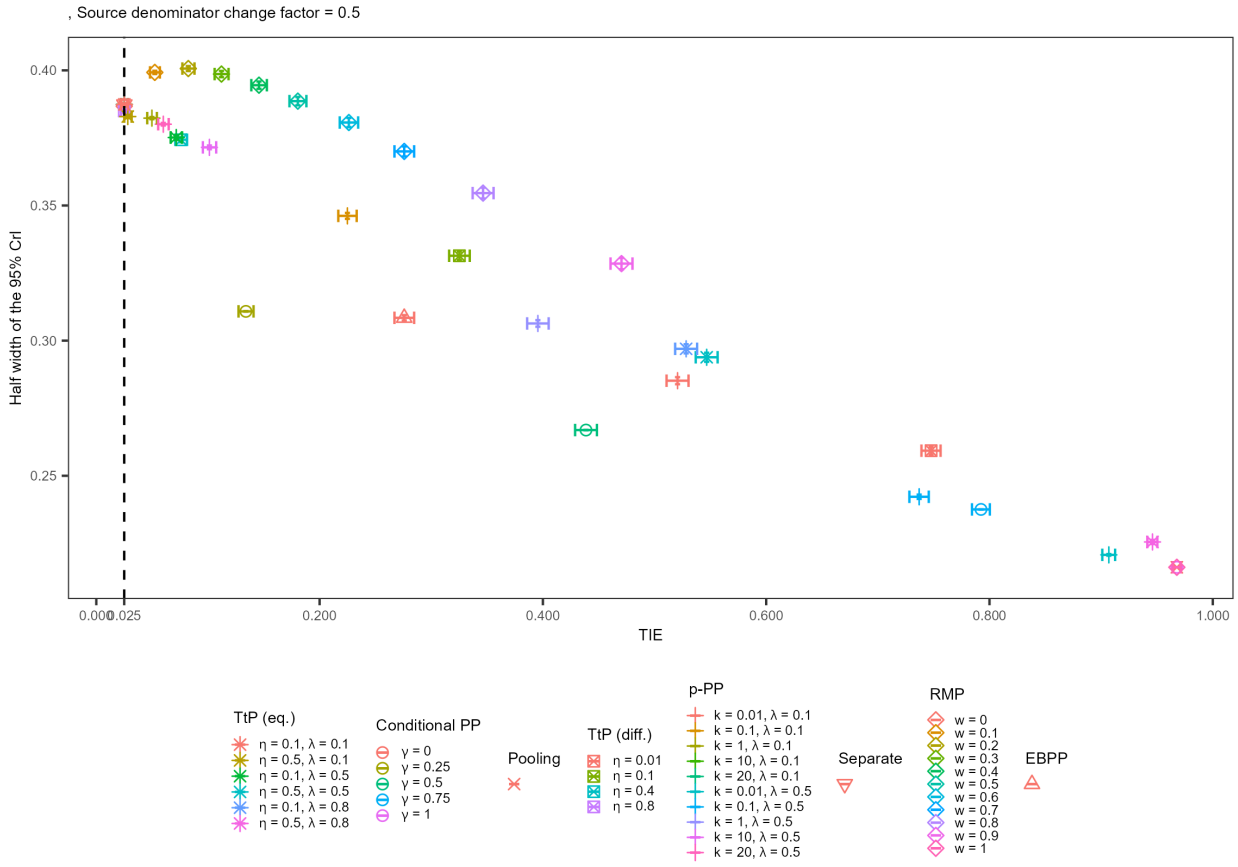

Figure 36: Precision as a function of type 1 error rate for Belimumab case study, with the source denominator change factor of 0.5, sample size arm of 281, without treatment effect, across methods

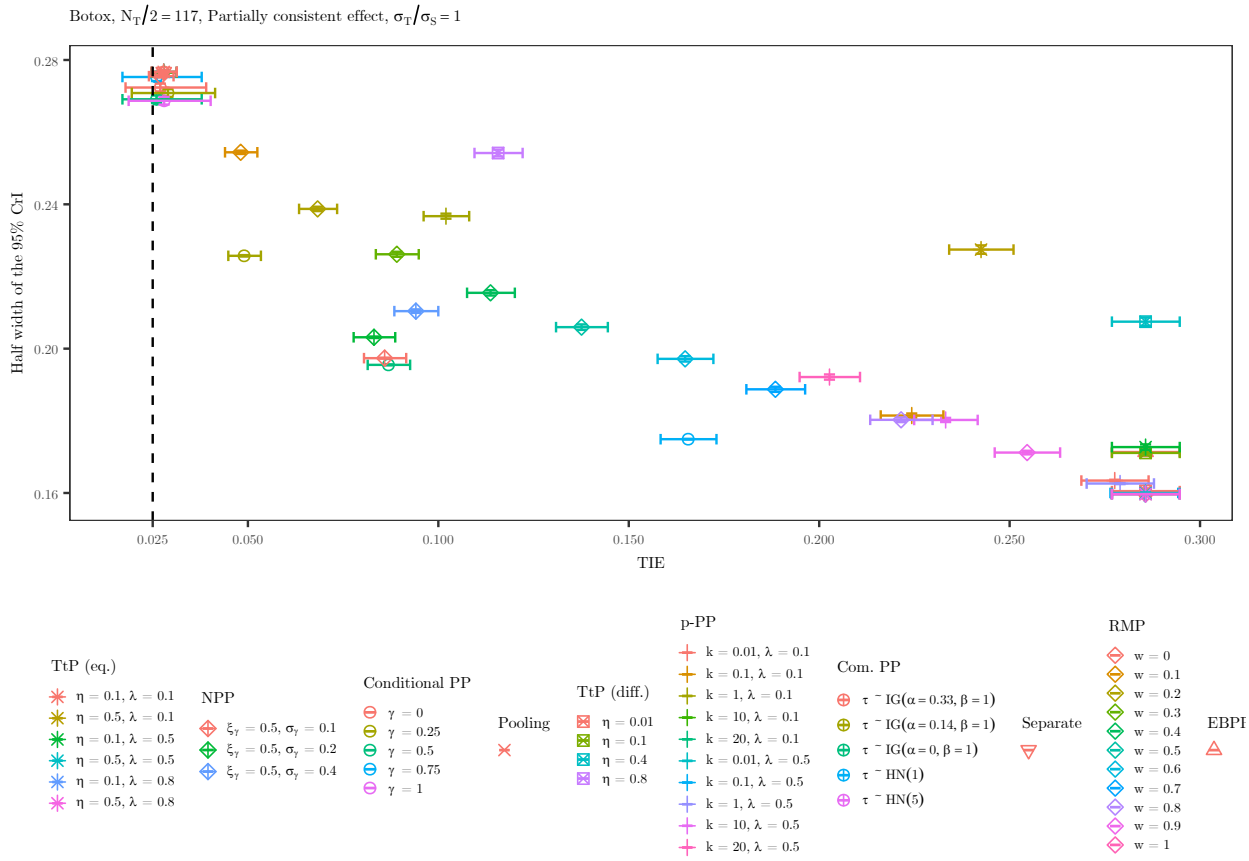

Figure 37: Precision as a function of type 1 error rate in the Botox case study with a partially consistent treatment effect sample size per arm of 58, across all methods/parameters combinations. Error bars correspond to the 95% Confidence Intervals. The dashed vertical line represents the nominal type 1 error rate of 0.025.

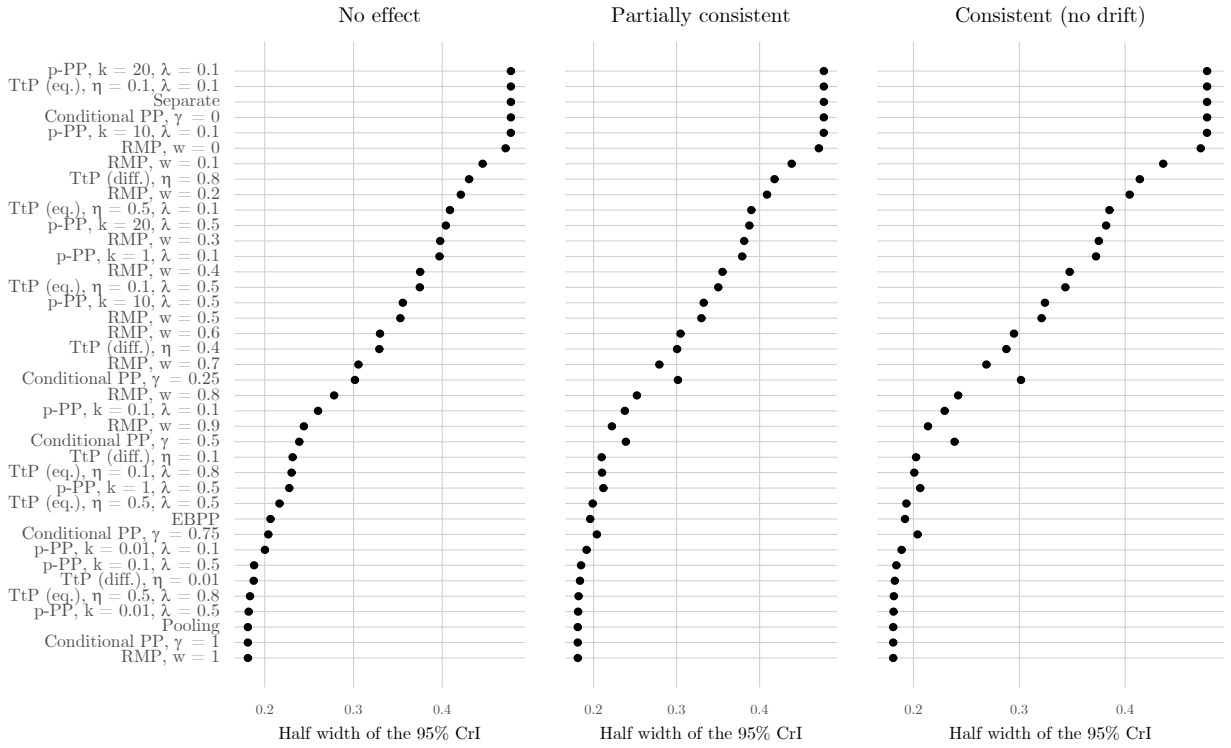

Figure 38: Comparison of the precision of the different methods for the three main treatment effects considered, in the Botox case study with 39 samples per arm in the target trial.

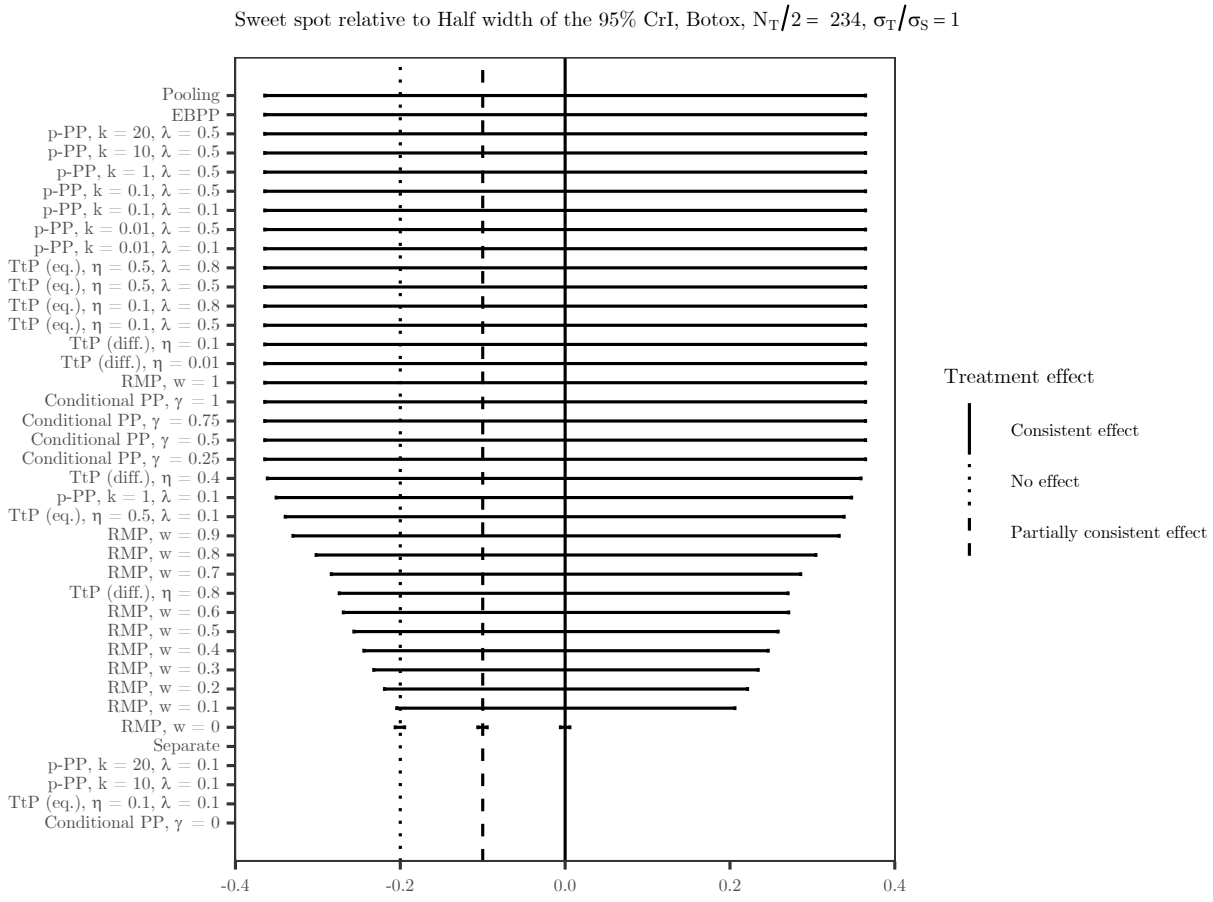

Figure 39: Sweet spot relative to precision for the different methods in the Botox case study, with a sample size per arm of 234 in the target study.

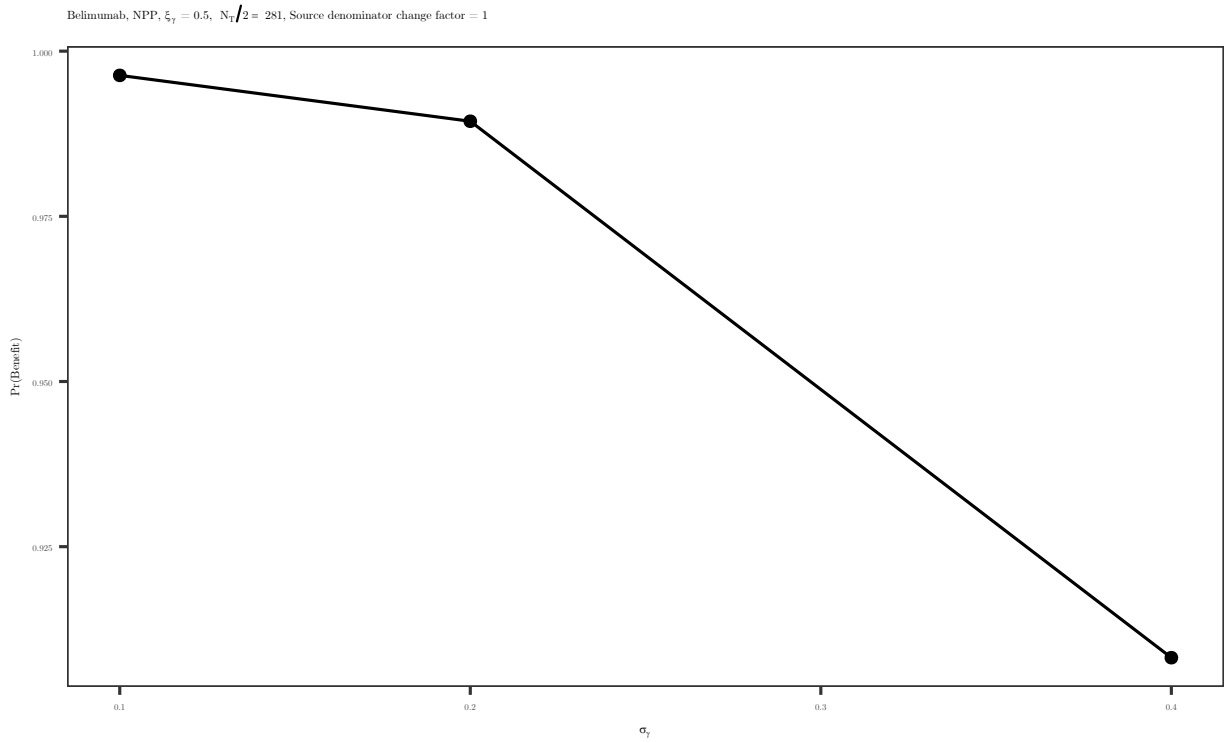

Figure 40: Probability of prior benefit as a function of the standard deviation of the hyperprior on the power parameter for the Normalized Power Prior, with a sample size of 281 without change introduced in the denominator of the source study summary measure (Belimumab case).

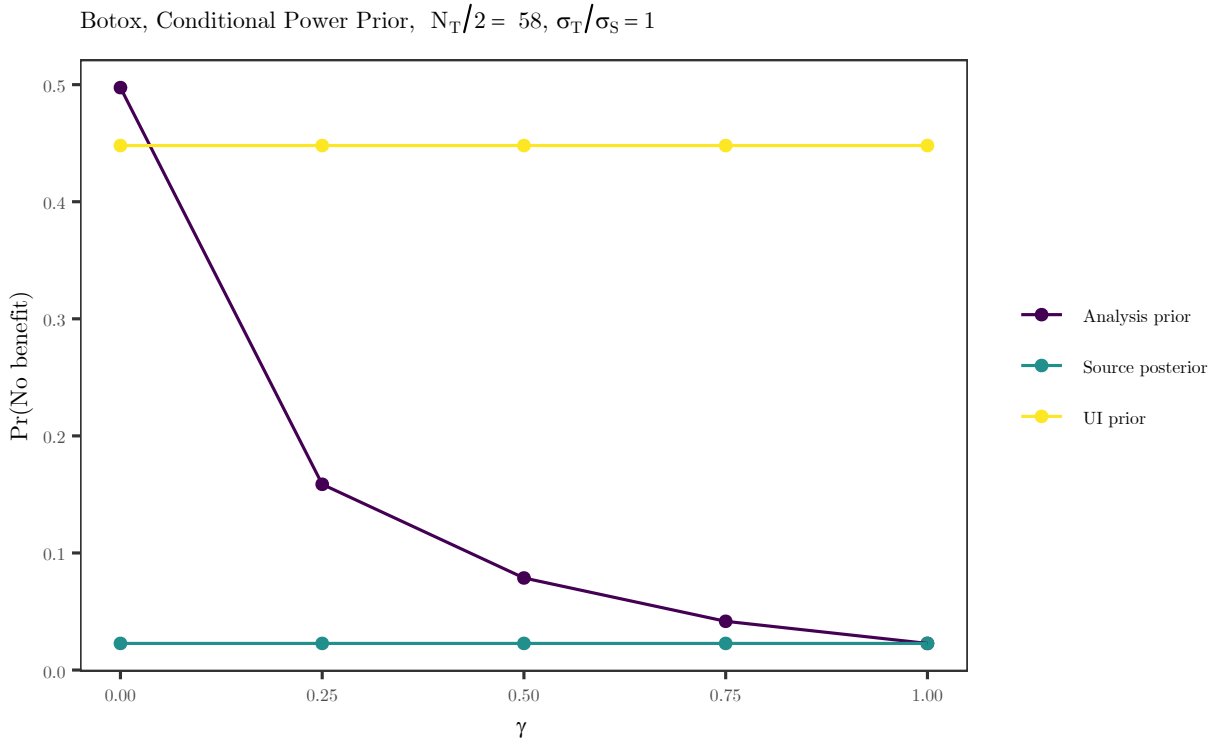

Figure 41: Prior probability of benefit as a function of the power parameter for the conditional power prior in the Botox case study.

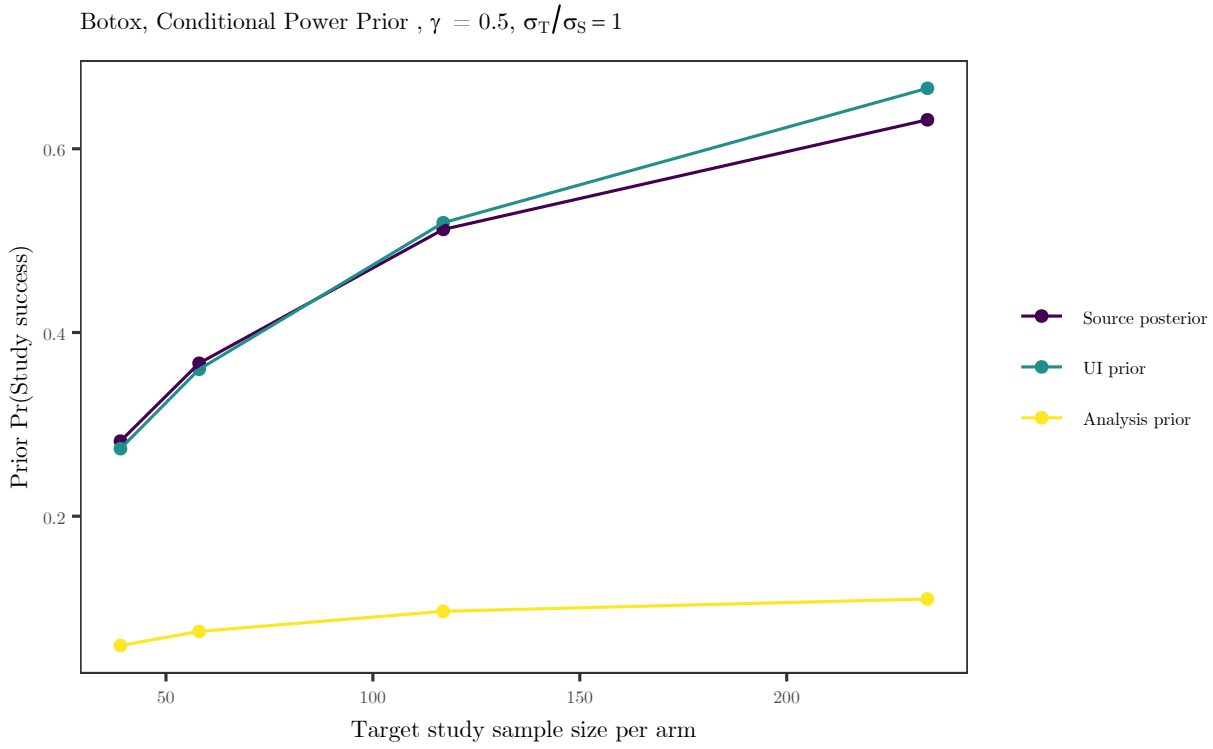

Figure 42: Prior probability of study success as a function of the sample size per arm for the conditional power prior in the Botox case study.

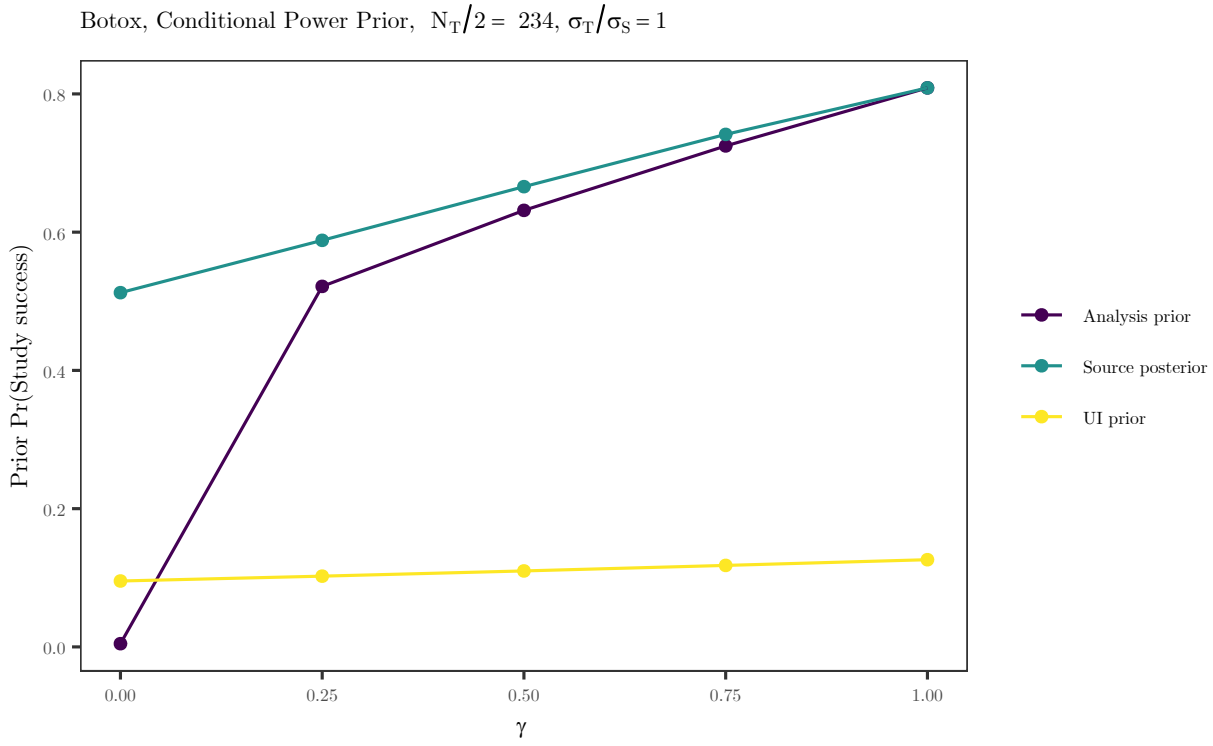

Figure 43: Prior probability of study success as a function of the power parameter for the conditional power prior in the Botox case study.

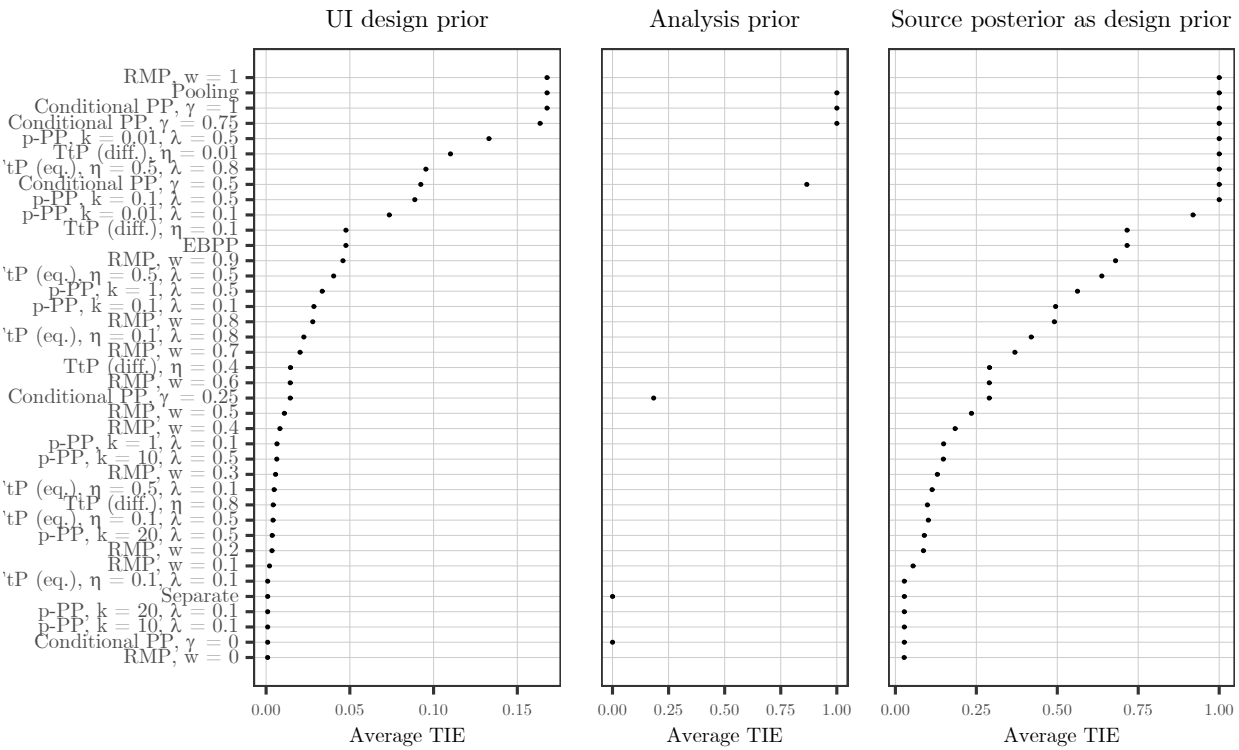

Figure 44: Bayesian TIE in the Belimumab case study with a target sample size per arm of 93.

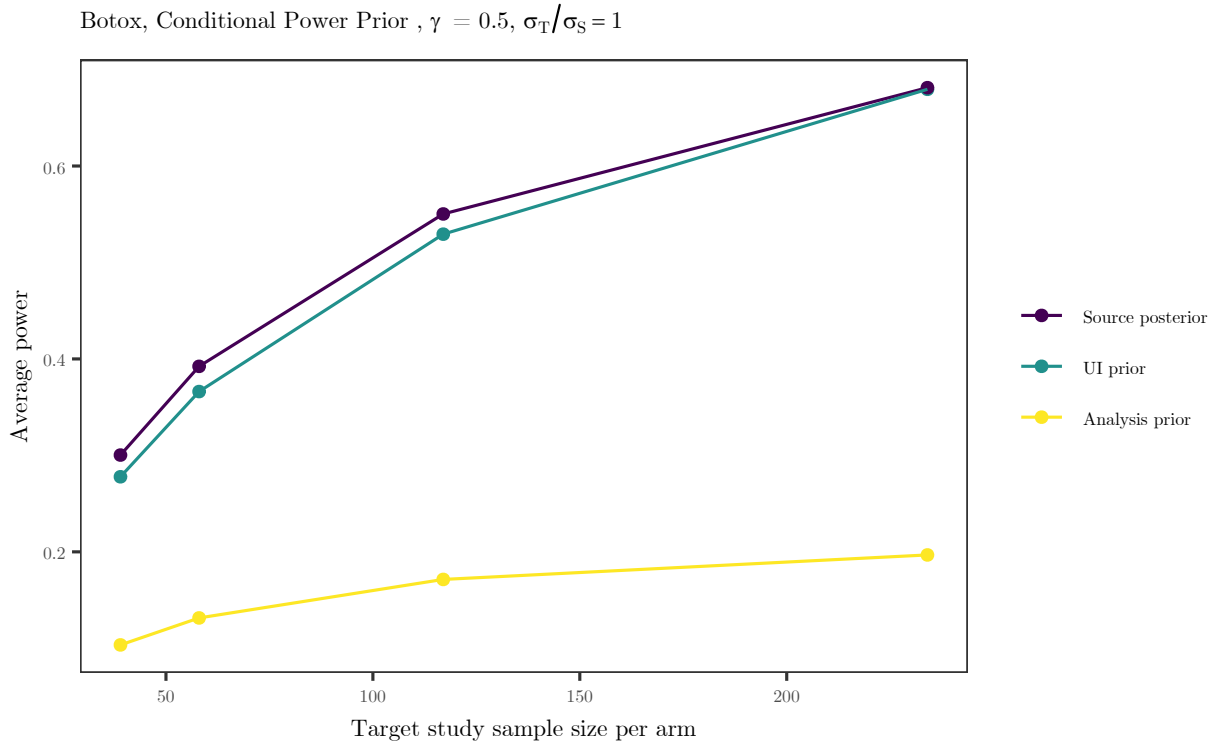

Figure 45: Average power as a function of the sample size per arm for the conditional power prior in the Botox case study

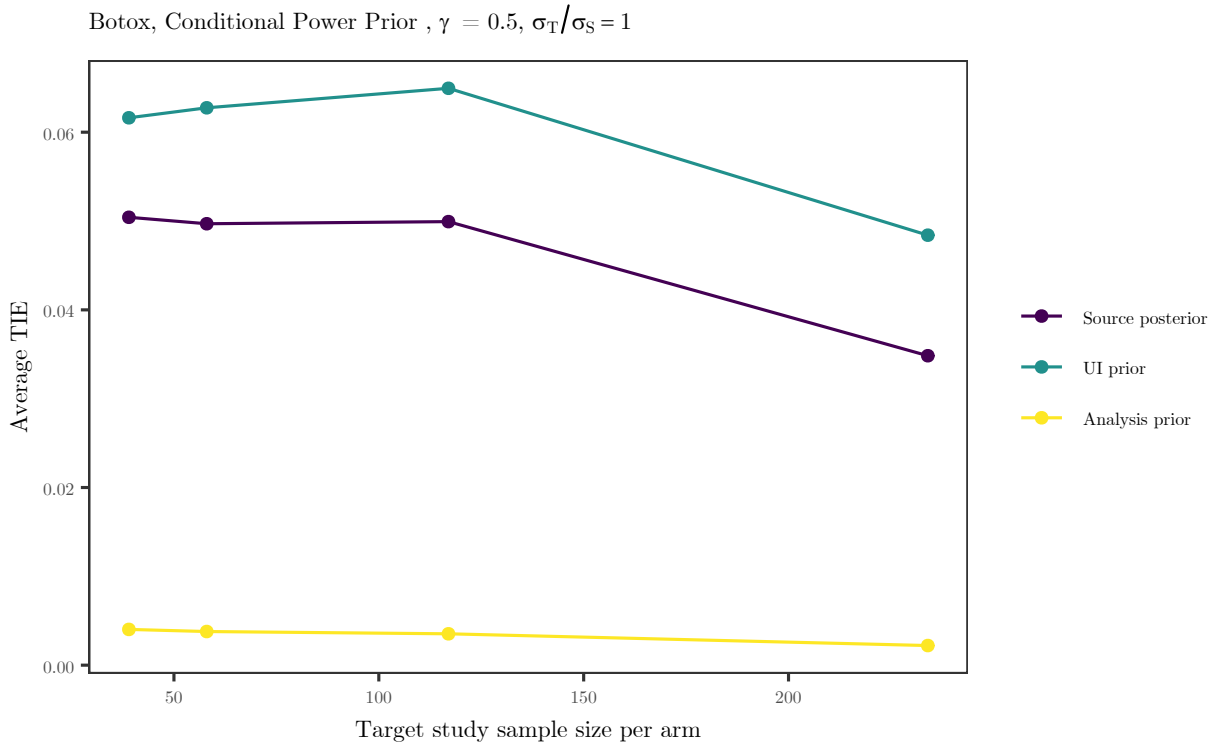

Figure 46: Average type 1 error rate as a function of the sample size per arm for the conditional power prior in the Botox case study

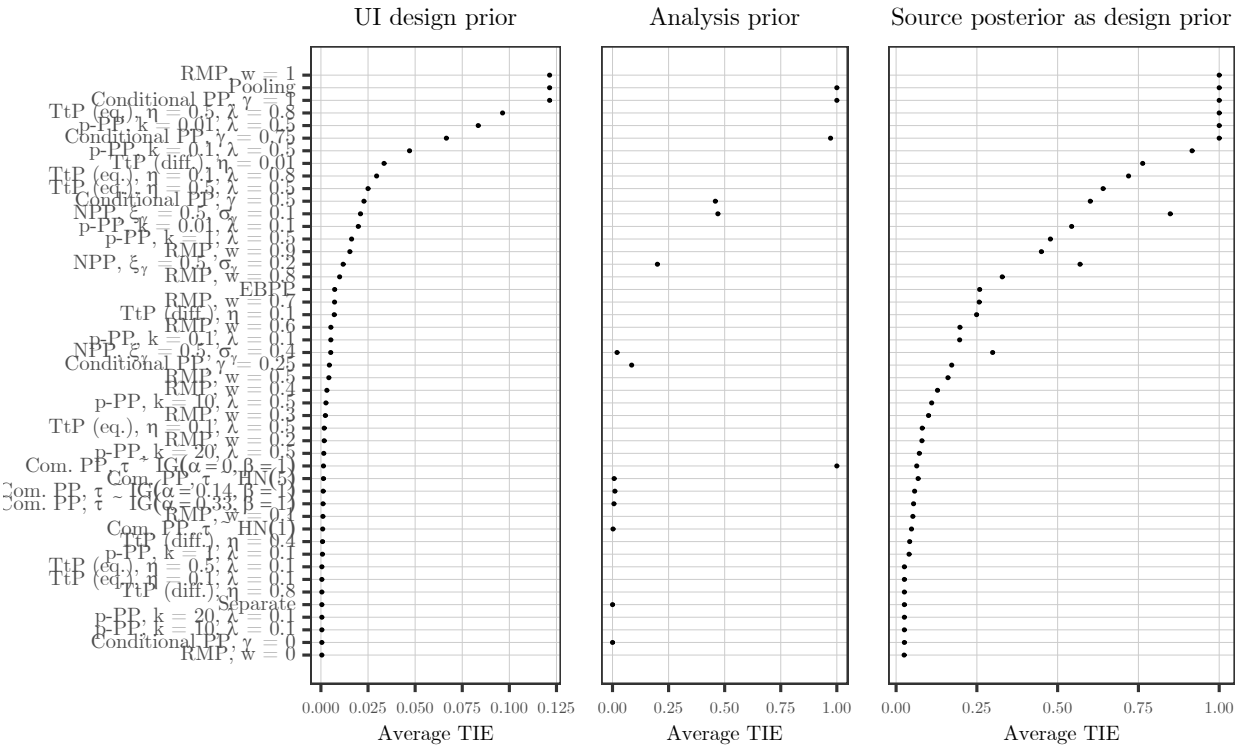

Figure 47: Average type 1 error rate of the different methods for the three different design priors considered, in the Belimumab case study with a sample size per arm of 281.

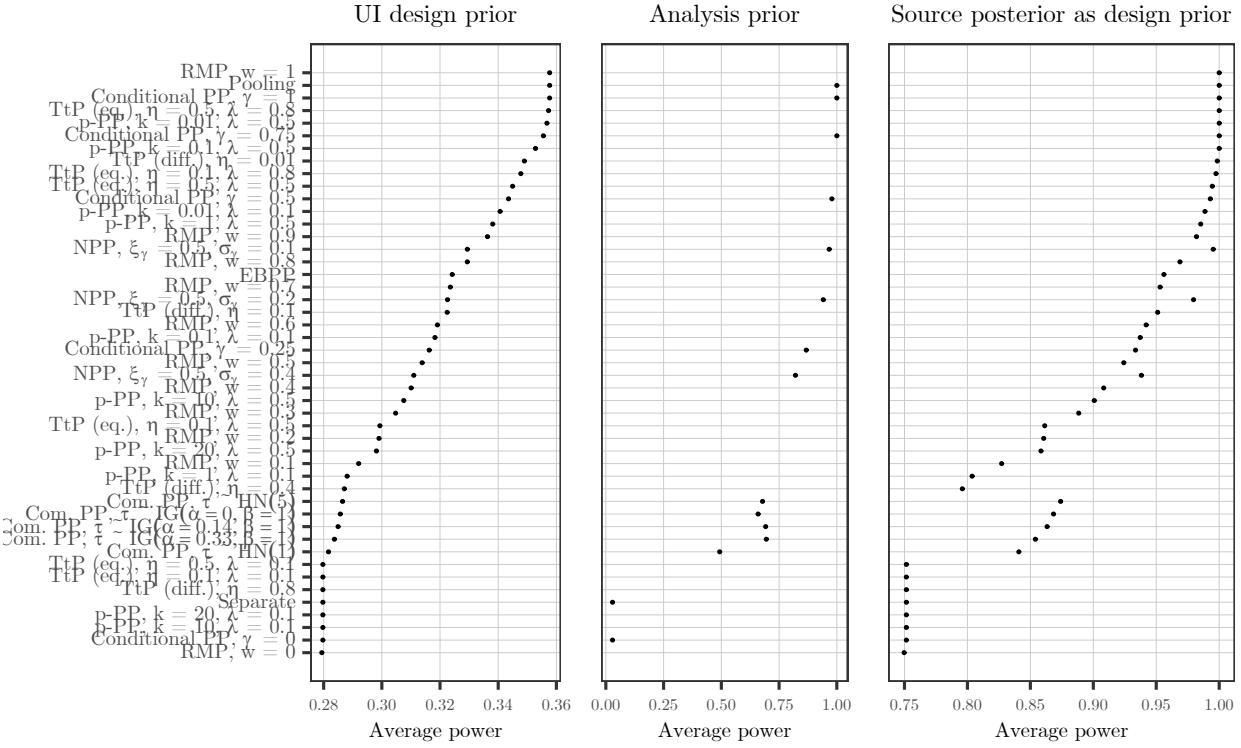

Figure 48: Average power of the different methods for the three different design priors considered, in the Belimumab case study with a sample size per arm of 281.

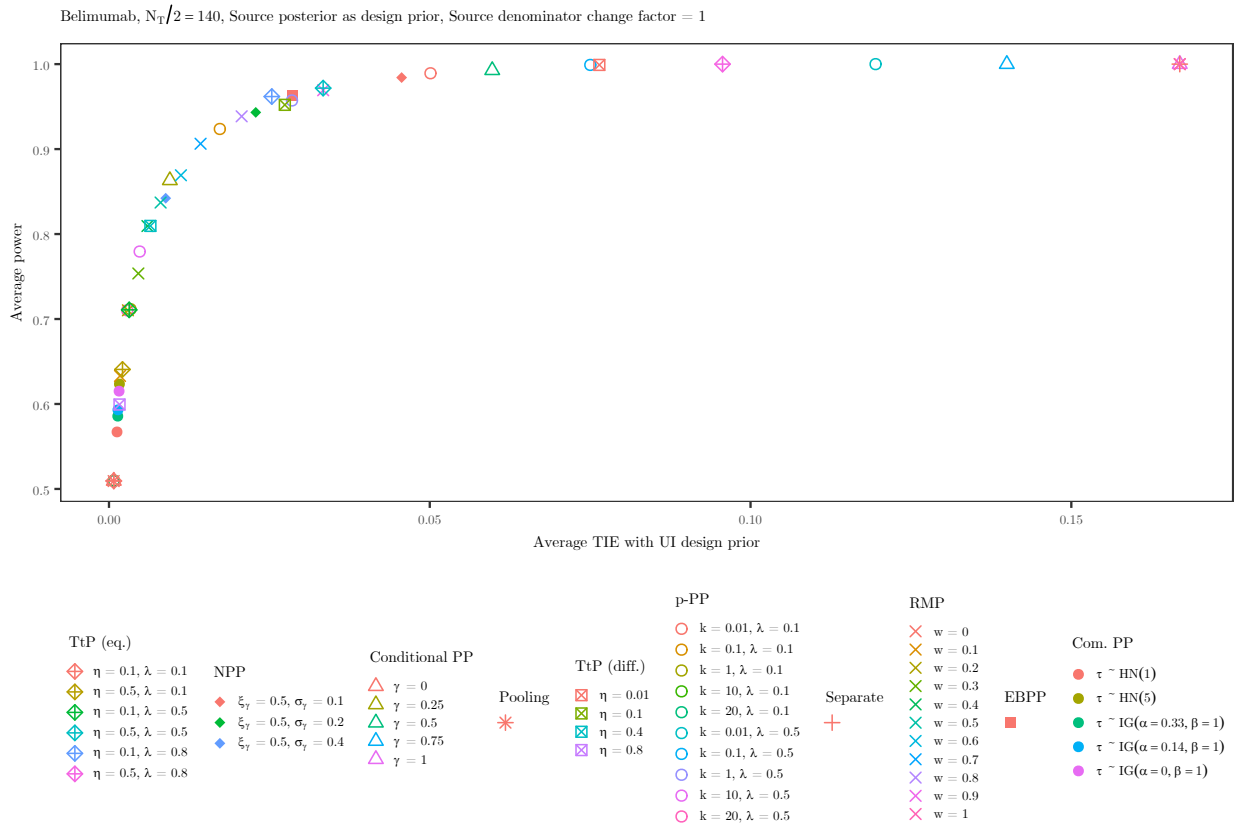

Figure 49: Average power as a function of average type 1 error rate with UI design prior, across all methods, for the Belimumab case study. The sample size per arm is 140, with the source posterior as design prior, without change introduced in the denominator of the source study summary measure.

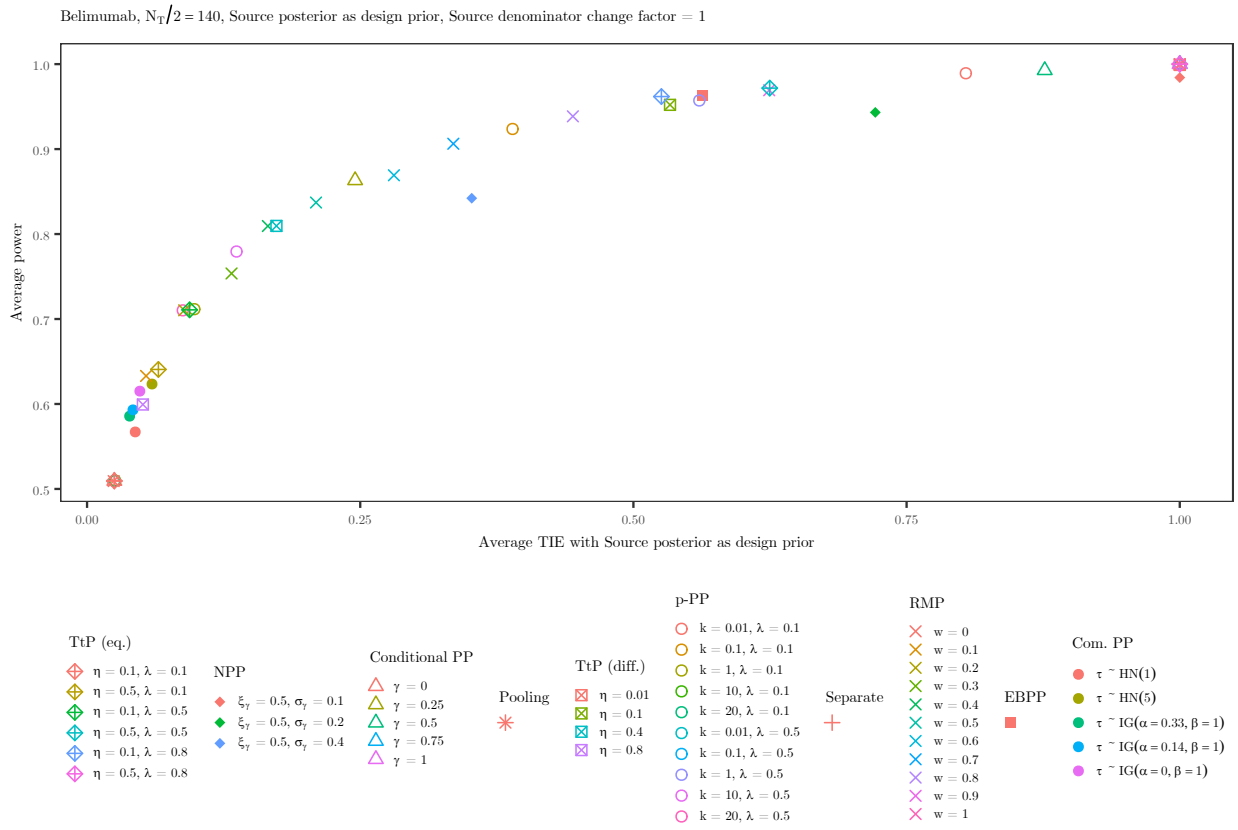

Figure 50: Average power as a function of average type 1 error rate with the source posterior as design prior, across all the methods, for the Belimumab case study. The sample size per arm is 140, without change introduced in the denominator of the source study summary measure.

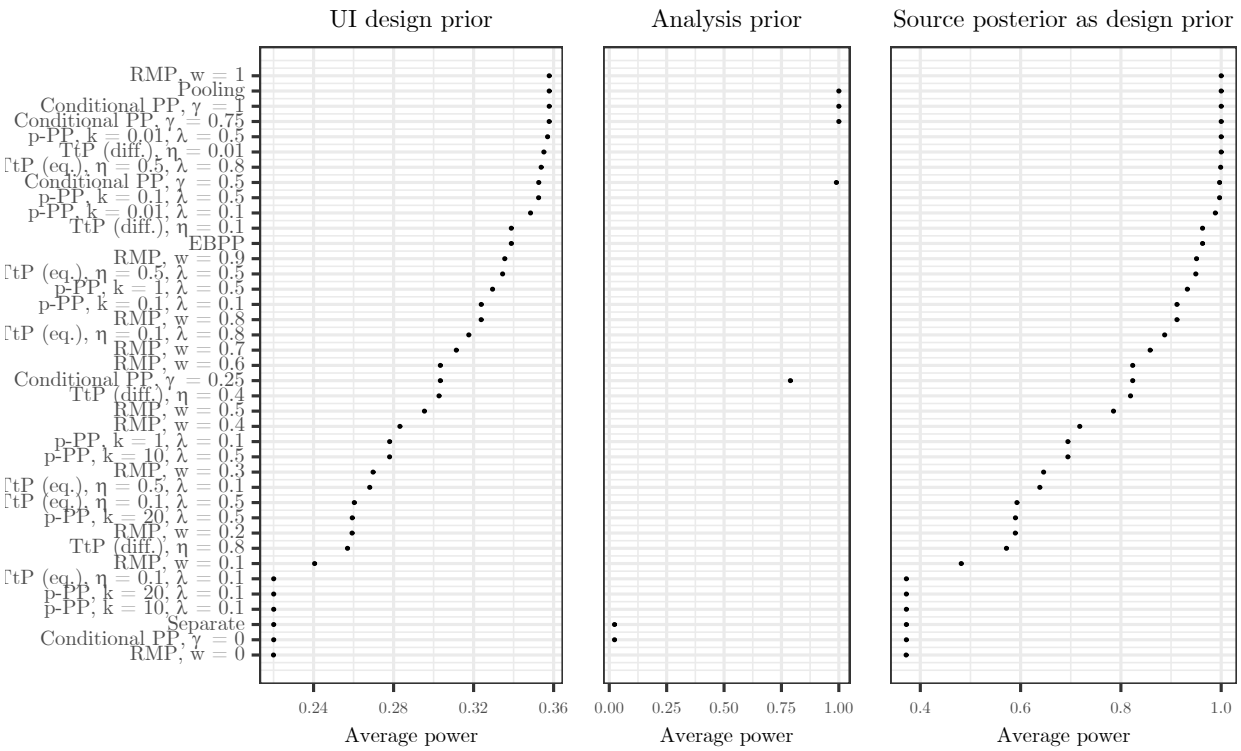

Figure 51: Bayesian power in the Belimumab case study with a target sample size per arm of 93.

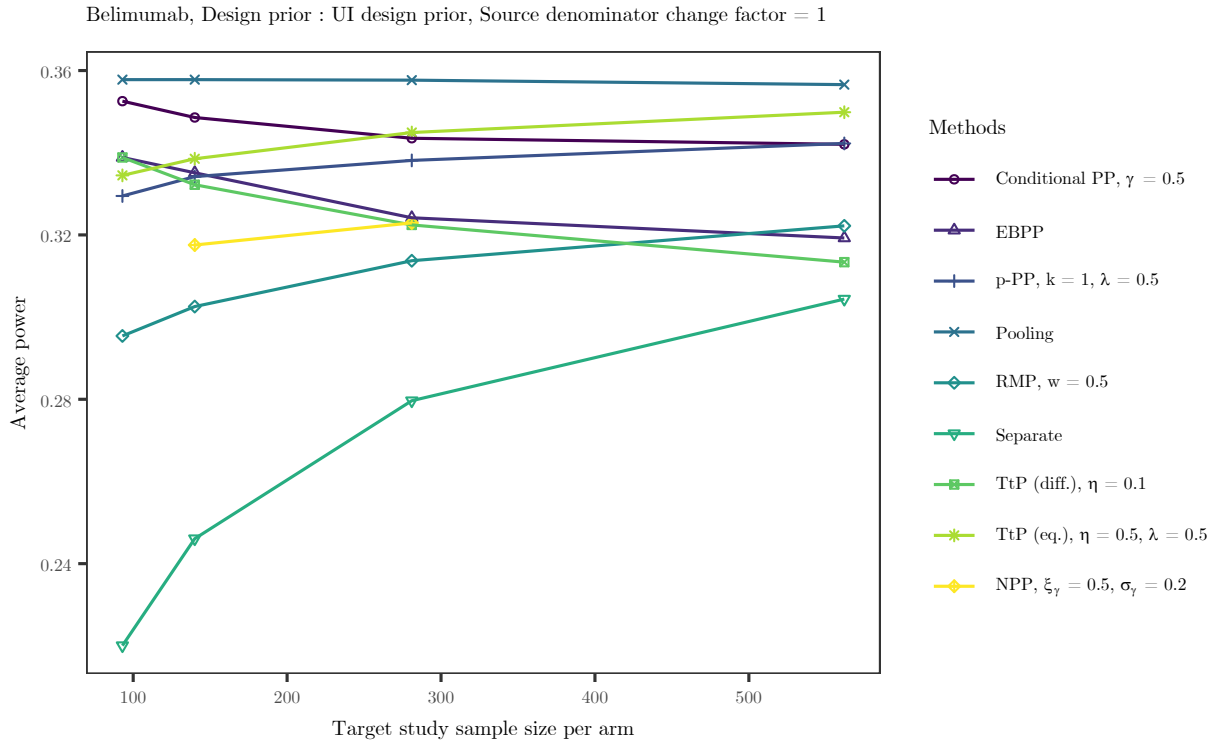

Figure 52: Average power as a function of Study sample size per arm with UI design prior, across all the methods, for the Belimumab case study without change introduced in the denominator of the source study summary measure

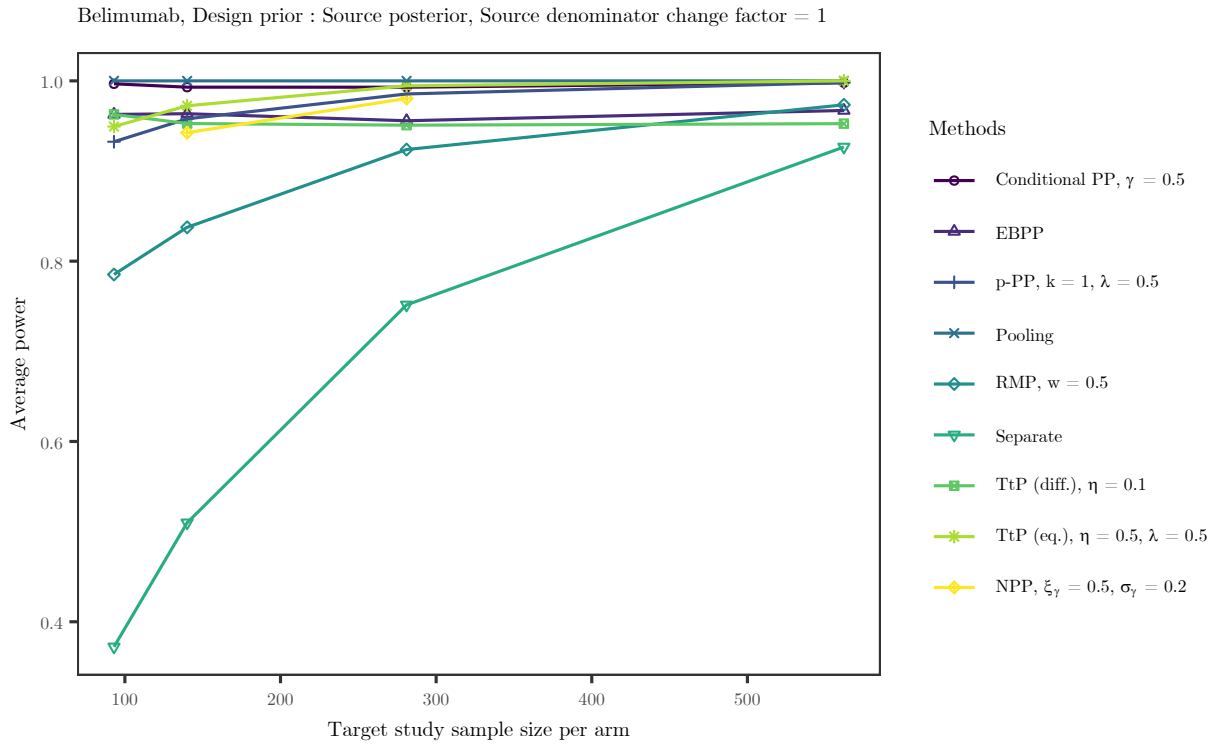

Figure 53: Average power as a function of Study sample size per arm with Source posterior as design prior, across all the methods, for the Belimumab case study without change introduced in the denominator of the source study summary measure

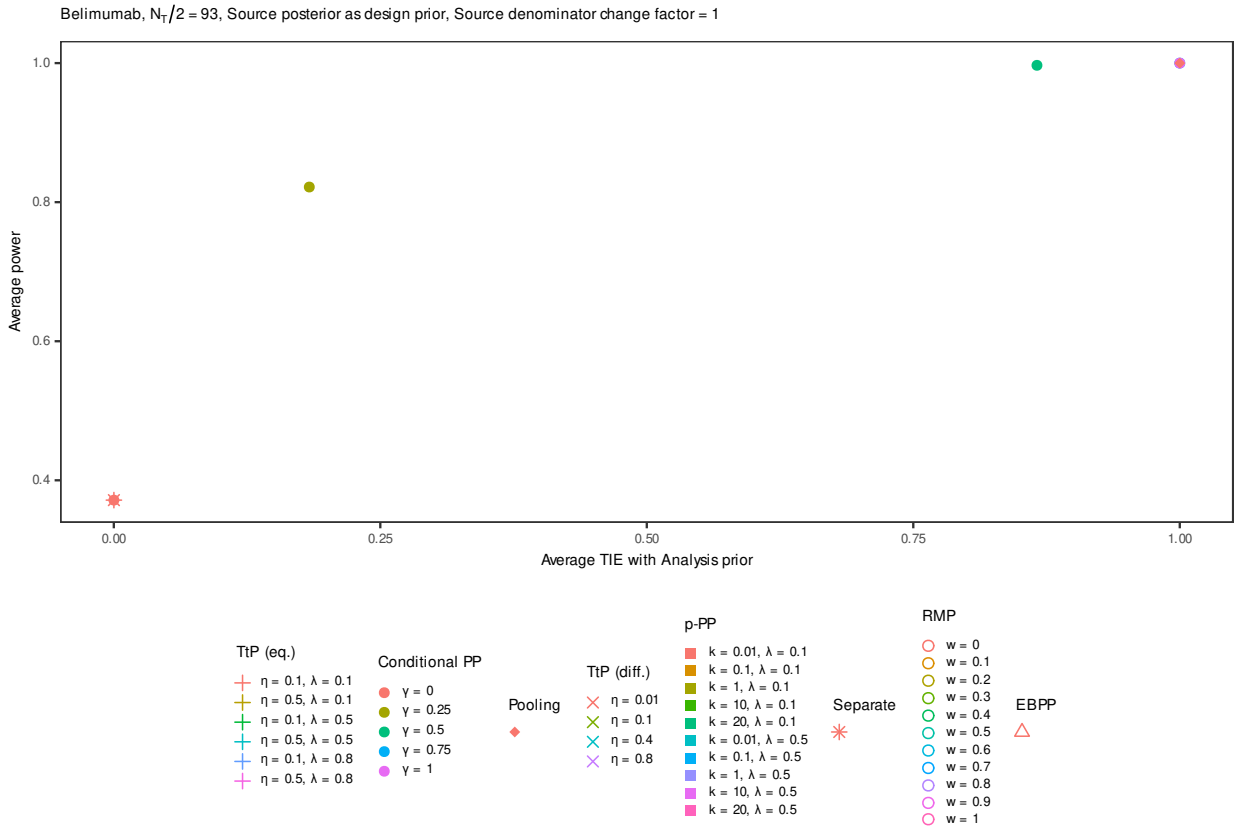

Figure 54: Average power as a function of Analysis prior average type 1 error rate with Source posterior as design prior, across all the methods, for the Belimumab case study. The sample size per arm is 93, without change introduced in the denominator of the source study summary measure

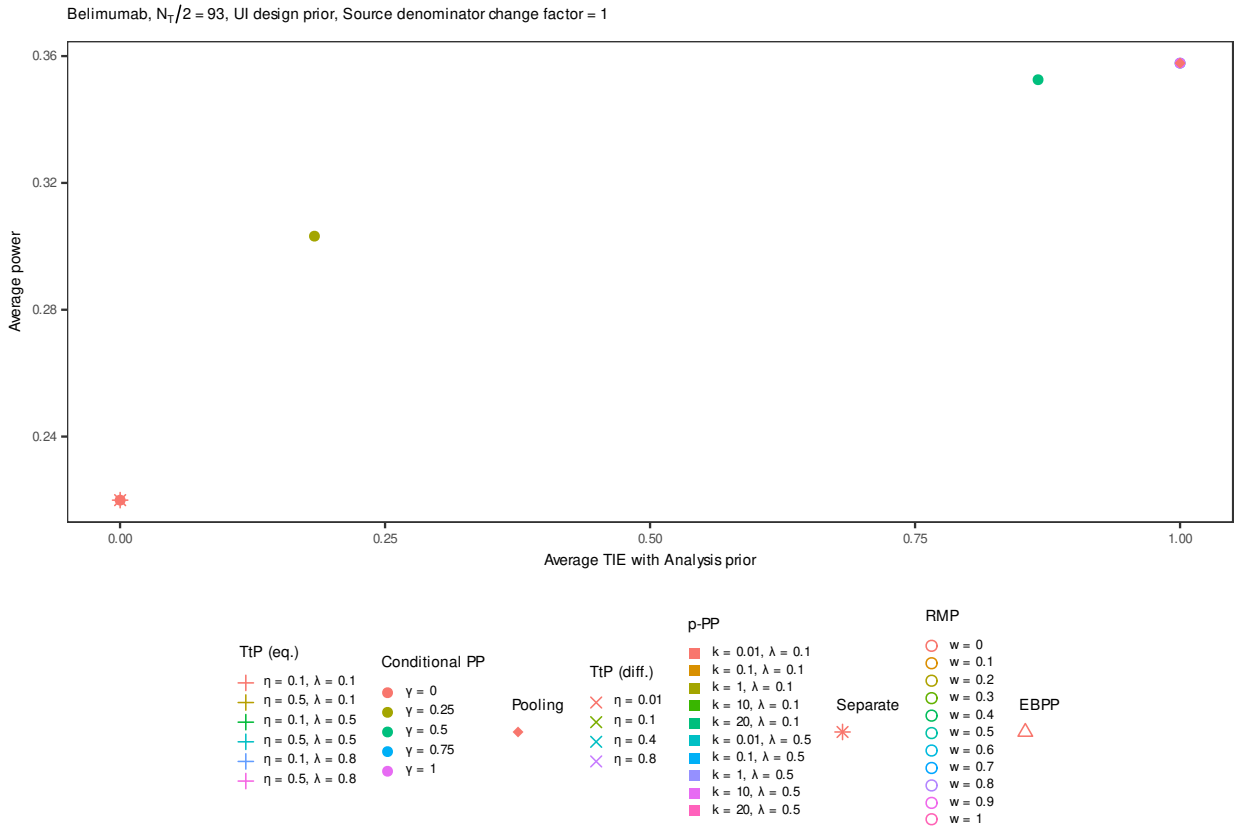

Figure 55: Average power as a function of Analysis prior average type 1 error rate with UI design prior, across all the methods, for the Belimumab case study. The sample size per arm is 93, without change introduced in the denominator of the source study summary measure

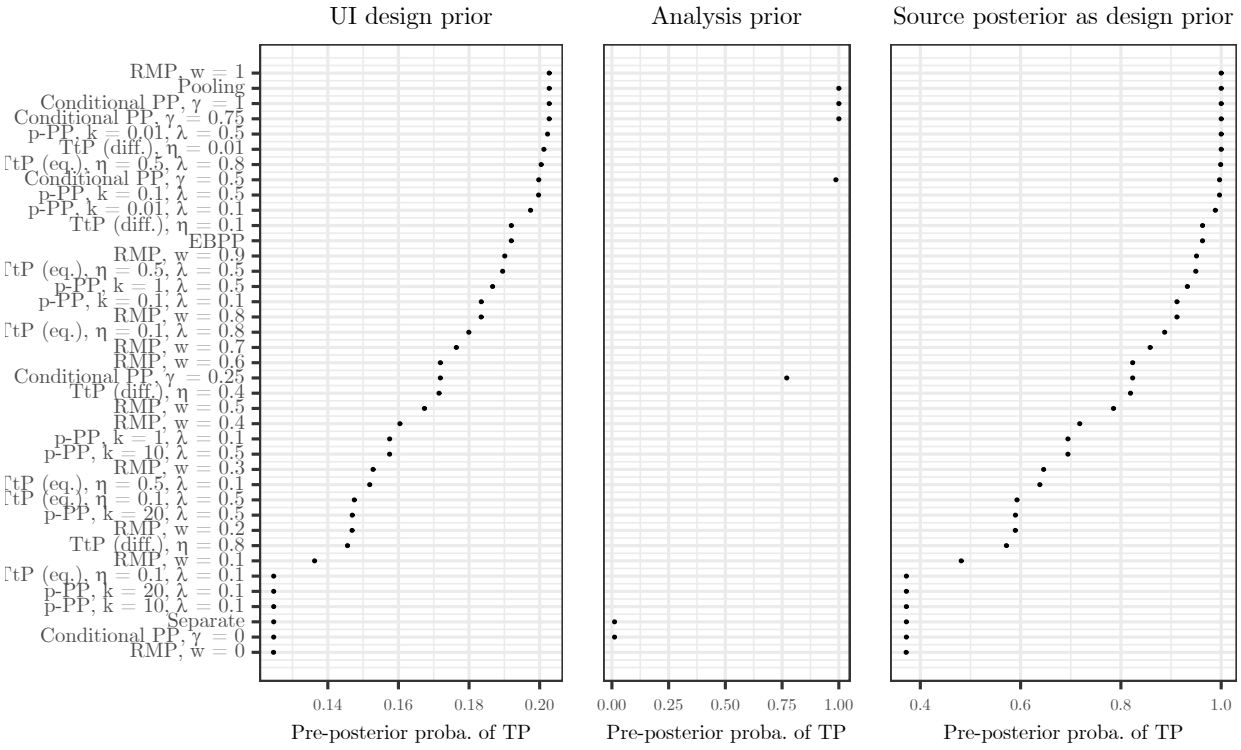

Figure 56: Pre-posterior probability of TP, for the three design prior, in Belimumab case study. The sample size per arm is 93.

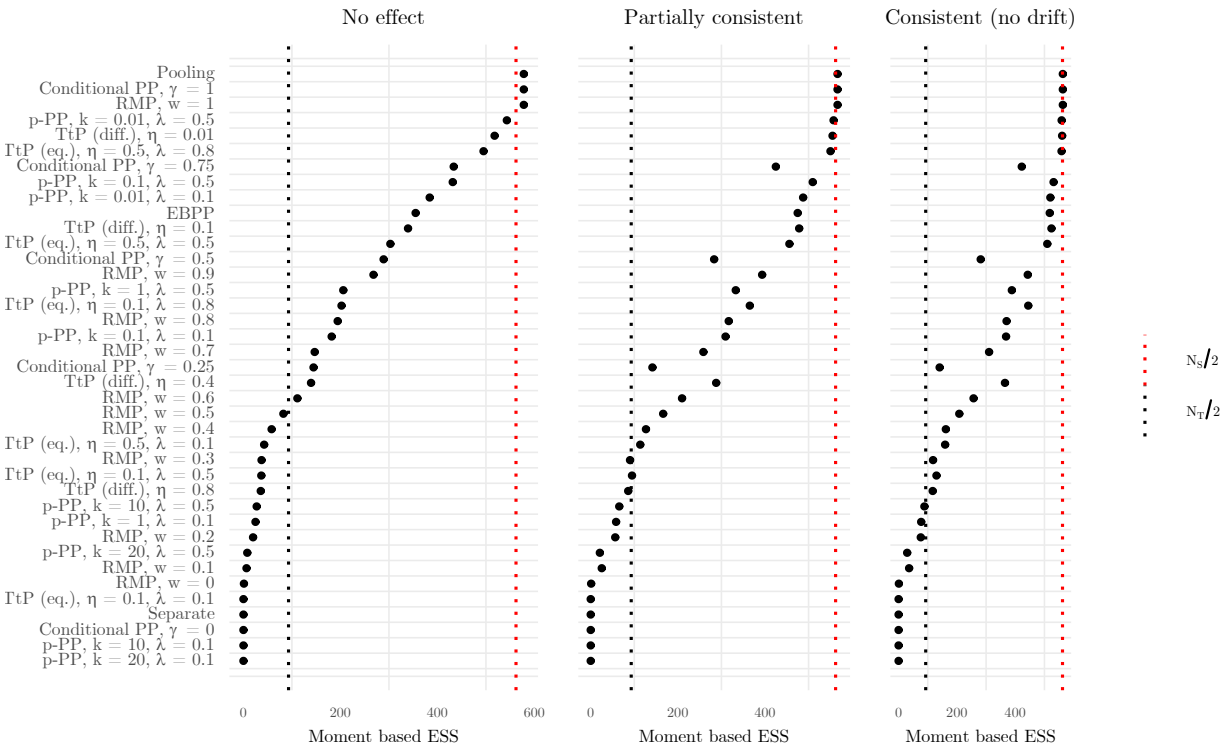

Figure 57: Mean Moment-based ESS across different methods in the Belimumab case study with a sample size per arm in the target study of 93. The black dotted line corresponds to the sample size per arm in the target study. The red dotted line corresponds to the sample size per arm in the source study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

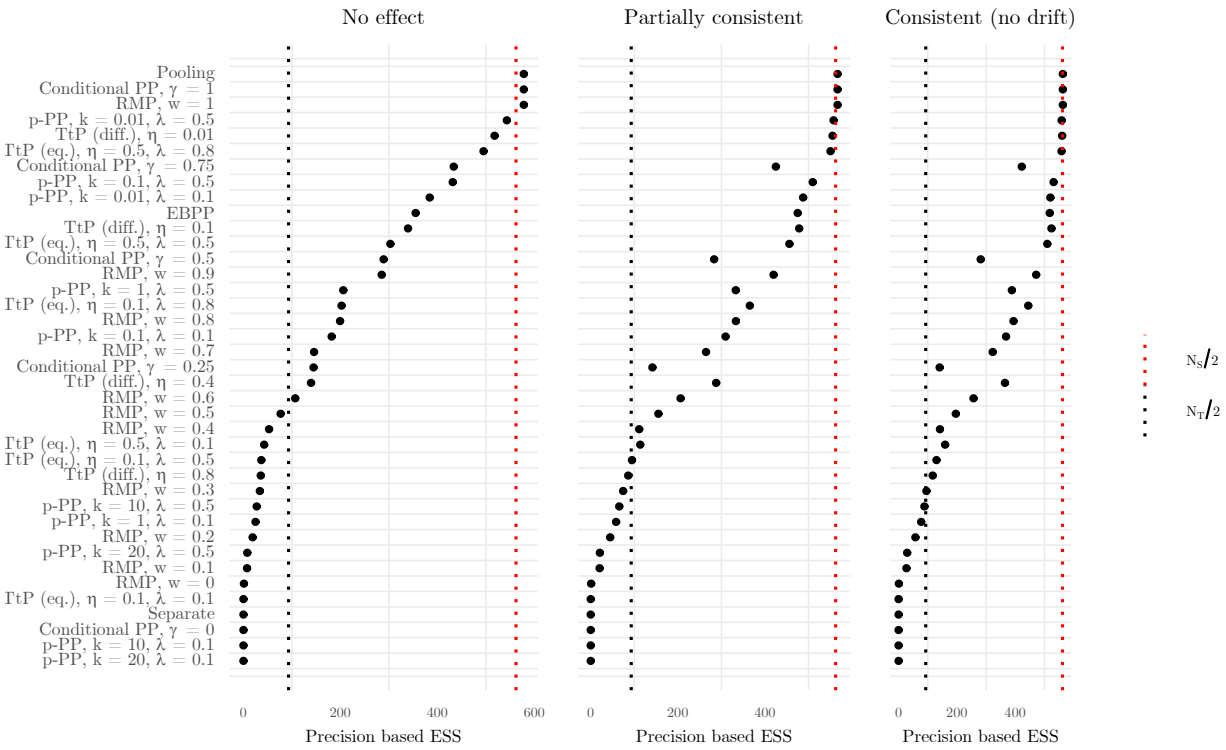

Figure 58: Mean Precision-based ESS across different methods in the Belimumab case study with a sample size per arm in the target study of 93. The black dotted line corresponds to the sample size per arm in the target study. The red dotted line corresponds to the sample size per arm in the source study. Error bars correspond to the 95% Confidence Interval of the Mean Precision-based ESS.

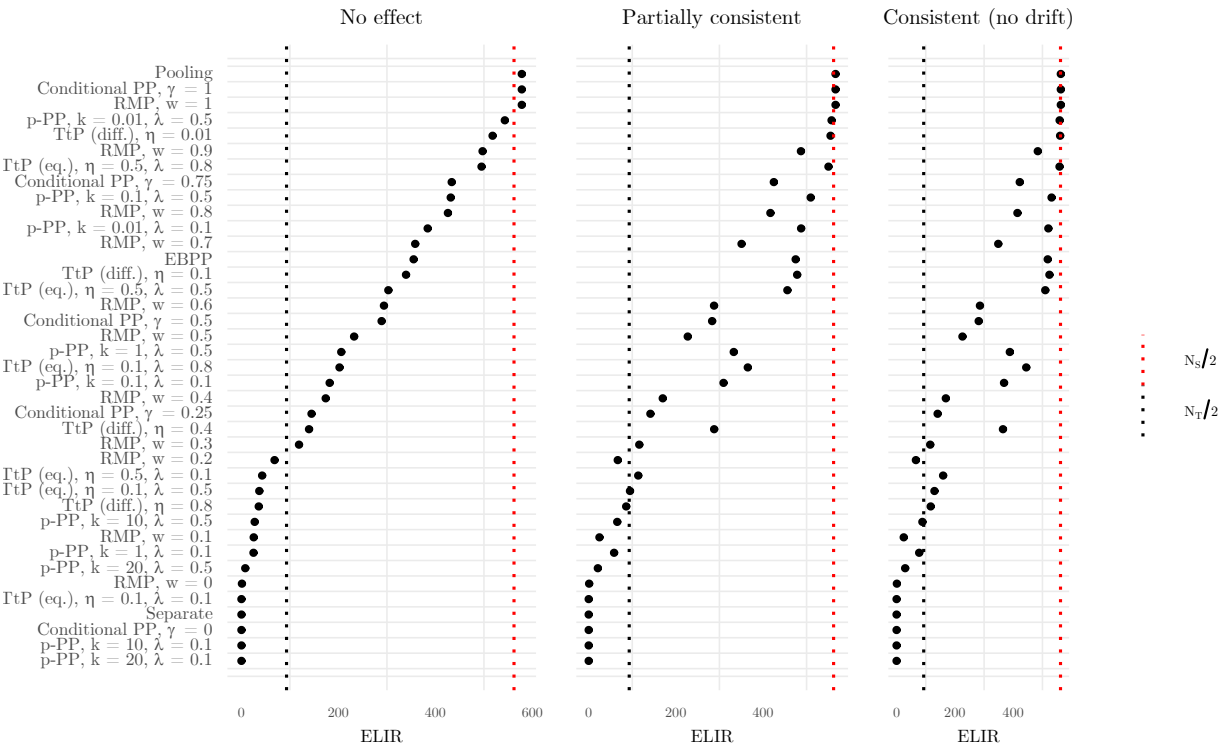

Figure 59: Mean ELIR ESS across different methods in the Belimumab case study with a sample size per arm in the target study of 93. The black dotted line corresponds to the sample size per arm in the target study. The red dotted line corresponds to the sample size per arm in the source study. Error bars correspond to the 95% Confidence Interval of the mean ELIR ESS.

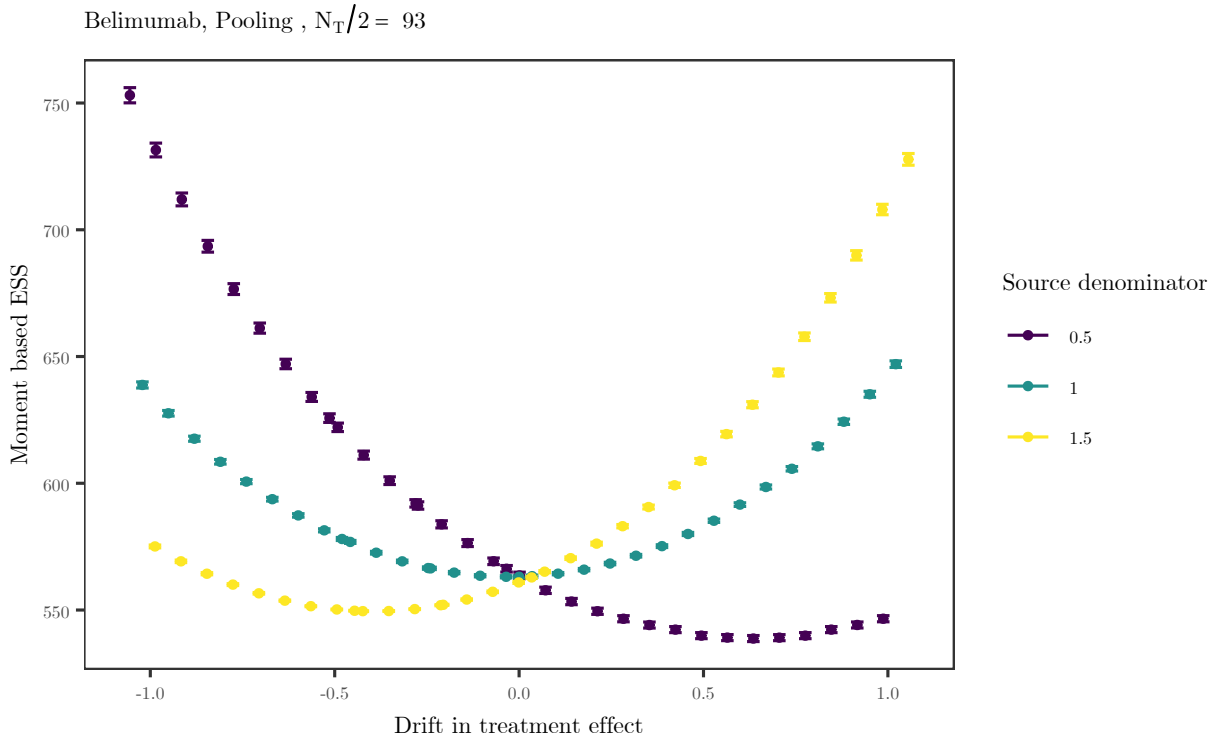

Figure 60: Mean Moment-based ESS as a function of drift in treatment effect in the Belimumab case study with a sample size per arm in the target study of 93 patients, for three change factors in the denominator of the source study summary measure. Error bars correspond to the 95% Confidence Interval of the Mean Moment-based ESS.

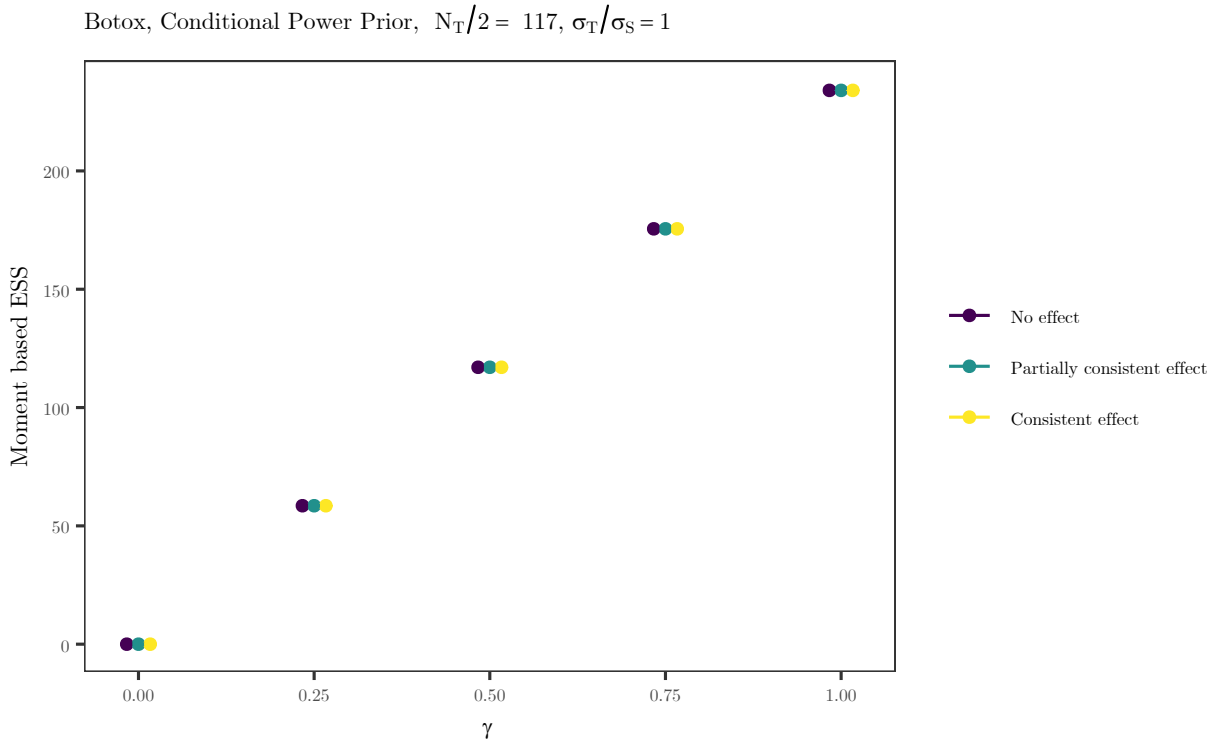

Figure 61: Mean Moment-based ESS as a function of the power parameter for the Conditional Power Prior, for the three main target treatment effect values. Error bars correspond to the 95% Confidence Interval of the Mean Moment-based ESS.

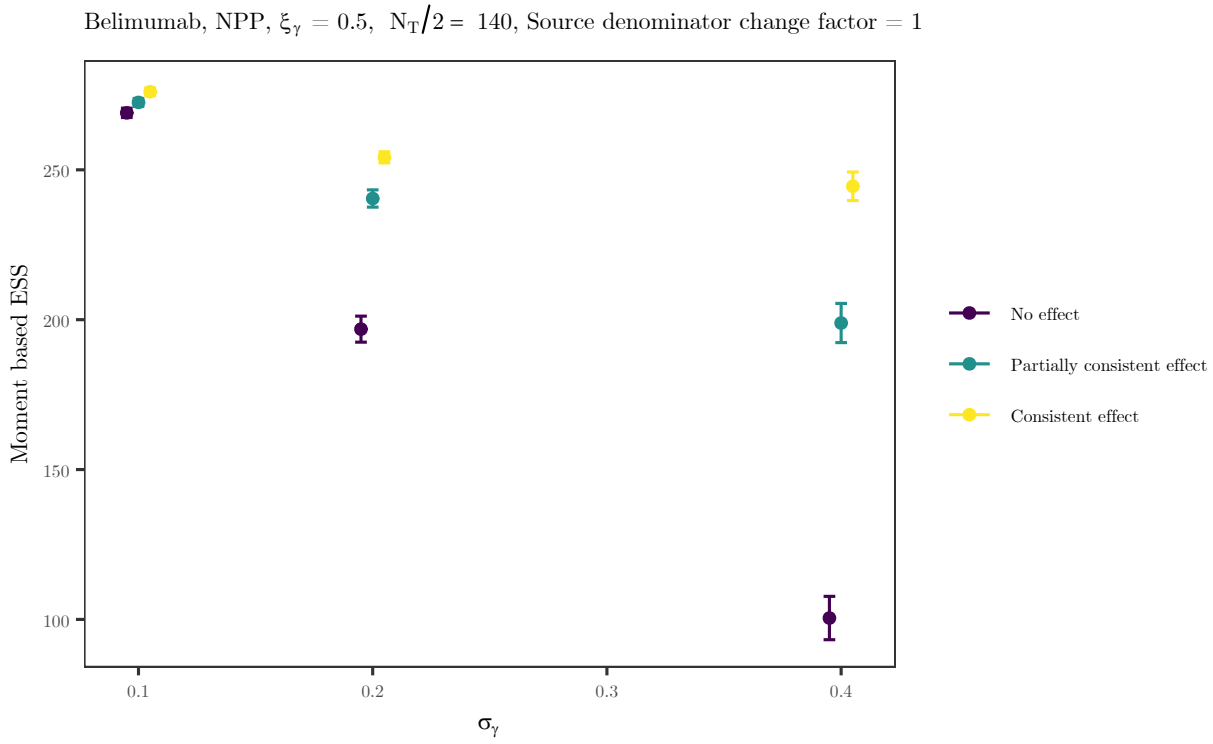

Figure 62: Mean Moment-based ESS as a function of the standard deviation of the hyperprior on the power parameter for the Normalized Power Prior, for the three main target treatment effect values, in the Belimumab case study, for a target sample size per arm of 281, without change in the source study summary measure denominator. Error bars correspond to the 95% Confidence Interval of the Mean Moment-based ESS.

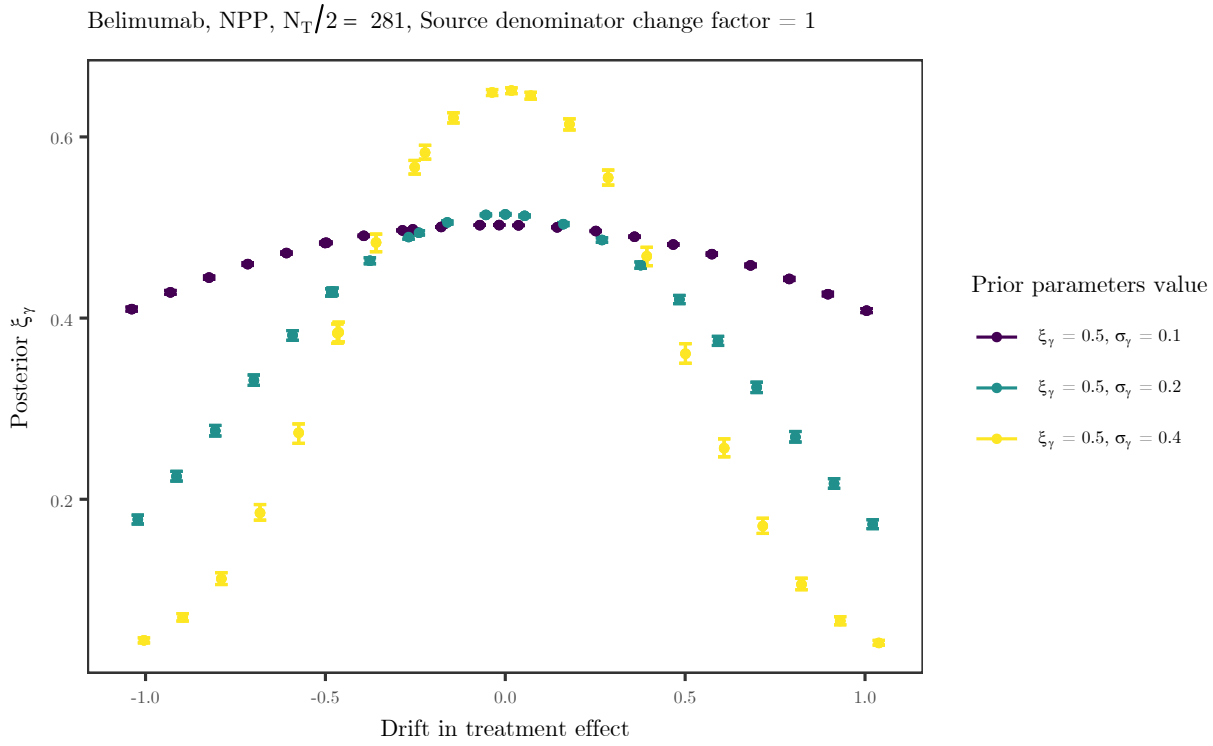

Figure 63: Posterior mean of the power parameter distribution in the Normalized Power Prior as a function of the drift, for different parameters of the Beta hyperprior on the power parameter, in the Belimumab case study, for a target sample size per arm of 281, without change in the source study summary measure denominator. Error bars correspond to the 95% Confidence Interval of the posterior mean.

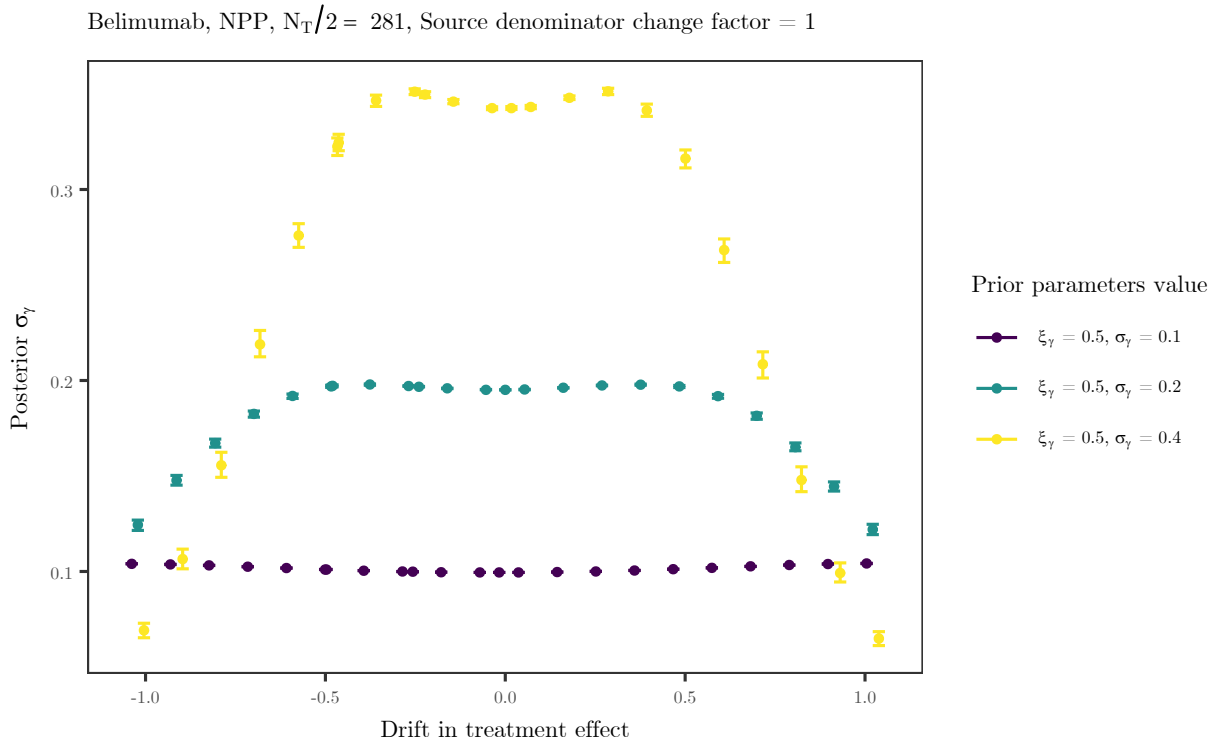

Figure 64: Posterior standard deviation of the power parameter distribution in the Normalized Power Prior as a function of the drift, for different parameters of the Beta hyperprior on the power parameter, in the Belimumab case study, for a target sample size per arm of 281, without change in the source study summary measure denominator. Error bars correspond to the 95% Confidence Interval of the posterior standard deviation.

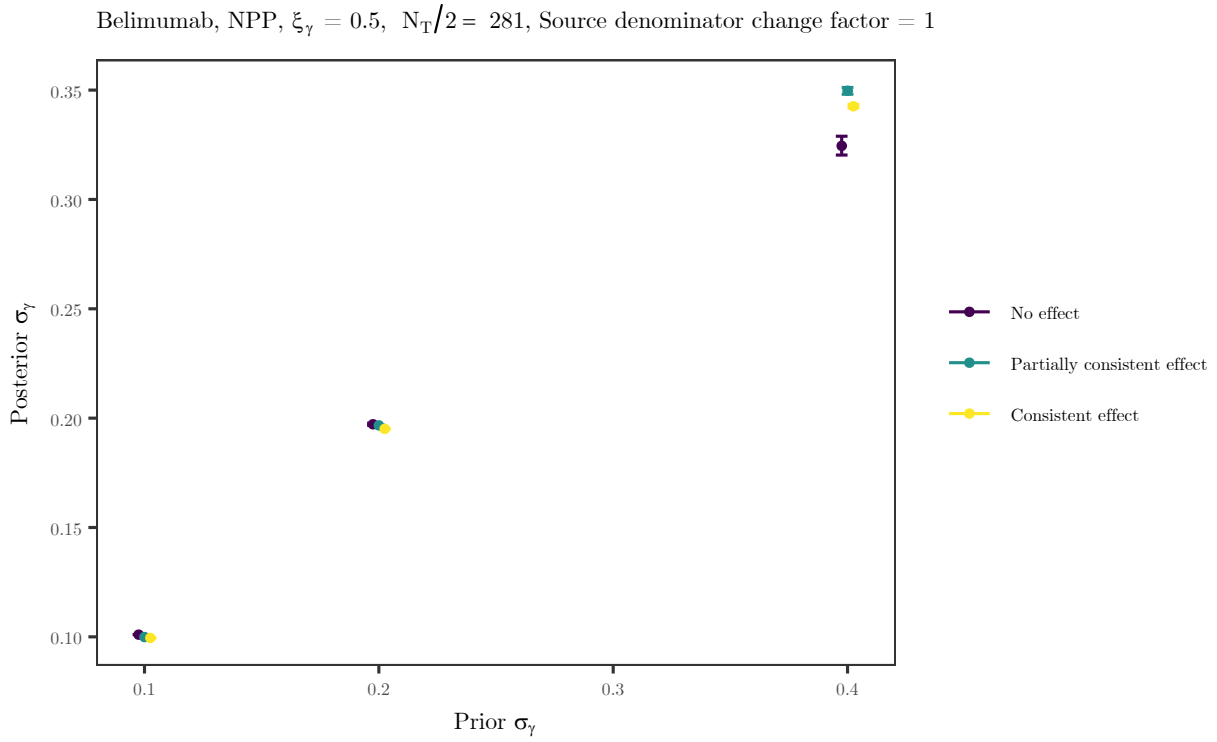

Figure 65: Standard deviation of the posterior distribution of the power parameter in the Normalized Power Prior, as a function of the standard deviation of the prior distribution of the power parameter in the Belimumab case study, for a target sample size per arm of 281, without change in the source study summary measure denominator. Error bars correspond to the 95% Confidence Interval of the standard deviation.

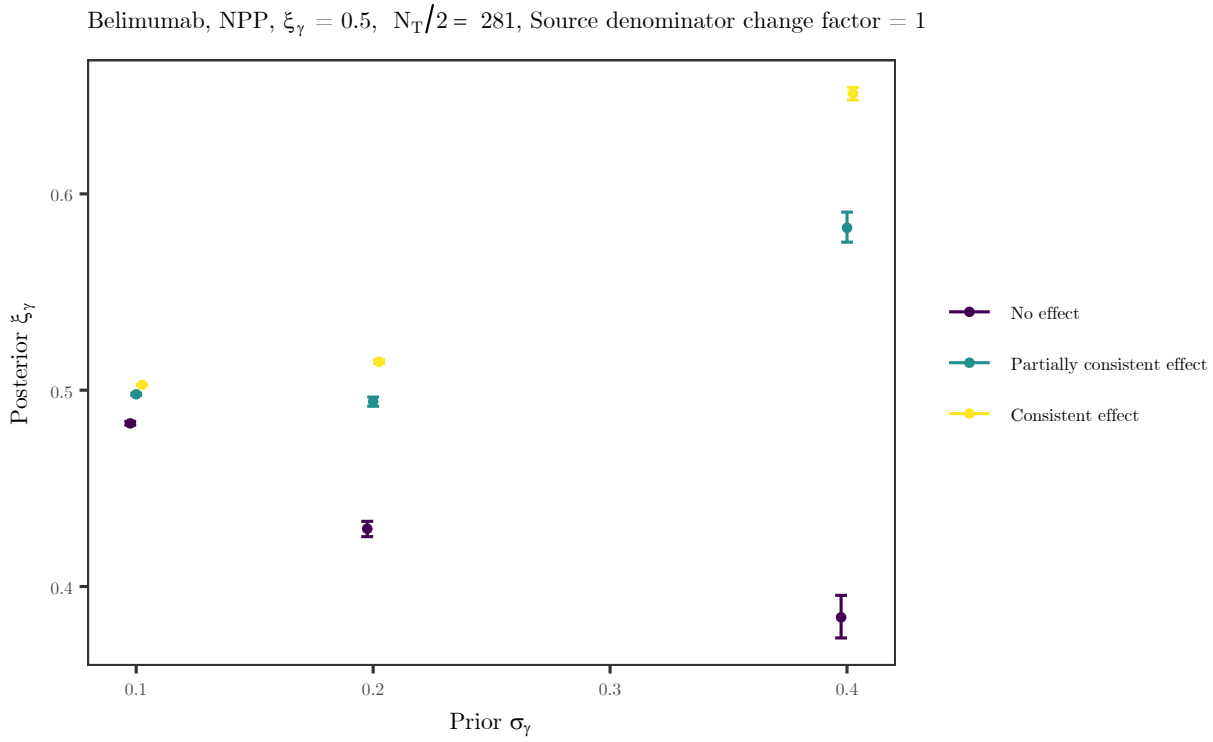

Figure 66: Mean of the posterior distribution of the power parameter with the Normalized Power Prior, as a function of the standard deviation of the prior distribution of the power parameter in the Belimumab case study, for a target sample size per arm of 281, without change in the source study summary measure denominator. Error bars correspond to the 95% Confidence Interval of the mean of the posterior distribution.

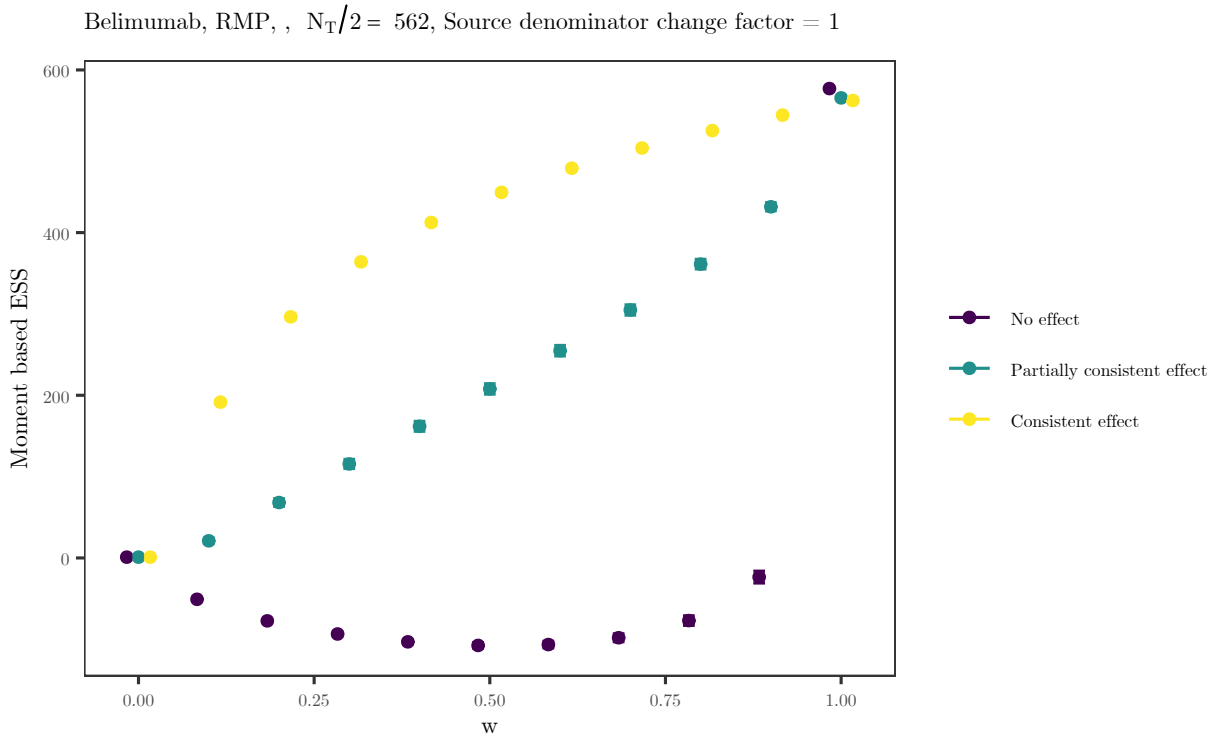

Figure 67: Mean Moment-based ESS as a function of the standard deviation of the prior weight w of the Normalized Power Prior, for the three main target treatment effect values, in the Belimumab case study, for a target sample size per arm of 281, without change in the source study summary measure denominator. In the absence of effect, the moment-based ESS can be negative. Error bars correspond to the 95% Confidence Interval of the Mean Moment-based ESS.

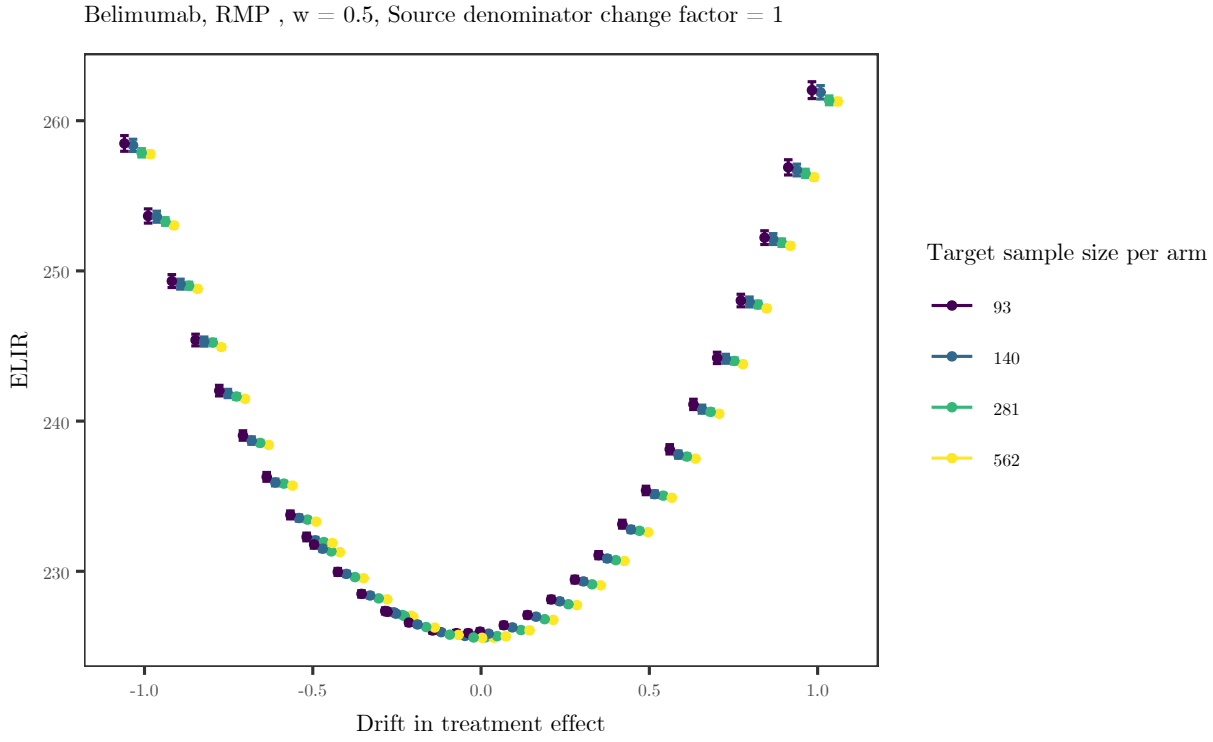

Figure 68: Mean ELIR ESS as a function of the drift in treatment effect for the RMP with prior weight $w = 0.5$, for different sample size per arm in the target study. Error bars correspond to the 95% Confidence Interval of the mean ELIR ESS.

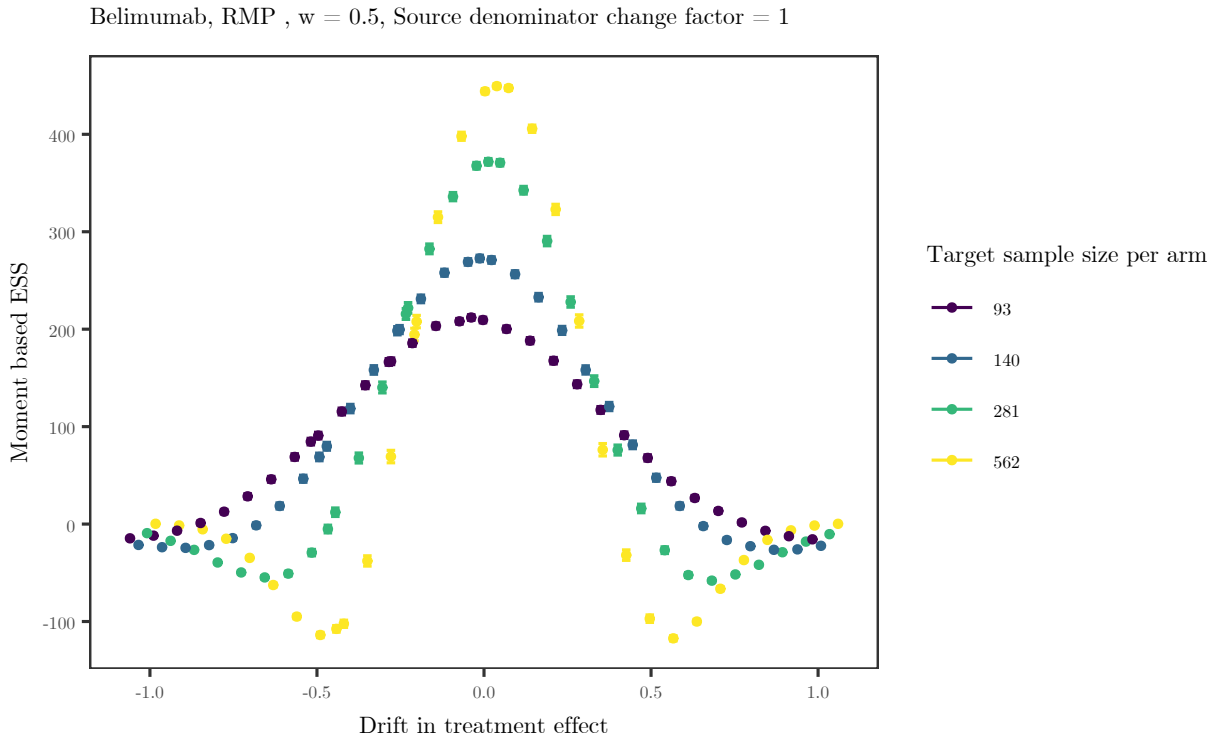

Figure 69: Mean Moment-based ESS as a function of the drift in treatment effect for the RMP, in the Belimumab case study, with a prior weight of 0.5 for various sample size per arm in the target study. Error bars correspond to the 95% Confidence Interval of the Mean Moment-based ESS.

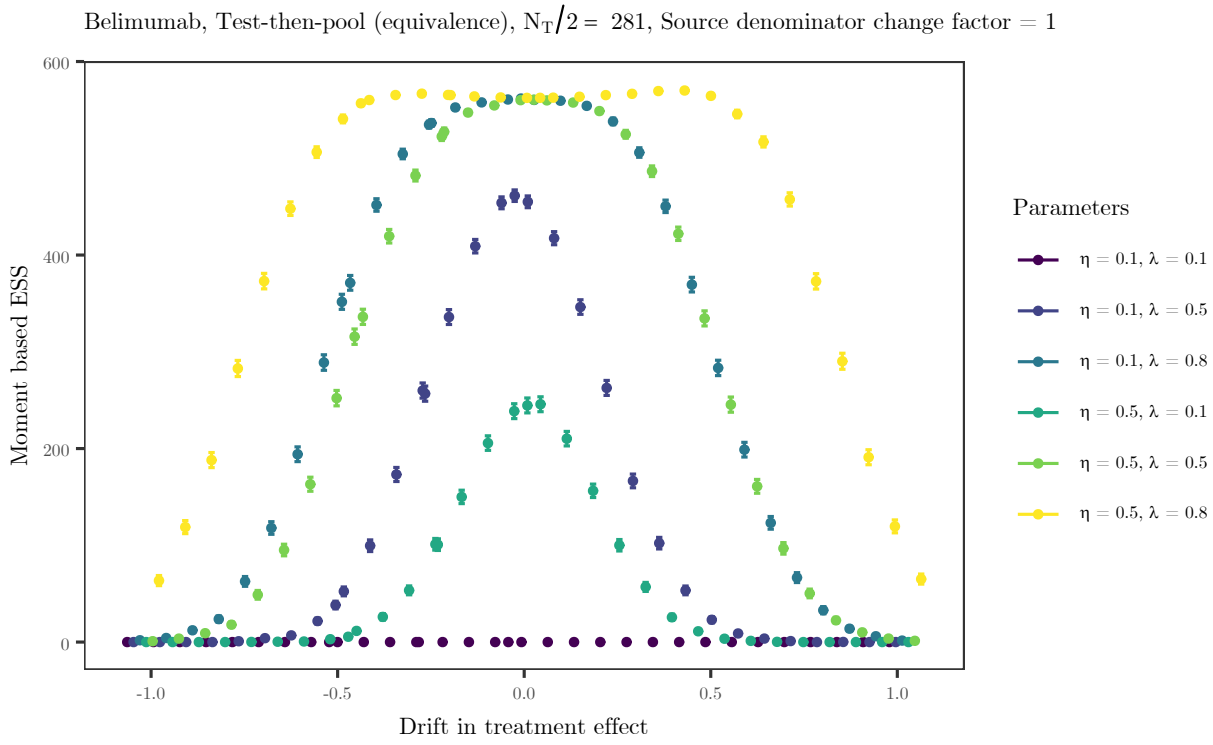

Figure 70: Mean Moment-based ESS of the test-then-pool method with an equivalence test as a function of the drift in treatment effect, for various parameters, in the Belimumab case study with 281 subjects per arm in the target study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

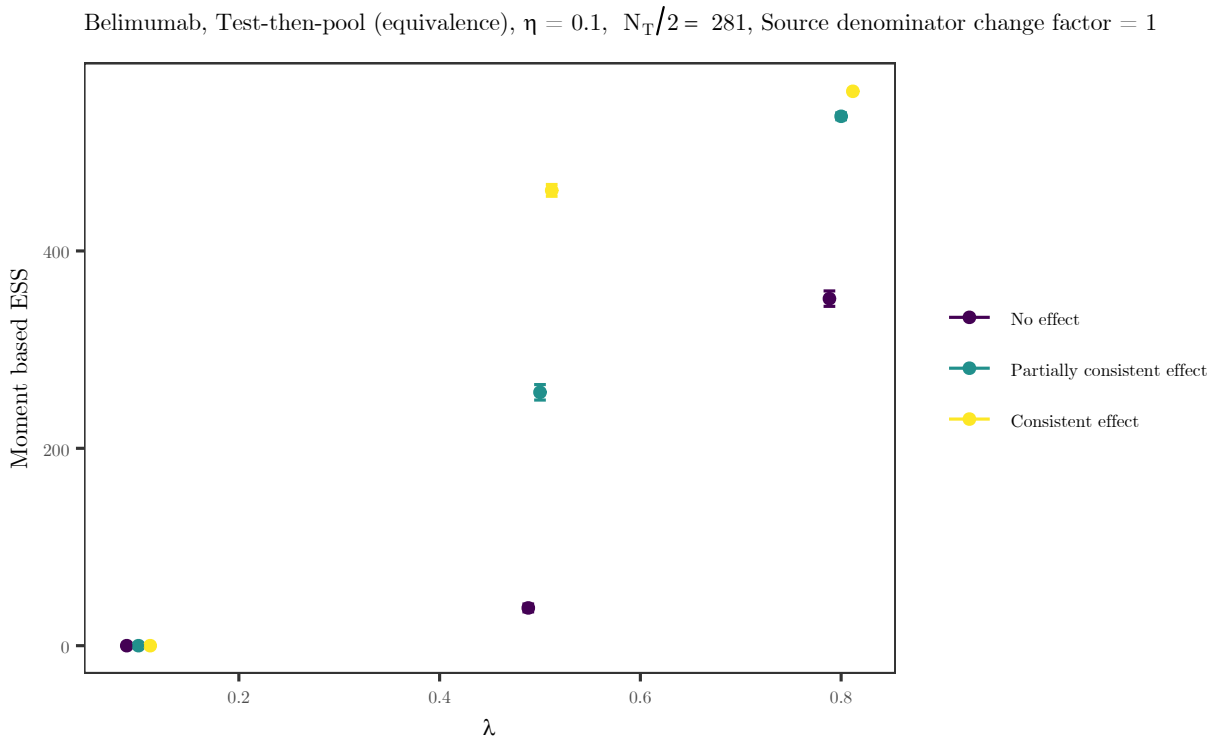

Figure 71: Mean Moment-based ESS of the test-then-pool method with an equivalence test as a function of the equivalence margin parameter λ, for the three main drift in treatment effect and significance threshold $\eta = 0.1$, in the Belimumab case study with 281 subjects per arm in the target study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

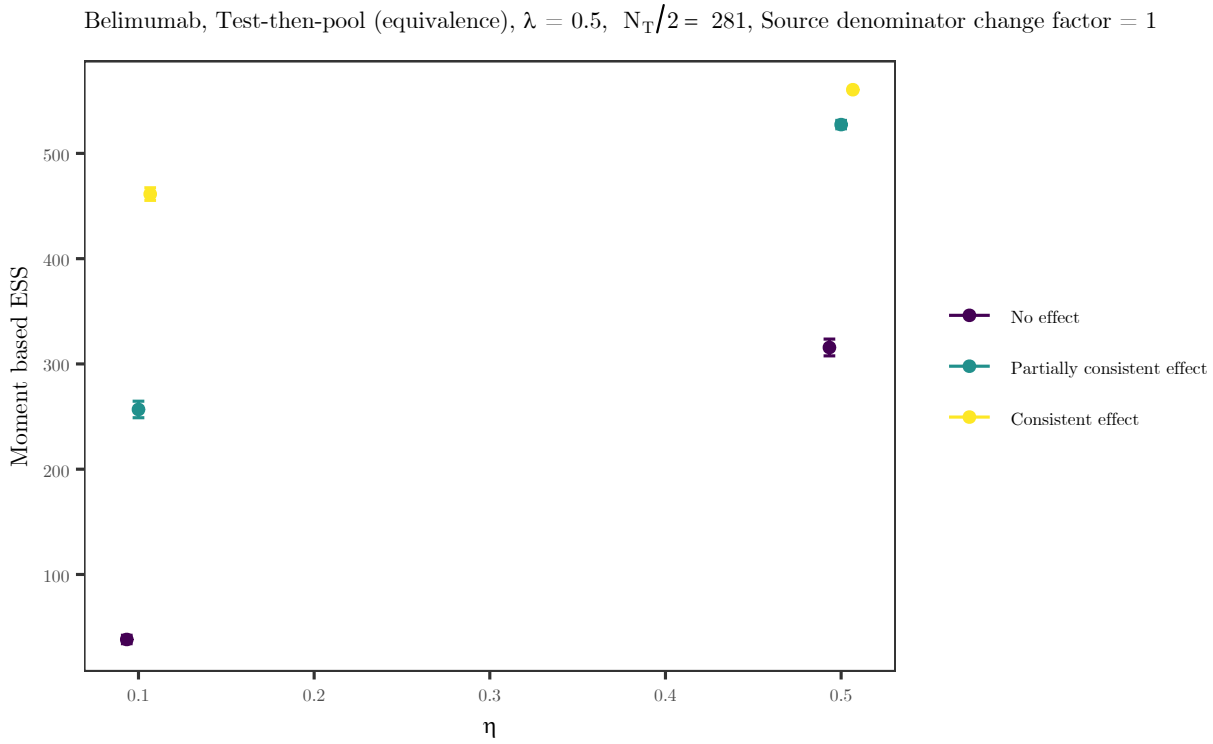

Figure 72: Mean Moment-based ESS of the test-then-pool method with an equivalence test as a function of the significance threshold parameter η, for the three main drift in treatment effect and equivalence margin $\lambda = 0.5$, in the Belimumab case study with 281 subjects per arm in the target study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

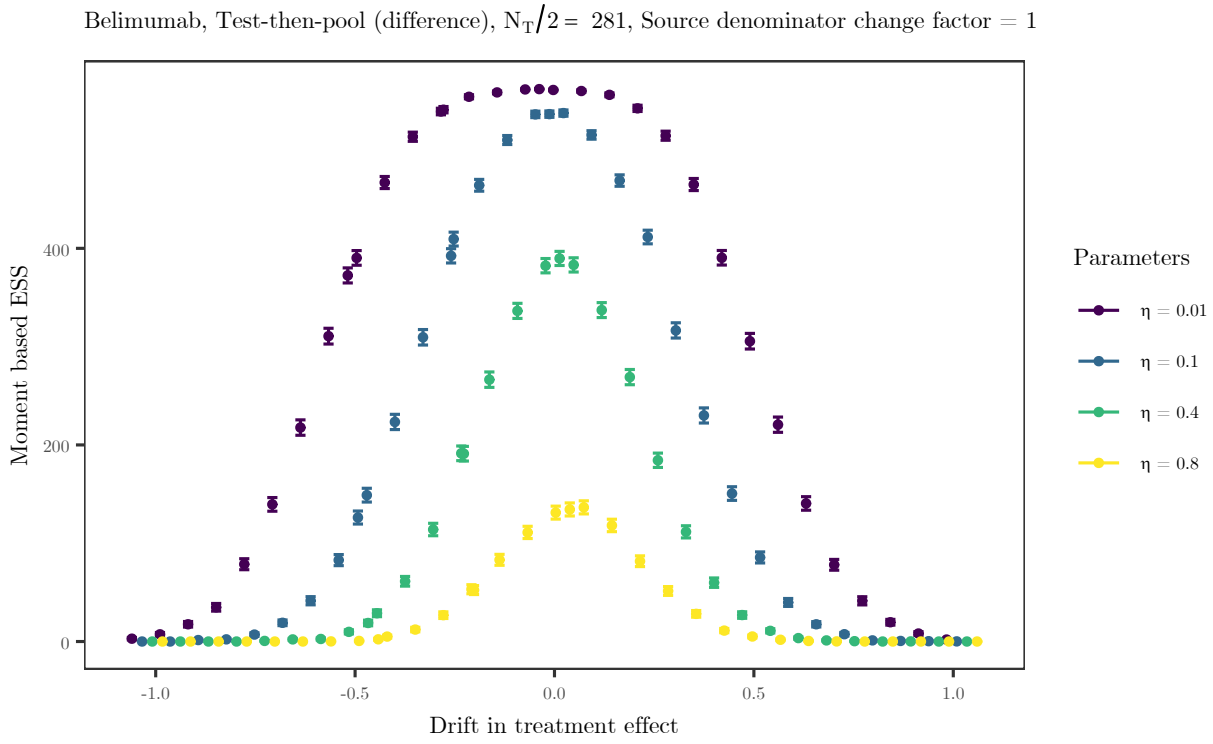

Figure 73: Mean Moment-based ESS of the test-then-pool method with a test for difference, as a function of the drift in treatment effect, for various values of the equivalence margin η, in the Belimumab case study with 281 subjects per arm in the target study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

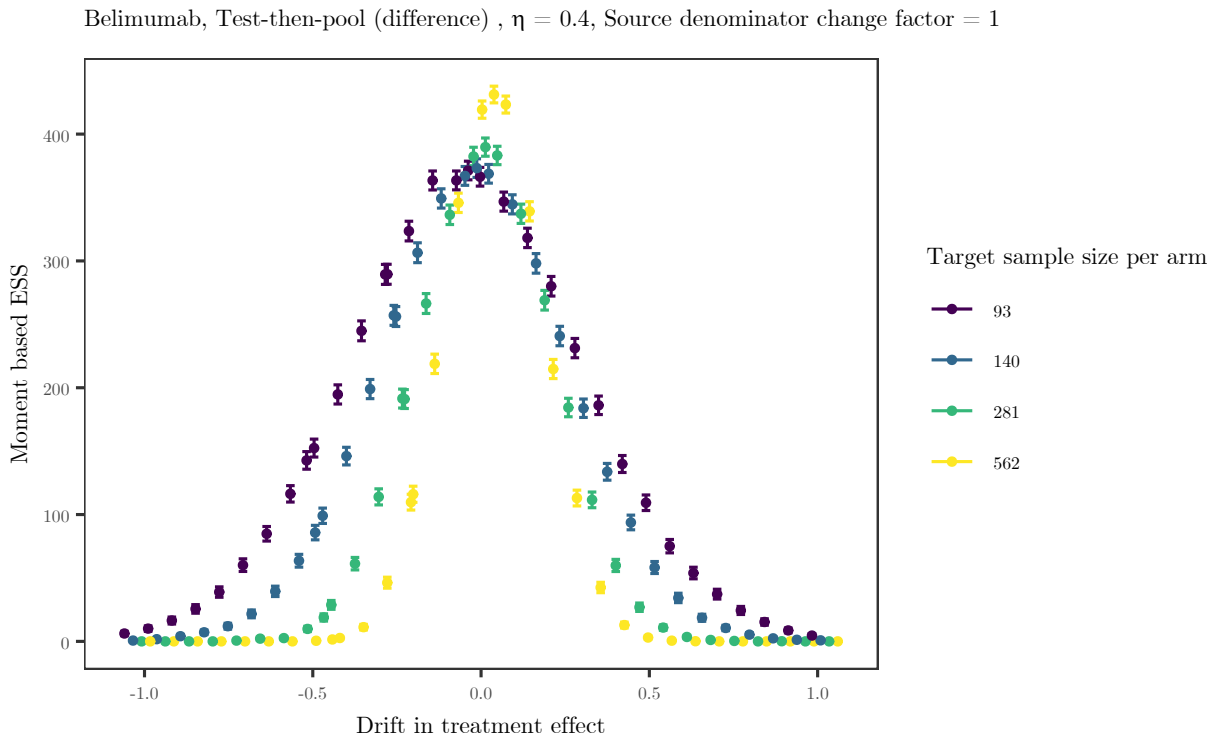

Figure 74: Mean Moment-based ESS of the test-then-pool method with test for difference, as a function of the drift in treatment effect, for various target study sample size per arm and equivalence margin $\eta = 0.4$, in the Belimumab case study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

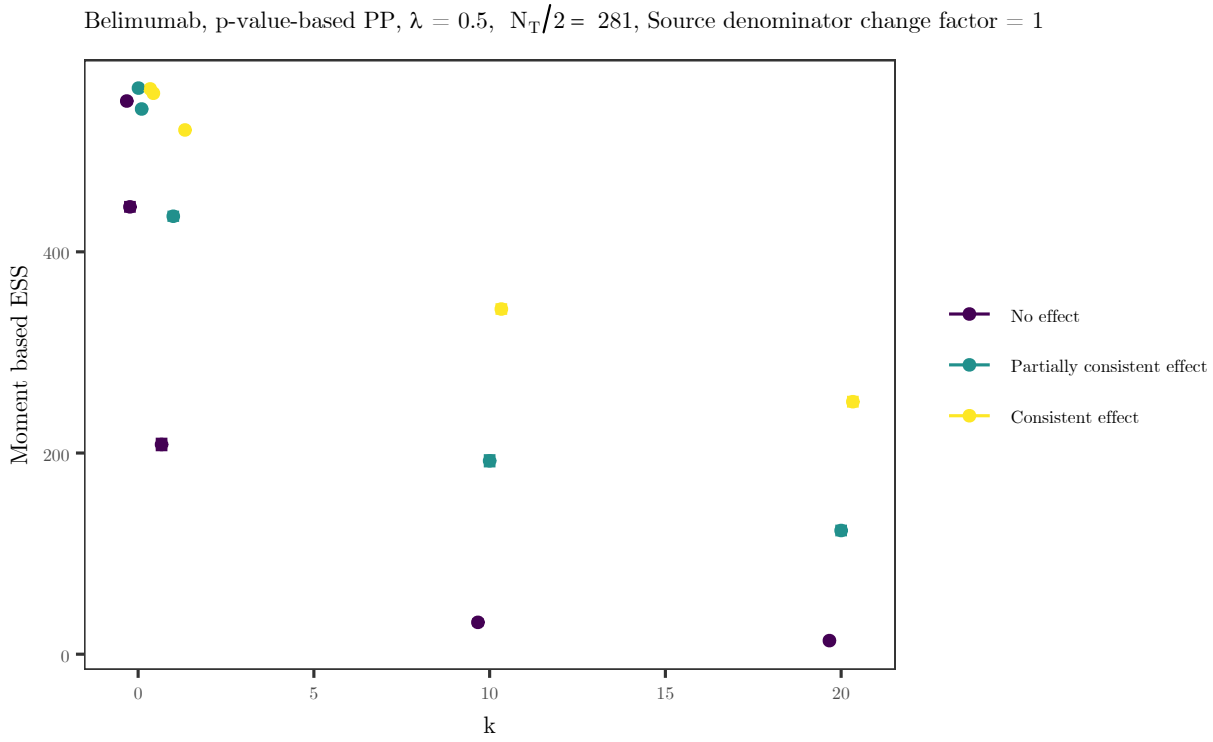

Figure 75: Mean moment-based ESS of the p-value based PP as a function of the shape parametr k, in the Belimumab case study with a sample size per arm in the target study of 281 patients, with $\lambda = 0.5$, for the three main treatment effect values in the target study. Error bars correspond to the 95% Confidence Interval of the Mean moment-based ESS.

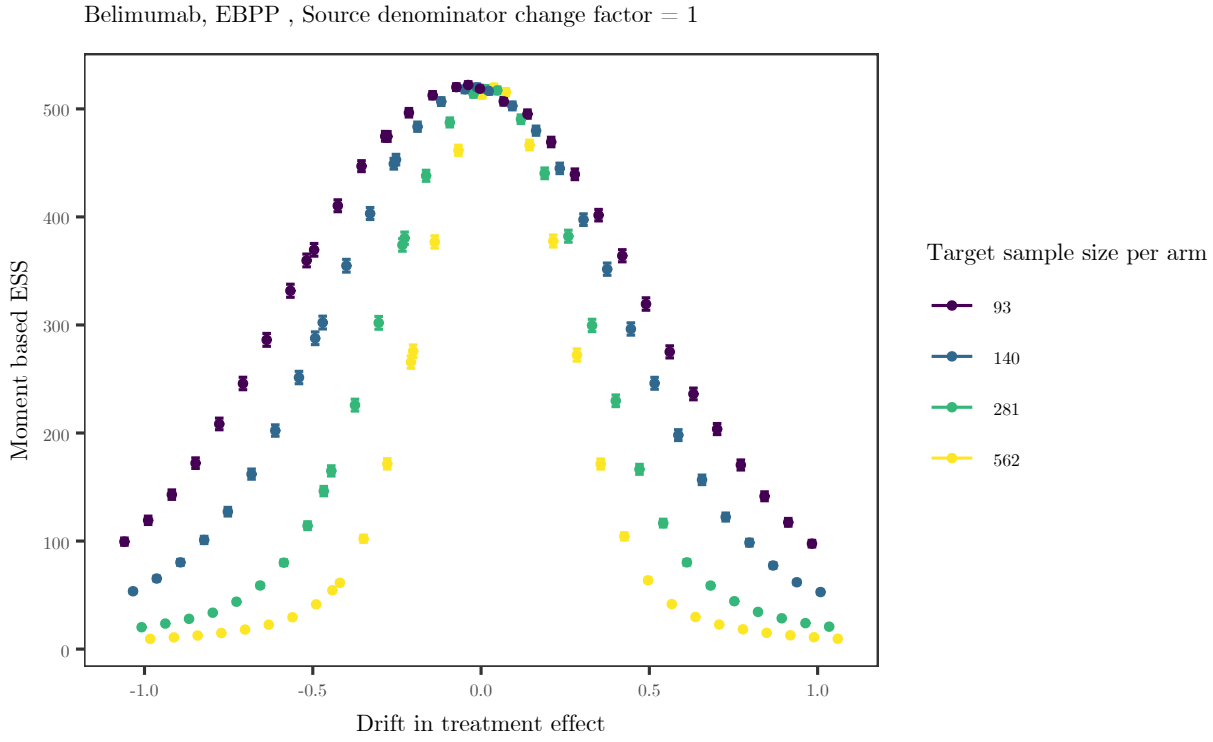

Figure 76: Mean Moment-based ESS of the Empirical Bayes Power Prior method as a function of the drift in treatment effect, for various target study sample size per arm, in the Belimumab case study. Error bars correspond to the 95% Confidence Interval of the mean Moment-based ESS.

12 Discussion

12.1 Type I error rate of borrowing methods

Our results show that borrowing treatment effects systematically leads to an increased type 1 error rate, to an extent that strongly depends on methods and method parameters. This is in agreement with previous literature, with e.g. Campbell (2017) noticing that "if the prior data makes the null hypothesis more unlikely, it may be no surprise that the type I error probability calculated under the unlikely null hypothesis is inflated". It is sometimes considered that Bayesian borrowing methods could be used as a principled way of incorporating prior information on the treatment effect from the source population, at the expense of increasing the acceptable type 1 error rate, but also with the benefit of increased power. The idea is that one may agree on a given analysis prior based on assumptions regarding the similarity of the source and target study, and this assumption would translate into an increased, yet potentially acceptable, type 1 error rate. However, our results show that methods do not behave equally for a given increase in type 1 error rate, with some methods providing less accurate estimates at equivalent type 1 error rate compared to other methods. This is the case, in particular, of the test-then-pool variants, the EBPP, and the p-value-based power prior, which do not perform as well as fixed borrowing methods when considering this criterion.

12.2 Power losses at equivalent type 1 error rate

Bayesian borrowing methods are sometimes motivated by potential power gains compared to frequentist methods, with some authors suggesting, in the case of historical control borrowing, that this can be achieved at equivalent or lower type 1 error rate Viele et al. (2014) and Yang et al. (2023). Indeed, intuitively, an informative prior containing information from the source population should improve the chance of meeting the decision criterion in the study in the target population. However, Kopp-Schneider et al. (2020), (preceded by Psioda and Ibrahim (2019) in the Gaussian case) showed that in terms of power gain, “approaches adaptively discounting prior information do not offer any advantage over a fixed amount of borrowing, or no borrowing at all", when a Uniformly Most Powerful (UMP) test exists, which is the case in most settings encountered in confirmatory trials. The argument can be summarized as follows: let us denote α_B the type 1 error rate obtained with borrowing, a borrowing method increases the type 1 error rate to α_B. Let us denote $1-\beta_B$ the power with borrowing. We can compute the power of $1-\beta(\alpha_B)$ of a frequentist method at level α_B. By definition, if the frequentist test is UMP, then $1-\beta(\alpha_B) \geq 1-\beta_B$. Our results also provide an experimental confirmation that no gains in power can be obtained at lower or equivalent type 1 error rate by using borrowing methods. Calderazzo et al. (2022) proposed a Bayesian decision-theoretic approach which, in particular, provides a rationale for type 1 error inflation. This approach implies explicitly specifying the cost associated with each type of error. These aspects were not investigated in our simulation study.

Moreover, Kopp-Schneider et al. (2023) show that, in some cases, Bayesian borrowing methods lead to non-UMP tests, so that their power at equivalent type 1 error rate is lower compared to frequentist methods. The behavior of some methods leading to reduced power compared to frequentist methods at equivalent TIE can therefore be interpreted as the corresponding test not being Uniformly Most Powerful. Such a behavior was previously noted by Kopp-Schneider et al. (2023) in case of "extreme borrowing". In these scenarios, the use of borrowing methods is therefore counter-productive. We observed, in particular, that test-then-pool variants tended to display power loss, in particular in case of consistent treatment effect. Moreover, it is not clear why static borrowing methods did not show such power loss in the Botox and Dapagliflozin case studies, despite extreme borrowing. This point would require further investigation.

The "sweet spot" is defined in Viele et al. (2014), in the context of external control borrowing, as a range of drift values between the control arms at which there is an increase in power and a decrease in TIE compared to their nominal values (i.e., compared to the power and TIE of a separate analysis of the target study data). However, in the context of borrowing treatment effect, it is unclear how to generalize this definition. Indeed, in this latter case, the TIE corresponds to the probability of success for a specific drift value. One may wonder whether there exists a range of drift values for which the probability of success is higher than the nominal power, or lower than the nominal TIE. However, we observed, in almost all scenarios and for all methods considered, that the power of the borrowing method is larger than the nominal power in the whole alternative hypothesis space, while the TIE error is inflated. These two observations are a consequence of the fact that the probability of success of the borrowing method is similar to the one of frequentist tests at equivalent TIE, and borrowing induces inflated TIE because borrowed data favours the alternative hypothesis.

12.3 Power gains at equivalent TIE

We reported in the results section that in the Mepolizumab and the Belimumab case studies, some scenarios could lead to apparent power gains with borrowing methods. We observed, that in several cases, the method of interest was a separate Bayesian analysis (for example, a conjugate power prior with a power parameter of zero). In our setting, a Bayesian separate analysis should be equivalent to a frequentist t-test, therefore, how can we interpret this unexpected result, and what does it mean in terms of the limitations of our study?

The reason we identified is the approximation to the sample standard deviation that is implied by the t-test. In our estimation of the power of borrowing methods in the Mepolizumab and Belimumab case studies, we generated data samples according to the true data-generating process. We then assumed a Gaussian likelihood with a known standard deviation for the analysis. However, when analytically determining the power of the t-test, we implicitly assumed a χ^2 distribution for the standard deviation. Therefore, while a Gaussian distribution is assumed for the mean when computing the power for the Bayesian analysis and the t-test, different distributions are assumed for the standard deviation. This explains why we only observed power gains in the Belimumab case study for small sample sizes, and the Mepolizumab case study, and not in the Dapagliflozin and Botox case studies (with Gaussian endpoints) and the Teriflunomide case study (in which the data generating process was approximated by sampling from a Gaussian). A way to circumvent this issue would be to determine the power of the frequentist analysis by simulation. We performed this as a supplemental analysis, in the Belimumab case, and observed that the number of scenarios with power gains decreased dramatically when determining the power by simulation, which confirms our hypothesis regarding the origin of the previously observed power gains.

This result highlights a key requirement when comparing Bayesian and frequentist methods: one must make sure that the comparison between a Bayesian borrowing method and a frequentist test is not impeded by assumptions derived from asymptotic results. A simple way to check this is to compare the frequentist method to a Bayesian method without extrapolation.

12.4 Robustness of adaptive borrowing methods to drift

An absolute increase in drift tends to increase bias, MSE, and reduces the coverage probability of credible intervals of borrowing methods. Moreover, positive drift tends to increase power and TIE rate. An important question regarding the use of such methods is therefore their robustness to drift, which can be defined as the tendency to maintain good operating characteristics as drift increases. There is a trade-off between robustness to drift and benefit from borrowing: a more robust method will favor a reduction in bias at the cost of a larger variance of the estimator. As a measure of robustness, we considered the sweet spots for the different frequentist operating characteristics. This corresponds, for a given operating characteristic, to the range of drift values for which borrowing brings an improvement compared to a separate analysis. Interestingly, the width of the sweet spot for MSE and coverage shrinks as the sample size increases (or the standard deviation) decreases in the target study. This may seem paradoxical, as one may consider that there is less risk in borrowing information if the ratio between the source and target studies' sample size is smaller.

We noticed that some borrowing methods displayed sweet spots relative to MSE that were significantly narrower than with a pooled analysis. These included the test-then-pool variants and the p-value-based Power Prior. This suggests that, while these methods do not systematically fully borrow external information, they are less robust to drift than a pooled analysis.

12.5 Bias and variance of Bayesian borrowing methods as a function of drift

Power gains cannot be obtained while maintaining or lowering the nominal type 1 error rate. The focus on type 1 error (and type 1 error control) and power in a prospective decision-making framework hardly makes sense from a Bayesian decision-theoretic point of view, leading to the seemingly paradoxical absence of power gains with borrowing. If we rather focus on treatment effect estimation accuracy, then we see that gains can be obtained in terms of MSE or coverage within a limited drift range (the MSE and coverage sweet spots). This highlights the key benefit of Bayesian borrowing: in this sweet spot, the bias due to drift is compensated by variance reduction due to the use of an informative prior. If our focus is on the accuracy of the estimation, one would therefore prefer a borrowing method such that the MSE does not increase too much outside the MSE sweet spot, and such that this sweet spot encompasses realistic drift values.

12.6 Comparing borrowing methods

The choice of model parameters modulating information borrowing allows for controlling the amount of borrowing and the response of the operating characteristics to drift. We saw that for methods such as the RMP, the Conditional Power Prior, the Commensurate Power Prior, and the test-then-pool variants, it is possible to adjust the borrowing parameters in the spectrum that goes from no borrowing to pooling. The NPP and the commensurate power prior differ in that regard, as they never fully pool the source and target study data. This may be explained by the fact that, with these priors, the prior on the heterogeneity parameters implies that some posterior probability is always assigned to the possibility of between-studies heterogeneity. At the design stage, during discussions with regulatory agencies, a rationale for the choice of parameters should be provided. Because of the sensitivity of operating characteristics to the method's parameters, it is difficult to directly compare different methods. One approach may be to consider an operating characteristic of main interest, for example, the type 1 error rate, and to compare methods anchored on this operating characteristic (e.g. a target TIE rate of 0.1). This requires calibrating the borrowing parameters to match the target value. We did not consider this in our simulation study design due to the implied computational burden, but future work could consider the following approach:

1. Define the operating characteristic for which we need equivalent value across methods to compare them, and define its target value.
2. In a given scenario, calibrate the method's parameters to reach the target value for the OC of interest.
    - Define the range of parameters considered
    - Define a small number of simulation replicates used only for calibration
    - Use an optimization algorithm to find the parameter values for which the method matches the target OC. Note that for methods, some OCs like the probability of success are monotonic functions of the parameters, which can be leveraged for efficient optimization. Otherwise, black-box derivative-free optimization algorithms (such as Bayesian optimization) could be used.
3. Run a simulation study with the calibrated parameters with a large number of replicates.

However, this approach implies a nested simulation, and can therefore be computationally highly expensive. However, it is practically feasible if the number of scenarios and methods to consider is small. An advantage is that, in addition to allowing a fair comparison between methods, it directly allows anchoring an OC of interest, such as type 1 error rate, to a pre-specified value.

We approached the comparison of borrowing methods by considering whether, at similar type 1 error rate, other characteristics would be more or less improved. Although we were not able to compare method at exactly the same type 1 error rates, since we included many methods and parameters, it was possible to make meaningful comparisons.

12.7 Pros and cons of the different methods.

From a practical perspective, key aspects to consider when using a borrowing method are the following:

- **Control of type 1 error:** It is difficult to control the type 1 error of adaptive borrowing methods, although some methods such as the PDCCPP (Nikolakopoulos et al. 2018) have been proposed that do so (see also Calderazzo and Kopp-Schneider (2022)). Overall, static borrowing methods, in combination with calibration, provide a straightforward way to control type 1 error to a pre-specified value. Static borrowing methods however may have limited or no ability to adapt to drift.
- **Interpretability of the method's parameters**: some methods have free parameters, the value or prior distribution of which are challenging to specify a priori due to the difficulty in interpreting them. This is the case of the Commensurate Power Prior, the Normalized Power Prior, or of the informative prior weight in the RMP. How to translate clinical knowledge into a value or a prior distribution for that parameter may be challenging.
- **Performance at similar type 1 error rate:** In this study, we compared frequentist OCs as a function of TIE across methods. We found that the p-value based PP, the EBPP, and the test-then-pool variants displayed poor performance compared to other methods when considering accuracy (MSE) and uncertainty calibration (coverage probability of the 95% CrI).
- **Underlying assumptions:** Some methods, such as the Conditional Power Prior, assume the treatment effect in both source and target populations in the same, whereas others include separate parameters for both and rely on the assumption of exchangeability between the source and target study.

- **Difficulty of inference** : Implementation of some of the methods investigated may often be difficult as they rely on using MCMC simulation techniques, which may impact the ability to perform inference due to convergence issues, especially in settings with limited data.
- **Computational burden:** this is especially important if a simulation study is required at the design stage, which is highly likely, and if parameters calibration is intended. Methods such as the Normalized Power Prior and the Commensurate Power Prior imply a much larger computational cost when used with a Gaussian likelihood, compared to other methods considered in this study. Moreover, when no Gaussian approximation is used, MCMC is usually required, which implies a much larger computational cost.

12.8 Influence of the type of endpoint

Since we did not vary the type of endpoint independently of other scenario parameters, it was not possible to systematically compare a method's behavior across endpoints. Furthermore, we often used transformations of summary measures suitable for a Normal approximation. However, when comparing the different case studies—particularly the Aprepitant case study, where the likelihood was non-Gaussian—we found no systematic differences in the observed patterns or the relative performance of methods attributable to the type of endpoint.

12.9 Bayesian operating characteristics of the analysis prior

For several of the adaptive borrowing methods investigated, the prior is not fully defined until the data in the target population have been observed. For these methods, it is therefore not possible to compute Bayesian metrics such as the prior probability of benefit or success, or the pre-posterior probabilities with the analysis prior as design prior. However, these quantities, particularly the prior probability of benefit, are important from a regulatory perspective. In these cases, at the design stage, one may consider using a UI design prior or the source posterior as design prior so as to provide bounds on the metric of interest. However, as we saw in our results, these bounds can be very large, therefore giving limited information. One may instead use design priors that correspond to the analysis prior of interest with some fixed hyperparameters, that do not depend on the target data. For example, in the RMP, the variance of the vague distribution should ideally be chosen based on the sample variance in the target study to make it 'worth two subjects from the target population'. Note that this approach is not practically feasible for all methods.

12.10 Measuring uncertainty on Bayesian operating characteristics

One of the advantages of our simulation study compared to previous work in the field, is the rigorous estimation and reporting of uncertainty associated with all inference metrics and frequentist operating characteristics. Indeed, many previous work compared their contributed methods to established methodologies, but the lack of uncertainty reporting limited interpretation.

In this study, however, we did not report uncertainty on Bayesian Operating Characteristics. Indeed, this would incur a very large computational cost when using Monte Carlo integration or numerical integration to estimate these OCs. However, the smooth and consistent patterns we observed when plotting these OCs suggest that the estimates we reported are not affected by substantial simulation error.

13 Conclusions and Recommendations

13.1 Summary

Despite the growing interest in the use of partial extrapolation methods in trials design and analysis, and their recommendation by regulatory authorities to overcome reduced sample size problems (Committee for Medicinal Products for Human Use 2006; Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development 2010; Parmar et al. 2016), their use for treatment effect borrowing, e.g. in rare diseases, remains limited (Partington et al. 2022). Indeed, in practice, no clear guideline exists as to the selection and evaluation of a Bayesian treatment effect borrowing method. This stems, in particular, from the lack of systematic comparison of existing methods in a unified simulation-based assessment framework.

In this study, we performed a large-scale simulation study using realistic scenarios based on real use cases. We explored a wide diversity of scenarios by varying the sample size of the clinical trial in the target population, the magnitude of the treatment effect, the link function between observed outcome and statistical

model for the treatment effect, the variance in the target study, the type of endpoint, as well as the parameters needed to specify the models.

- This study allowed us to empirically confirm previous theoretical results (Kopp-Schneider et al. 2020) by showing that all models are equivalent to the frequentist approach regarding power at equivalent type 1 error, irrespective of the type 1 error rate. This implies that the improved power is simply bought at the expense of type 1 error inflation, and that none of the models outperform frequentist approaches.
- We showed that in numerous scenarios, borrowing could lead to a test that is not UMP and consequently to power loss compared to frequentist tests at equivalent type 1 error.
- We showed that, if type 1 error control is key, then a calibration-based Power Prior may be the most promising approach.
- We determined how the amount of information borrowed from the source population depends on the method parameters and the scenario. In particular, we determined how the borrowing strength varies with the drift between the target and source studies, and how this impacts on operating characteristics.
- We evaluated and compared other operating characteristics of the Bayesian approach that have been proposed.
- We showed that one quantity is key to characterize borrowing methods: the prior ESS for a given drift value, which depends on how adaptive the method is to discrepancies between the source and target studies data.
- By comparing frequentist OCs at similar type 1 error rate, we showed that the p-value based Power Prior, the Test-then-Pool variants and the EBPP methods underperformed compared to other methods -including fixed borrowing methods- in terms of MSE and coverage.
- We identified key requirements for simulation studies involving borrowing methods, including comparison of OCs at equivalent type 1 error, MC error quantification, and the risk in not taking into account differing approximations implied in the computation of power between Bayesian and frequentist methods.

13.2 Simulation studies for evaluating Bayesian borrowing designs

In a seminal paper, Pocock (1976) discussed qualitative comparability criteria to assess if a control group from one study should be used to elicit an informative Bayesian prior for another study (see also Hatswell et al. (2020) for an update) :

1. Control treatment should be the same in both studies.
2. Both studies should have the same requirements for patient eligibility/inclusion and be contemporary.
3. The methods of treatment evaluation must be the same.
4. The distributions of important patient characteristics affecting outcomes should be the same.
5. The historical study must have been performed in the same organization with largely the same clinical investigators.
6. There must be no other indications leading one to expect differing results between the randomized and historical controls.

These intuitive criteria could also be applicable to the borrowing of treatment effects from source trials. But, in practice, in most settings, they will either not be fulfilled or not be verifiable. To overcome this, the potential for heterogeneity between the source and target trial data should carefully be assessed in order to justify borrowing. We observed in our simulation study that, when borrowing information, drift may indeed lead to an overall degradation in operating characteristics, in particular TIE inflation.

In most cases, in particular when using non-conjugate models, analytically controlling the type I error rate of a design making use of dynamic borrowing is intractable (see however Nikolakopoulos et al. (2018) and Calderazzo et al. (2022)). This can be problematic as regulatory agencies tend to strongly prefer statistical methods that do so (Collignon et al. 2018). As a consequence, and because of the uncertainty associated with the performance of a given method, it is usually recommended to run extensive simulation studies tailored to the specific problem at hand and to the available source data. The simulations studies should focus on assessing the operating characteristics of the borrowing method (Committee for Medicinal Products for

Human Use 2022), in line with what is commonly required for Bayesian analysis of clinical trials (US Food and Drug Administration 2010; US Food and Drug Administration 2016; US Food and Drug Administration 2019). For example, the EMA states that "a common approach to addressing the risk of TIE rate inflation when information is borrowed is to carry out multiple simulation studies to quantify this effect." (EMA. Qualification opinion for prognostic covariate adjustment (PROCOVA). EMA; 2022.). This is important both for comparing existing methods, selecting a prior and its hyperparameters (e.g., between-trials variance, power parameter, mixture weights), and estimating the sample size that can be spared.

Our implementation could be reused to carry out such simulation studies in similar scenarios (that is, a single source, potentially based on a meta-analysis of multiple source studies, and without covariates). The simulation framework we developed is highly modular, follows software engineering good practices and interfaces with existing R packages that implement borrowing methods such as RBesT. It makes it simple for users to configure a wide variety of scenarios and could be a valuable tool to extensively investigate designs that use Bayesian borrowing through simulation, analyze clinical trial data using borrowing methods, or perform sensitivity analyses.

However, we noticed that exploring a variety of scenarios in a non-interactive way could be a tedious task. Moreover, the current implementation of the simulation study requires users to manually edit configuration files, which requires some level of understanding of the methods. However, the package we developed throughout this project could be used to build an interactive tool (such as an R Shiny dashboard for example), that could easily be used by statisticians to facilitate communication with clinicians. In pilot experiments, we noticed that 1000 replicates already gave meaningful results. Therefore, determining operating characteristics with reasonable uncertainty is possible, for most methods, within a few seconds, which would make such a dashboard sufficiently reactive for users to explore various designs and methods. Inclusion of covariates could be performed, for example, by interfacing the code with the *hdbayes* package released in April 2024 (`https://github.com/ethan-alt/hdbayes`).

13.3 Choice of partial extrapolation methods

We observed that the p-value-based Power Prior and the test-then-pool variants displayed a much larger MSE at a similar type 1 error compared to other methods. These methods, as well as the EBPP, also showed a strong reduction in the coverage probability of the 95% CrI in case of drift. These elements provide a strong argument against the use of such methods. In each case study, it was not possible to identify a method that would systematically perform better compared to other methods in terms of power gains, estimation accuracy, and coverage. For example, we observed that the RMP with prior weight in the range 0.1 to 0.9 displayed a more robust coverage probability compared to other adaptive borrowing methods, but similar to the conditional power prior. However, a surprising result is the overall good performance of the Conditional Power Prior, a fixed borrowing method, compared to adaptive borrowing method. Over all scenarios and case studies, the Conditional Power Prior was among the best-performing methods when comparing at equivalent type 1 error rate, performing better than the RMP in terms of MSE in many cases. This may seem counterintuitive, as one may expect adaptive borrowing methods to incur lower MSE in the presence of drift. However, one has to consider the fact that, when comparing methods at equivalent type 1 error rate, comparison is performed after adaptation, and therefore at similar prior ESS.

A sensible criterion for judging whether a borrowing method is appropriate is the following: as the target sample size increases and as drift goes to zero, more information should be borrowed. We observed that some methods, such as the NPP, do not fully borrow external information even in the absence of drift. This may seem reasonable, as the similarity between the source and target populations may occur due to random fluctuations. However, the inference process also weights data points based on the variance, so it is not clear whether such a discounting in the prior is needed in this case. That said, this is fine if it helps with the regulatory concern of ensuring that the posterior distribution is driven by the results in the target population in the presence of drift from the source population.

13.4 Sensitivity analyses

From a regulatory perspective, transparency of the methods and interpretability of the borrowing parameters are crucial. At the analysis stage, conducting so-called "credibility analyses" by varying the borrowing parameters enhances transparency by showing how much borrowing is needed to reject the null hypothesis. The credibility of this tipping point is then evaluated by subject-matter experts. A major inconvenience with this approach is that it requires the borrowing parameter to be intepretable in terms of similarity between the source and target. This approach has been applied in a successful clinical trial. Indeed, the FDA approved belimumab in children with systemic lupus erythematosus based on a randomized controlled

trial (NCT01649765) that evaluated belimumab versus placebo in 93 pediatric patients. Determination of efficacy was supported by the use of a robust mixture prior borrowing from the established efficacy of belimumab from two phase 3 adult studies (Pottackal et al. 2019). A tipping point analysis was conducted which consisted in determining the minimum weight to assign to the informative component of the mixture prior in order to reject the null hypothesis.

One may wish to verify that conclusions drawn from Bayesian analysis incorporating external data remain robust across a range of credible assumptions regarding the heterogeneity between the source and target studies and that any sensitivity of the results to these assumptions is well understood. To address this need, Best et al. (2021) proposed a method they describe as a type of "analysis of credibility" or "reverse-Bayes" method, with the goal of assessing "the properties of the prior distribution needed to achieve a certain posterior statement for the given data" (Matthews 2018; Held 2019; Held et al. 2022). Best et al. (2021) applied this approach to a robust mixture prior in a pediatric example to identify the minimum prior probability $Pr(M_{\text{source}})$ needed to be assigned to the patient-level exchangeability assumption to yield statistically significant evidence of treatment benefit in children. The credibility of this tipping point can then be evaluated. In a sense, the idea is to assess the credibility of a conclusion. Essentially, the approach consists in evaluating the credibility of a conclusion by estimating the level of skepticism required to yield a non-significant result. Importantly, Best et al. (2021) note that this approach can be used post-hoc (at the analysis stage) or as part of a pre-planned sensitivity analysis.

When applying a Bayesian borrowing method to a specific scenario, it seems essential to determine the sensitivity of inference results to hypotheses regarding the similarity between the source and target studies, in particular regarding the drift. Therefore, methods should be computationally feasible for extensive simulations to be performed within in a reasonable time. Therefore, computational considerations, often neglected in the Bayesian borrowing literature, are key to ensuring applicability of these methods. Another reason to favor computationally inexpensive methods is the potential use of calibration procedures, in which the method's parameters are adjusted so as to match a specific type 1 error rate. This requires running a derivative-free optimization algorithm (typically a grid search, although more efficient approaches such as Bayesian optimization could be used), which comes at a prohibitive computational cost if inference itself is computationally expensive. These technical conveniences however must be weighed against other relevant aspects, such as the overall performance of the proposed approach in terms of operating characteristics.

We found it especially difficult from a technical point of view to work with cases that do not allow for analytical posterior evaluation or fast numerical integration, that is in the Aprepitant case study and with the Commensurate Power Prior. This stems, in particular, from a bug in CmdStanR that incurred prohibitive computational costs, except on an AWS machine. This raises the question as to whether other approximate Bayesian inference approaches, such as Laplace approximation or variational inference, could be used instead of MCMC in this context. These methods allow for much faster inference but do not benefit from the convergence guarantees of MCMC methods. Moreover, there seems to lack R packages that support the more recent methods for variational inference.

13.5 Choice of likelihood

A systematic comparison of the influence of the likelihood on the performance of borrowing methods was out of the scope of this study. In the Aprepitant case study, the use of a likelihood that more closely reflected the true data-generating process implied a lot of additional computational cost compared to using a Gaussian likelihood for the treatment effect, because of the requirement for MCMC inference. Moreover, methods are usually designed for Gaussian likelihood, and adaptation and implementation in order not to rely on a Gaussian approximation can be complex. Therefore, determining the sample size regimes in which a Gaussian approximation does not impact inference too much could be useful in practice.

13.6 Measuring the prior Effective Sample Size

Our study also highlighted the limitation of the different ESS methods we used. In particular, the ESS measure should account for differences between the source and target study sample standard deviation, so that the prior ESS of a separate or pooled analysis does not depend on the drift. Indeed, in the Belimumab case study for example, an increase in drift induced an increase of the posterior standard error on the summary measure. Given that the ESS measures we used assumed a Gaussian distribution of the treatment effect, this corresponded to an increase in the corresponding standard deviation of the Gaussian. At equivalent information content in the prior but increased estimated standard deviation, this corresponds to a larger prior ESS. We therefore recommend that sanity checks based on pooled or separate analysis be included in simulation studies that aim at determining the prior ESS.

Distribution of the summary statistics	**Normal**	**Binomial**
Source prior	**Normal**	**Uniform**
Conditional power prior	Analytic	MCMC
Normalized PP	MCMC	MCMC
Empirical Bayes PP	Analytic	MCMC
p-value based PP	Analytic	MCMC
PDCCPP	Analytic + Optimization	NA
Commensurate PP	MCMC	MCMC
Robust mixture prior	Analytic	MCMC

Table 15: Methods used for inference. "Optimization" corresponds to cases where the method requires (gradient-free) optimization to determine the value of some parameters.

13.7 Funding

This project was funded through the reopening of competition no. 02 under framework contract following procurement procedure EMA/2020/46/TDA (Lot 3).

13.8 Acknowledgements

This work was granted access to the HPC/AI resources of the National Computer Center for Higher Education (CINES) under the allocation 2024-AD010315186 made by the Grand Equipement National de Calcul Intensif (GENCI).

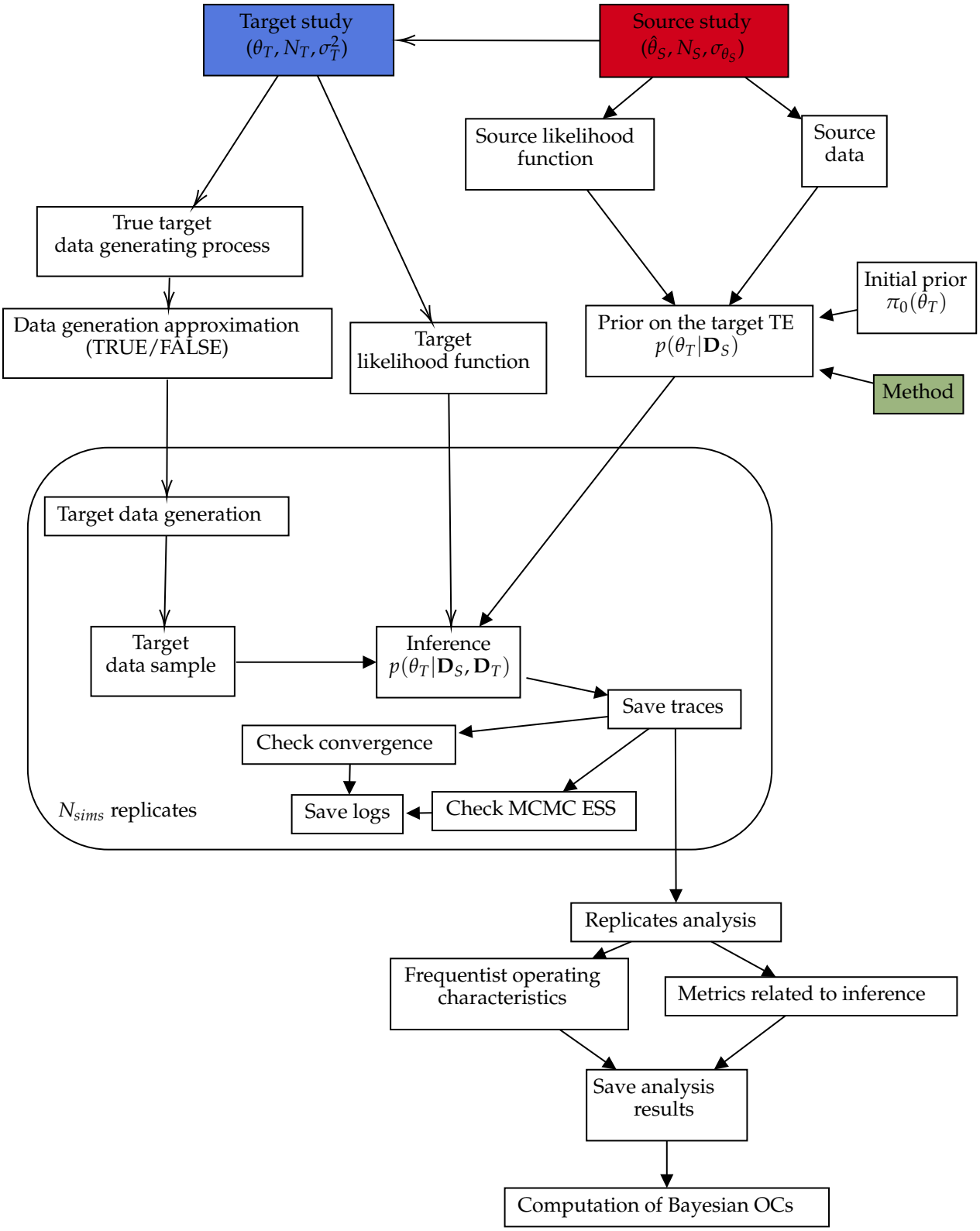

Figure 77: Summary of the simulation study pipeline. Colored boxes correspond to components of the configuration that will be varied.

References

F. M. Al Amer, C. G. Thompson, and L. Lin (2021). "Bayesian Methods for Meta-Analyses of Binary Outcomes: Implementations, Examples, and Impact of Priors". *International Journal of Environmental Research and Public Health* 18.7, 3492.

C. J. Bailey, J. L. Gross, D. Hennicken, N. Iqbal, T. A. Mansfield, and J. F. List (2013). "Dapagliflozin add-on to metformin in type 2 diabetes inadequately controlled with metformin: a randomized, double-blind, placebo-controlled 102-week trial". *BMC medicine* 11, 43.

C. J. Bailey, J. L. Gross, A. Pieters, A. Bastien, and J. F. List (2010). "Effect of dapagliflozin in patients with type 2 diabetes who have inadequate glycaemic control with metformin: a randomised, double-blind, placebo-controlled trial". *Lancet (London, England)* 375.9733, 2223–2233.

N. Best, M. Ajimi, B. Neuenschwander, G. Saint-Hilary, and S. Wandel (2023). *Beyond the classical type I error: Bayesian metrics for Bayesian designs using informative priors*.

N. Best, R. G. Price, I. J. Pouliquen, and O. N. Keene (2021). "Assessing efficacy in important subgroups in confirmatory trials: An example using Bayesian dynamic borrowing". *Pharmaceutical Statistics* 20.3, 551–562.

F. Bovis, M. Ponzano, A. Signori, I. Schiavetti, P. Bruzzi, and M. P. Sormani (2022). "Reinterpreting Clinical Trials in Children With Multiple Sclerosis Using a Bayesian Approach". *JAMA Neurology* 79.8, 821.

C. Brard, L. V. Hampson, N. Gaspar, M.-C. Le Deley, and G. Le Teuff (2019). "Incorporating individual historical controls and aggregate treatment effect estimates into a Bayesian survival trial: a simulation study". *BMC medical research methodology* 19.1, 85.

H. I. Brunner, C. Abud-Mendoza, M. Mori, C. A. Pilkington, R. Syed, S. Takei, D. O. Viola, R. A. Furie, S. Navarra, F. Zhang, D. L. Bass, G. Eriksson, A. E. Hammer, B. N. Ji, M. Okily, D. A. Roth, H. Quasny, and N. Ruperto (2021). "Efficacy and safety of belimumab in paediatric and adult patients with systemic lupus erythematosus: an across-study comparison". *RMD Open* 7.3, e001747.

H. I. Brunner, C. Abud-Mendoza, D. O. Viola, I. Calvo Penades, D. Levy, J. Anton, J. E. Calderon, V. G. Chasnyk, M. A. Ferrandiz, V. Keltsev, M. E. Paz Gastanaga, M. Shishov, A. L. Boteanu, M. Henrickson, D. Bass, K. Clark, A. Hammer, B. N. Ji, A. Nino, D. A. Roth, H. Struemper, M.-L. Wang, A. Martini, D. Lovell, and N. Ruperto (2020). "Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial". *Annals of the Rheumatic Diseases* 79.10, 1340–1348.

S. Calderazzo and A. Kopp-Schneider (2022). *Robust incorporation of historical information with known type I error rate inflation*.

S. Calderazzo, M. Wiesenfarth, and A. Kopp-Schneider (2022). "A decision-theoretic approach to Bayesian clinical trial design and evaluation of robustness to prior-data conflict". *Biostatistics (Oxford, England)* 23.1, 328–344.

G. Campbell (2017). "Bayesian methods in clinical trials with applications to medical devices". *Communications for Statistical Applications and Methods* 24.6, 561–581.

M.-H. Chen, J. G. Ibrahim, H. Amy Xia, T. Liu, and V. Hennessey (2014). "Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program: Bayesian Meta-experimental Design". *Statistics in Medicine* 33.9, 1600–1618.

T. Chitnis et al. (2021). "Safety and efficacy of teriflunomide in paediatric multiple sclerosis (TERIKIDS): a multicentre, double-blind, phase 3, randomised, placebo-controlled trial". *The Lancet Neurology* 20.12, 1001–1011.

Y. Chu and Y. Yuan (2018). "A Bayesian basket trial design using a calibrated Bayesian hierarchical model". *Clinical Trials (London, England)* 15.2, 149–158.

C. Chuang-Stein and S. Kirby (2017). *Quantitative Decisions in Drug Development*. Springer Series in Pharmaceutical Statistics. Cham: Springer International Publishing.

O. Collignon, F. Koenig, A. Koch, R. J. Hemmings, F. Pétavy, A. Saint-Raymond, M. Papaluca-Amati, and M. Posch (2018). "Adaptive designs in clinical trials: from scientific advice to marketing authorisation to the European Medicine Agency". *Trials* 19.1, 642.

E. M. A. Committee for Medicinal Products for Human Use (2006). *Guideline on Clinical Trials in Small Populations*.

E. M. A. Committee for Medicinal Products for Human Use (2022). *Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™)*.

C. Confavreux, P. O'Connor, G. Comi, M. S. Freedman, A. E. Miller, T. P. Olsson, J. S. Wolinsky, T. Bagulho, J.-L. Delhay, D. Dukovic, P. Truffinet, and L. Kappos (2014). "Oral teriflunomide for patients with relapsing multiple sclerosis (TOWER): a randomised, double-blind, placebo-controlled, phase 3 trial". *The Lancet Neurology* 13.3, 247–256.

P. Diemunsch, T. J. Gan, B. K. Philip, M. J. Girao, L. Eberhart, M. G. Irwin, J. Pueyo, J. E. Chelly, A. D. Carides, T. Reiss, J. K. Evans, F. C. Lawson, and Aprepitant-PONV Protocol 091 International Study Group (2007). "Single-dose aprepitant vs ondansetron for the prevention of postoperative nausea and vomiting: a randomized, double-blind phase III trial in patients undergoing open abdominal surgery". *British Journal of Anaesthesia* 99.2, 202–211.

Y. Duan, K. Ye, E. P. Smith, and E. Smith (2006). "Evaluating water quality using power priors to incorporate historical information". *Environmetrics* 17.1, 95–106.

M. Feißt, J. Krisam, and M. Kieser (2020). "Incorporating historical two-arm data in clinical trials with binary outcome: A practical approach". *Pharmaceutical Statistics* 19.5, 662–678.

I. Gravestock and L. Held (2019). "Power priors based on multiple historical studies for binary outcomes." *Biometrical Journal. Biometrische Zeitschrift* 61.5, 1201–1218.

I. Gravestock, L. Held, and COMBACTE-Net consortium (2017). "Adaptive power priors with empirical Bayes for clinical trials: Adaptive Power Priors with Empirical Bayes for Clinical Trials". *Pharmaceutical Statistics* 16.5, 349–360.

J. B. Greenhouse and L. Waserman (1995). "Robust bayesian methods for monitoring clinical trials". *Statistics in Medicine* 14.12, 1379–1391.

B. Han, J. Zhan, Z. John Zhong, D. Liu, and S. Lindborg (2017). "Covariate-adjusted borrowing of historical control data in randomized clinical trials". *Pharmaceutical Statistics* 16.4, 296–308.

A. Hatswell, N. Freemantle, G. Baio, E. Lesaffre, and J. van Rosmalen (2020). "Summarising salient information on historical controls: A structured assessment of validity and comparability across studies". *Clinical Trials (London, England)* 17.6, 607–616.

L. Held (2019). "The assessment of intrinsic credibility and a new argument for $p < 0.005$". *Royal Society Open Science* 6.3, 181534.

L. Held, R. Matthews, M. Ott, and S. Pawel (2022). "Reverse-Bayes methods for evidence assessment and research synthesis". *Research Synthesis Methods* 13.3, 295–314.

B. P. Hobbs, B. P. Carlin, S. J. Mandrekar, and D. J. Sargent (2011). "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials". *Biometrics* 67.3, 1047–1056.

M. D. Hoffman and A. Gelman (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo".

B. Holzhauer, C. Wang, and H. Schmidli (2018). "Evidence synthesis from aggregate recurrent event data for clinical trial design and analysis". *Statistics in Medicine* 37.6, 867–882.

B. Hupf, V. Bunn, J. Lin, and C. Dong (2021). "Bayesian semiparametric meta-analytic-predictive prior for historical control borrowing in clinical trials". *Statistics in Medicine* 40.14, 3385–3399.

J. G. Ibrahim and M.-H. Chen (2000). "Power prior distributions for regression models". *Statistical Science* 15.1, 46–60.

J. G. Ibrahim, M.-H. Chen, H. A. Xia, and T. Liu (2012). "Bayesian meta-experimental design: evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes". *Biometrics* 68.2, 578–586.

Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development (2010). *Rare Diseases and Orphan Products: Accelerating Research and Development*. Ed. by M. J. Field and T. F. Boat. The National Academies Collection: Reports funded by National Institutes of Health. Washington (DC): National Academies Press (US).

L. Jiang, L. Nie, and Y. Yuan (2021). "Elastic priors to dynamically borrow information from historical data in clinical trials". *Biometrics* 79.1, 49–60.

F. Jiao, W. Tu, S. Jimenez, V. Crentsil, and Y.-F. Chen (2019). "Utilizing shared internal control arms and historical information in small-sized platform clinical trials". *Journal of Biopharmaceutical Statistics* 29.5, 845–859.

H. Jin and G. Yin (2021). "Unit information prior for adaptive information borrowing from multiple historical datasets". *Statistics in Medicine* 40.25, 5657–5672.

M. Jin, Q. Li, and A. Kaur (2021). "Bayesian Design for Pediatric Clinical Trials with Binary Endpoints When Borrowing Historical Information of Treatment Effect". *Therapeutic Innovation & Regulatory Science* 55.2, 360–369.

A. M. Kaizer, J. S. Koopmeiners, and B. P. Hobbs (2018). "Bayesian hierarchical modeling based on multisource exchangeability". *Biostatistics (Oxford, England)* 19.2, 169–184.

O. Keene, N. Best, R. Price, and I. Pouliquen (2020). "Use of a novel Bayesian borrowing statistical method to assess efficacy of mepolizumab in adolescents". In: *Paediatric asthma and allergy*. ERS International Congress 2020 abstracts. European Respiratory Society, 667.

S. E. Kern (2009). "Challenges in conducting clinical trials in children: approaches for improving performance". *Expert Review of Clinical Pharmacology* 2.6, 609–617.

A. Kopp-Schneider, S. Calderazzo, and M. Wiesenfarth (2020). "Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control". *Biometrical Journal. Biometrische Zeitschrift* 62.2, 361–374.
A. Kopp-Schneider, M. Wiesenfarth, L. Held, and S. Calderazzo (2023). "Simulating and reporting frequentist operating characteristics of clinical trials that borrow external information". *arXiv*.
J. Lim, L. Wang, N. Best, J. Liu, J. Yuan, F. Yong, L. Zhang, R. Walley, A. Gosselin, R. Roebling, and K. Viele (2020). "Reducing Patient Burden in Clinical Trials Through the Use of Historical Controls: Appropriate Selection of Historical Data to Minimize Risk of Bias". *Therapeutic Innovation & Regulatory Science* 54.4, 850–860.
G. F. Liu (2018). "A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint". *Pharmaceutical Statistics* 17.1, 61–73.
R. A. J. Matthews (2018). "Beyond 'significance': principles and practice of the Analysis of Credibility". *Royal Society Open Science* 5.1, 171047.
T. P. Morris, I. R. White, and M. J. Crowther (2019). "Using simulation studies to evaluate statistical methods". *Statistics in Medicine* 38.11, 2074–2102.
B. Neelon and A. J. O'Malley (2010). "Bayesian Analysis Using Power Priors with Application to Pediatric Quality of Care". *Journal of biometrics & biostatistics* 2010.1, 1–9.
B. Neuenschwander, M. Branson, and D. J. Spiegelhalter (2009). "A note on the power prior: A NOTE ON THE POWER PRIOR". *Statistics in Medicine* 28.28, 3562–3566.
B. Neuenschwander, S. Weber, H. Schmidli, and A. O'Hagan (2020). "Predictively consistent prior effective sample sizes". *Biometrics* 76.2, 578–587.
S. Nikolakopoulos, I. van der Tweel, and K. C. B. Roes (2018). "Dynamic borrowing through empirical power priors that control type I error: Dynamic Borrowing with Type I Error Control". *Biometrics* 74.3, 874–880.
P. O'Connor, J. S. Wolinsky, C. Confavreux, G. Comi, L. Kappos, T. P. Olsson, H. Benzerdjeb, P. Truffinet, L. Wang, A. Miller, and M. S. Freedman (2011). "Randomized Trial of Oral Teriflunomide for Relapsing Multiple Sclerosis". *New England Journal of Medicine* 365.14, 1293–1303.
A. O'Hagan, J. W. Stevens, and M. J. Campbell (2005). "Assurance in clinical trial design". *Pharmaceutical Statistics* 4.3, 187–201.
H. G. Ortega, M. C. Liu, I. D. Pavord, G. G. Brusselle, J. M. FitzGerald, A. Chetta, M. Humbert, L. E. Katz, O. N. Keene, S. W. Yancey, and P. Chanez (2014). "Mepolizumab Treatment in Patients with Severe Eosinophilic Asthma". *New England Journal of Medicine* 371.13, 1198–1207.
H. Pan, Y. Yuan, and J. Xia (2017). "A Calibrated Power Prior Approach to Borrow Information from Historical Data with Application to Biosimilar Clinical Trials". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.5, 979–996.
J. Pan, V. Bunn, B. Hupf, and J. Lin (2022). "Bayesian Additive Regression Trees (BART) with covariate adjusted borrowing in subgroup analyses". *Journal of Biopharmaceutical Statistics* 32.4, 613–626.
M. K. B. Parmar, M. R. Sydes, and T. P. Morris (2016). "How do you design randomised trials for smaller populations? A framework". *BMC Medicine* 14.1, 183.
G. Partington, S. Cro, A. Mason, R. Phillips, and V. Cornelius (2022). "Design and analysis features used in small population and rare disease trials: A targeted review". *Journal of Clinical Epidemiology* 144, 93–101.
S. Pawel, F. Aust, L. Held, and E.-J. Wagenmakers (2023). "Normalized power priors always discount historical data". *Stat* 12.1, e591.
S. J. Pocock (1976). "The combination of randomized and historical controls in clinical trials." *Journal of Chronic Diseases* 29.3, 175–188.
G. Pottackal, J. Travis, R. Neuner, R. Rothwell, G. Levin, L. Nie, J. Niu, A. Marathe, and N. Nikolov (2019). "Application of Bayesian Statistics to Support Approval of Intravenous Belimumab in Children with Systemic Lupus Erythematosus in the United States". In: Arthritis Rheumatol.
M. A. Psioda and J. G. Ibrahim (2019). "Bayesian clinical trial design using historical data that inform the treatment effect". *Biostatistics (Oxford, England)* 20.3, 400–415.
M. A. Psioda and X. Xue (2020). "A Bayesian Adaptive Two-stage Design for Pediatric Clinical Trials". *Journal of Biopharmaceutical Statistics* 30.6, 1091–1108.
J. van Rosmalen, D. Dejardin, Y. van Norden, B. Löwenberg, and E. Lesaffre (2018). "Including historical data in the analysis of clinical trials: Is it worth the effort?" *Statistical Methods in Medical Research* 27.10, 3167–3182.
C. Röver, S. Wandel, and T. Friede (2019). "Model averaging for robust extrapolation in evidence synthesis". *Statistics in Medicine* 38.4, 674–694.
F. T. Salman, C. DiCristina, A. Chain, and A. S. Afzal (2019). "Pharmacokinetics and pharmacodynamics of aprepitant for the prevention of postoperative nausea and vomiting in pediatric subjects". *Journal of Pediatric Surgery* 54.7, 1384–1390.

H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander (2014). "Robust meta-analytic-predictive priors in clinical trials with historical control information: Robust Meta-Analytic-Predictive Priors". *Biometrics* 70.4, 1023–1032.

D. J. Schuirmann (1987). "A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability". *Journal of Pharmacokinetics and Biopharmaceutics* 15.6, 657–680.

N. Shehadeh, T. Barrett, P. Galassetti, C. Karlsson, J. Monyak, N. Iqbal, and W. V. Tamborlane (2023). "Dapagliflozin or Saxagliptin in Pediatric Type 2 Diabetes". *NEJM Evidence*.

Y. Shi, W. Li, and G. F. Liu (2023). "A novel power prior approach for borrowing historical control data in clinical trials". *Statistical Methods in Medical Research* 32.3, 9622802221146309.

T. C. Smith, D. J. Spiegelhalter, and A. Thomas (1995). "Bayesian approaches to random-effects meta-analysis: A comparative study". *Statistics in Medicine* 14.24, 2685–2699.

D. J. Spiegelhalter and L. S. Freedman (1986). "A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion". *Statistics in Medicine* 5.1, 1–13.

L. Su, X. Chen, J. Zhang, and F. Yan (2022). "Comparative Study of Bayesian Information Borrowing Methods in Oncology Clinical Trials". *JCO precision oncology* 6.6, e2100394.

L. Thompson, J. Chu, J. Xu, X. Li, R. Nair, and R. Tiwari (2021). "Dynamic borrowing from a single prior data source using the conditional power prior". *Journal of Biopharmaceutical Statistics* 31.4, 403–424.

US Food and Drug Administration (2010). *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials - Guidance for Industry and FDA Staff*.

US Food and Drug Administration (2016). *Adaptive Designs for Medical Device Clinical Studies - Guidance for Industry and FDA Staff*.

US Food and Drug Administration (2019). *Adaptive Designs for Clinical Trials of Drugs and Biologics - Guidance for Industry*.

US House of Representatives (2015). *The 21st Century Cures Act. 2015. Report No.: R44071.*

K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, S. Micallef, S. Roychoudhury, and L. Thompson (2014). "Use of historical control data for assessing treatment effects in clinical trials." *Pharmaceutical Statistics* 13.1, 41–54.

K. Viele, L. M. Mundy, R. B. Noble, G. Li, K. Broglio, and J. D. Wetherington (2018). "Phase 3 adaptive trial design options in treatment of complicated urinary tract infection". *Pharmaceutical Statistics* 17.6, 811–822.

Y. Wang, J. Travis, and B. Gajewski (2022). "Bayesian adaptive design for pediatric clinical trials incorporating a community of prior beliefs". *BMC medical research methodology* 22.1, 118.

K. Weber, R. Hemmings, and A. Koch (2018). "How to use prior knowledge and still give new data a chance?" *Pharmaceutical Statistics* 17.4, 329–341.

P. Yang, Y. Zhao, L. Nie, J. Vallejo, and Y. Yuan (2023). *SAM: Self-adapting Mixture Prior to Dynamically Borrow Information from Historical Data in Clinical Trials*.

Appendices

Appendix A Standard deviation of the sample quantiles

To determine the standard deviation of sample quantiles, we follow the following reasoning: let Y be a continuous random variable with probability density function f, for which we have a sample of size n. We are interested in determining the distribution of the sample median and 0.975th quantile, denoted X_q (with $q_1 = 0.5$ and $q_2 = 0.975$ respectively). We adapt the reasoning developed by Dr William A. Huber in `https://stats.stackexchange.com/a/86804/919`.

Let's denote G_q the c.d.f. of $Beta(\alpha, \beta)$, with $\alpha = qn + 1$ and $\beta = (1-q)n + 1$. Then, the c.d.f. of X_q in x is $G_q(F(x))$, so that the p.d.f. of X_q is: $\frac{\partial G_q \circ F}{\partial x}(x) = g_q(F(x))f(x)$.

So the p.d.f. of the sample quantile is $g_q(F(x))f(x)$.

Now we are interested in approximating the variance of this distribution.

By denoting $\mu_q = F^{-1}(q)$, we have, for sufficiently well-behaved F:

$$\begin{aligned} F(x) &= F(\mu_q + (x - \mu_q)) \\ &\approx F(\mu_q) + F'(\mu_q)(x - \mu_q) \\ &\approx q + f(\mu_q)(x - \mu_q) \end{aligned} \tag{34}$$

So, assuming f is continuous near μ_q, the p.d.f. of X_q is approximately : $g_q(q + f(\mu_q)(x - \mu_q))f(\mu_q)$. This is essentially a shift of the location and scale of the Beta distribution. The variance of $Beta(\alpha, \beta)$ is :

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

so that the variance of the sample quantile is approximately:

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)f(F^{-1}(q))^2},$$

So, for large n, this variance can be approximated as : $\frac{q(1-q)}{nf(F^{-1}(q))^2}$. So for two different quantiles q_1 and q_2, the ratio of standard error on the sample quantile is approximately :

$$\sqrt{\frac{q_1(1-q_1)}{q_2(1-q_2)}}\frac{f(F^{-1}(q_2))}{f(F^{-1}(q_1))}$$

For the standard normal distribution, with $q_1 = 0.5$ and $q_2 = 0.975$, this gives a ratio of 0.47.