Robust incorporation of historical information with known type I error rate inflation

Silvia Calderazzo | Manuel Wiesenfarth | Annette Kopp-Schneider

Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany

Correspondence
Silvia Calderazzo, Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 581, 69120 Heidelberg, Germany.
Email: s.calderazzo@dkfz.de

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract
Bayesian clinical trials can benefit from available historical information through the specification of informative prior distributions. Concerns are however often raised about the potential for prior-data conflict and the impact of Bayes test decisions on frequentist operating characteristics, with particular attention being assigned to inflation of type I error (TIE) rates. This motivates the development of principled borrowing mechanisms, that strike a balance between frequentist and Bayesian decisions. Ideally, the trust assigned to historical information defines the degree of robustness to prior-data conflict one is willing to sacrifice. However, such relationship is often not directly available when explicitly considering inflation of TIE rates. We build on available literature relating frequentist and Bayesian test decisions, and investigate a rationale for inflation of TIE rate which explicitly and linearly relates the amount of borrowing and the amount of TIE rate inflation in one-arm studies. A novel dynamic borrowing mechanism tailored to hypothesis testing is additionally proposed. We show that, while dynamic borrowing prevents the possibility to obtain a simple closed-form TIE rate computation, an explicit upper bound can still be enforced. Connections with the robust mixture prior approach, particularly in relation to the choice of the mixture weight and robust component, are made. Simulations are performed to show the properties of the approach for normal and binomial outcomes, and an exemplary application is demonstrated in a case study.

KEYWORDS
Bayesian trial design, borrowing of historical information, robust borrowing, type I error rate

1 INTRODUCTION

While the adoption of Bayesian clinical trial designs is on the rise, particularly in early phase trials, additional reporting and “reasonable” control of frequentist operating characteristics has often been a requirement of regulatory authorities (Food & Drug Administration, 2010). A key advantage of Bayesian designs is the possibility to include in the analysis relevant historical information about the model parameters through the specification of informative prior distributions. If
such information is consistent with the information collected during the current trial, improvement of the design operating characteristics, for example, in terms of test error rates and estimation error, can generally be achieved. However, if prior-data conflict is present, losses can be significant. To deal with this problem, robustification approaches have been proposed, in the form of static or dynamic (i.e., depending on the currently observed data) discounting of the historical information conveyed by the prior. While both static and dynamic robustification approaches can decrease the impact of prior-data conflict, no power gains can be achieved when strict control of type I error (TIE) rate is required if a uniformly most powerful (UMP) test is available (Kopp-Schneider et al., 2020). Available robustification approaches, for example, the robust mixture prior (see, e.g., Berger & Berliner, 1986; Schmidli et al., 2014), the power prior (see, e.g., Ibrahim & Chen, 2000), the commensurate prior (see, e.g., Hobbs et al., 2012) require the specification of additional parameters and/or distributions. Such choices, are not typically intuitively related to their impact on the control of frequentist operating characteristics. Selection of a borrowing weight based on explicit control of TIE rate, is investigated in the power prior context by Nikolakopoulos et al. (2018). The approach does however not allow a simple interpretation of the weight in terms of TIE rate inflation.

The purpose of the present work is to investigate and characterize a compromise solution between Bayesian and frequentist test decisions controlling TIE rate at a prespecified “standard” level in one-arm studies. The proposed compromise solution can be obtained as a Bayes test decision with adjusted posterior probability test thresholds, or equivalently as a frequentist test decision with an adjusted rejection region. The solution is derived in the spirit of Berger (1985) who states that in a normal prior—normal likelihood case “the Bayesian method can be thought of as providing a rational way of choosing the size of the test,” but with the aim of achieving a gradual compromise between Bayesian and frequentist test decisions. A compromise of this type was first investigated in Hodges and Lehmann (1952), where restricted Bayes solutions were introduced as a general tool to characterize decisions minimizing the integrated risk under a maximum frequentist risk constraint. While the core principle, that is, a compromise on test decision thresholds, stays the same, we construct such a compromise by introducing a weighting of historical information which linearly relates to the TIE rate inflation.

It has been noted that different amounts of tolerated TIE rate inflation can be acceptable in different situations: indeed, more informative or optimistic priors may induce a stronger TIE rate inflation without necessarily representing lower quality information (Travis et al., 2023). One possibility to cope with this situation is to take into account observed divergences between current and historical data outcomes when selecting a borrowing weight. We therefore additionally propose a mechanism for dynamic weight choice which is tailored to testing.

Section 2 outlines the decision-theoretic background and reviews the restricted Bayes approach of Hodges and Lehmann (1952). Section 3 describes the proposed approach and its relationship to other robust borrowing approaches. Section 4 presents exemplary applications to simulated data for normal outcomes. Section 5 shows an application to a paediatric trial. We conclude the article with a discussion of the results and an outlook on future work.

2 | HYPOTHESIS TESTING UNDER THE WEIGHTED 0-1 LOSS

2.1 | Definitions and setup

We shortly introduce the concepts and notation which will be used in the following. Note that a similar setup has been considered in Calderazzo et al. (2022) to which we refer for additional details. We consider the situation of testing \( H_0 : \theta \leq \theta_0 \) versus the alternative hypothesis \( H_1 : \theta > \theta_0 \). We denote by \( y \) the observed data having probability density function \( f(y|n, \theta) \) indexed by \( \theta \) and sample size \( n \), and by \( \pi(\theta) \) the prior distribution for \( \theta \). A test decision \( d \) has to be taken for rejecting \((d = 1)\) or keeping \((d = 0)\) the null hypothesis. A 0-\( \kappa \) loss function \( L(\theta, d_\kappa(y)) \) assigns unit cost to a TIE, a cost \( \kappa \) to a type II error, and no cost if no test error is made. Note that only the ratio between the cost of a type I and a type II error is relevant, so the unit cost for the TIE is assumed without loss of generality.

If no prior information or belief about \( \theta \) is available, a test decision can be undertaken following the frequentist approach. Here, the aim is to minimize over the parameter space \( \Theta \) the maximum of the frequentist risk

\[
R(\theta, d_\kappa(y)) = I(\theta \leq \theta_0)\beta(\theta) + \kappa I(\theta > \theta_0)(1 - \beta(\theta)),
\]

where

\[
\beta(\theta) = P^I(d_\kappa(y) = 1|\theta)
\]
denotes the rejection probability, $I$ is the indicator function, and $P^f$ is the probability computed with respect to the distribution $f$, that is, the frequentist risk is obtained by averaging the loss function with respect to the data density $f$. The maximum of $R(\theta, d_\nu(y))$ over $\Theta$ is given by the maximum between $\beta(\theta_0)$ and $\kappa[1 - \beta(\theta_0)]$, if $\beta(\theta)$ is strictly increasing in $\theta$, as in, for example, the case of normally distributed observations (see Figure 1). The maximum frequentist risk is then minimized by taking $d_\nu(y)$ such that $\beta(\theta_0) = \kappa[1 - \beta(\theta_0)]$. Such a decision will be referred to in the following as “Frequentist decision” (FD). It follows that the maximum TIE rate $\beta(\theta_0)$ is equal to $\tau = \kappa/(1 + \kappa)$. The frequentist approach is often described as an “ultra-pessimistic” approach, which aims at minimizing the loss under the worst-case scenario.

The Bayesian approach associates uncertainty to the parameter while conditioning on the observed data, and thus aims at minimizing the posterior expected loss

$$\rho(\pi, d_\nu(y)|y) = P^\pi(\theta \leq \theta_0|y)I(d_\nu(y) = 1) + \kappa P^\pi(\theta > \theta_0|y)I(d_\nu(y) = 0).$$

The posterior expected loss $\rho$ is obtained by averaging the loss function with respect to the posterior density $\pi(\theta|y)$, and is minimized by the “Bayes decision” (BD) with respect to prior $\pi$, $d_\nu^\pi(y)$, which rejects $H_0$ if

$$P^\pi(\theta \leq \theta_0|y) < \tau$$

with $\tau = \kappa/(1 + \kappa)$. The underlying assumption is that $\pi$ is a reliable summary of the available information or prior beliefs about the parameter $\theta$. Uncertainty about the data outcomes can be taken into account by computing the integrated risk

$$r(\pi, d_\nu(y)) = \int_{-\infty}^{\theta_0} \beta(\theta)\pi(\theta)d\theta + \kappa \int_{\theta_0}^{\infty} (1 - \beta(\theta))\pi(\theta)d\theta,$$

where the loss function is averaged with respect to both the data and the prior distribution. The integrated risk is also minimized by the BD $d_\nu^\pi(y)$ which minimizes the posterior expected loss for each $y \in Y$ (see, e.g., Robert, 2007), and leads to a rejection probability $\beta^\pi(\theta) = P^f(d_\nu^\pi(y) = 1|\theta)$.

### 2.2 Restricted Bayes decisions

Often, a compromise between FDs and BDs induced by an informative prior distribution is sought. In the proposal of Hodges and Lehmann (1952), this compromise is characterized as a “restricted Bayes solution” (henceforth referred to as the “restricted Bayes decision,” RBD), in which the optimal decision is the one minimizing the integrated risk with respect
to an informative prior distribution $\pi$, subjected to a constraint on the maximum frequentist risk (see Berger, 1985, for an overview). Formally, $d^\nu$ is defined to be an RBD with respect to a prior distribution $\pi$ subjected to $\sup_{\theta \in \Theta} R(\theta, d) = C$ if, for a given constant $\eta \in [0, 1]$, it minimizes

$$
\eta r(\pi, d) + (1 - \eta) \sup_{\theta \in \Theta} R(\theta, d^\nu) = C. 
$$

The value of $C$ thus represents the maximum frequentist risk one is willing to accept when incorporating prior information in the decision. The RBD can also be identified as the BD with respect to a specific prior distribution, that is, a distribution of the form $\nu = \eta \pi + (1 - \eta) \mu_0$, for given constant $\eta \in [0, 1]$ and distribution $\mu_0$, such that $\int_{\Theta} R(\theta, d^\nu) \mu_0(\theta) d\theta = \sup_{\theta \in \Theta} R(\theta, d^\nu)$ (Hodges & Lehmann, 1952).

If the frequentist risk has a unique maximum, then $\mu_0$ corresponds to a point-mass prior assigning all probability to the point at which such maximum is achieved (Bayram & Gezici, 2011). However, in the context of clinical trial design, interest is typically placed on constraining TIE rate, rather than the maximum frequentist risk. If the maximum TIE rate is achieved at $\theta_0$ (i.e., $\beta(\theta)$ is strictly increasing in $\theta$), then we define the “TIE rate restricted Bayes decision” (TI-RBD) as the one minimizing

$$
\eta r(\pi, d_\kappa) + (1 - \eta) R(\theta_0, d_\kappa) 
$$

for a given $\eta \in [0, 1]$, and corresponding to the BD with respect to $\nu = \eta \pi + (1 - \eta) \mu_0$, where $\mu_0$ is the point-mass distribution at $\theta_0$ (see Appendix A.1 for a proof). Recall that we adopt the notation $d_\kappa$ for a decision under the $0-\kappa$ loss. Note that the TI-RBD differs from the BD only if prior information is favoring the alternative hypothesis, as we are not capping the frequentist risk, but the TIE rate (see Figure 1 for a graphical example). It is worth noting that an asymmetry in terms of the relative importance of type I and type II error rates is thus introduced in formulation (2).

Formulation (2) is only reported for completeness, as it highlights the underlying trade-offs of the strategy. A more straightforward implementation of the TI-RBD comes from noting that for families of distributions with a monotone likelihood ratio, a UMP test exists at each chosen TIE rate level (Lehmann, 1986). Moreover, for data distributions with monotone likelihood ratio and nondegenerate prior distributions, posterior densities are stochastically ordered with respect to the data outcomes (Milgrom, 1981; Whitt, 1979). This means in turn that the decision to keep or reject $H_0$ is a monotone function of the data (through the sufficient statistics), and therefore the test is UMP with TIE rate $\beta^\pi(\theta_0)$. Note that such a result has been proven for fixed and nondegenerate priors. When the prior is dynamically adapted according to the observed commensurability of current and historical information, the resulting test may not be UMP. Extreme situations where UMP property of the Bayesian test is lost under a dynamic prior choice are shown in Kopp-Schneider et al. (2020, 2023). The TI-RBDs may therefore be obtained by modifying the rejection region of the frequentist test in the “direction” suggested by the historical information. Such an approach can also be adopted to solve (1), as exemplified by Hodges and Lehmann (1952). Here, we discuss and show the applicability of such an approach to a wider set of situations targeting (2), and to different choices of the prior distribution $\pi$.

3 COMPROMISE DECISION

To highlight the relationship between maximum TIE rate, cost of each test decision, and incorporation of prior information, recall that the BD $d^\pi_\kappa$ under prior $\pi$ is to reject $H_0$ if $P^\pi(\theta \leq \theta_0|y) < \tau$, where $\tau = \kappa/(1 + \kappa)$. Assuming $\beta^\pi(\theta)$ to be increasing for $\theta \leq \theta_0$, denote $\beta^\pi(\theta_0)$ the maximum TIE rate of such a procedure; this quantity is typically not analytically available and depends on both the prior density $\pi$ and the cost ratio $\kappa$. A prior $\pi_0$ such that $P^\pi_0(\theta \leq \theta_0|y)$ is equal to the frequentist $p$-value, that is, the probability (under the data distribution $f$) to obtain a data outcome as or more extreme than the observed $y$ under the null hypothesis, is readily available for normal and binomial outcomes, corresponding to an improper uniform prior for the former, and a Beta$(0, 1)$ for the latter (Kopp-Schneider et al., 2019; Lecoutre, 2007). If such a prior $\pi_0$ is adopted for the analysis, then we have that the BD $d^\pi_\kappa$ induces a UMP test with TIE rate $\beta^\pi_0(\theta_0) = \tau$, which exactly coincides with posterior probability threshold.

The normal prior–normal likelihood model allows showing such a parallelism between frequentist and Bayesian test decisions (Berger, 1985). In particular, assume prior $\pi = N(\mu_\pi, \sigma_\pi)$ and data $y \sim N(\theta, \sigma/\sqrt{n})$. The BD can be obtained by
adopting a rejection threshold

$$\tau^\pi = 1 - \Phi \left( \frac{\sigma(\theta_0 - \mu_\pi)}{\sqrt{n \sigma^2_\pi}} + z_{1-\tau} \sqrt{1 + \frac{\sigma^2}{n \sigma^2_\pi}} \right), \quad (3)$$

under a vague prior $\pi_0 = N(\mu_\pi, \sigma_\pi)$, with $\sigma_\pi \to \infty$, and where $z_{1-\tau}$ denotes the $1 - \tau$ quantile of the standard normal distribution. The threshold $\tau^\pi$ coincides with the TIE rate of the test, and can be seen as induced by a revised cost ratio $\kappa^\pi$ which takes into account the prior probabilities assigned to each hypothesis by the informative prior distribution. It has therefore been suggested as a “rationally chosen” TIE rate for frequentist tests (Berger, 1985).

A compromise decision (CD) may therefore be reached either by tuning the frequentist test TIE rate level or, equivalently, the posterior probability threshold for rejection under $\pi_0$, so that it belongs to an interval which includes the standard TIE rate value $\beta^{\pi_0}(\theta_0)$, for example, 0.025 or 0.05, and the prior-induced TIE rate value $\beta^{\pi}(\theta_0)$ as extremes. A CD is here obtained by adopting a posterior decision threshold of the type

$$\tau^w = (1 - w)\tau + w\tau^\pi, \quad (4)$$

under $\pi_0$, where $w \in [0,1]$. The TIE rate is thus equal to $\beta^w(\theta_0) = \tau^w$. We denote the CD by $\alpha^{\pi_0}_w$, where $\kappa^w = \tau^w/(1 - \tau^w)$.

### 3.1 Dynamic borrowing

A major difficulty when borrowing historical information is the selection of the borrowing weight, which describes the degree of trust assigned to historical information. A reasonable solution would be to choose the weight based on a priori considerations on the commensurability of the historical and current study populations as measured by, for example, relevant covariates, but any other measure of discrepancy between the historical and current population may be used as well.

Here, we propose a dynamic approach in which the weight $w$ is defined according to the agreement between the posterior tail probabilities of the informative and vague prior distribution. Formally,

$$\hat{w} = 1 - |P^{\pi}(\theta > \theta_0 | y) - P^{\pi_0}(\theta > \theta_0 | y)|, \quad (5)$$

where $\pi_0^*$ is a prior specification with the same variance/informativeness as the informative prior, but with prior location agreeing with the observed data. Tuning the noninformative prior variance to match that of the informative prior one allows achieving full borrowing when no heterogeneity is observed between the informative prior and observed data location. Note that $\hat{w}$, in contrast with typical dynamic borrowing mechanisms, is specifically tailored to testing: If interest lies in estimation, a different discounting mechanism should be adopted to avoid potentially large bias. The estimated weight can also be seen as a measure of sensitivity of the Bayesian inference to the informative prior, as compared to a prior perfectly consistent with the observed data. This weight has indeed some resemblance with the “local sensitivity” measure introduced in O’Hagan (2010). A further connection of the CD approach with such a measure is discussed in the conclusions. Note, also, that since the weight choice, and therefore the threshold, is in this case dependent on the current data outcome, the exact TIE rate would have to be calculated via simulation. However, it is still possible to combine a dynamically estimated weight and a stricter control of the TIE rate, that is, we choose

$$\tau^\hat{w} = \min\{(1 - \hat{w})\tau + \hat{w}\tau^\pi, \tau^{\text{bound}}\},$$

where $\tau^{\text{bound}}$ is the maximum allowable TIE rate. This effectively constrains TIE rate also with a data-dependent weight: Indeed, if $z_{1-\hat{w}} \leq z_{1-\tau^{\text{bound}}}$, the TIE rate of the procedure will always be in $[0, \tau^{\text{bound}}]$.

Finally, note that another situation in which (4) would not directly provide the TIE rate value of the procedure is one in which historical information is considered random (not yet realized or unblinded). This situation is analytically investigated in Kopp-Schneider et al. (2023).
3.2 Connection with other robust borrowing approaches

The CD $d_{\pi_0}^\nu$ for any chosen fixed $w$ is also a TI-RBD satisfying (2) for an appropriately chosen $\eta$. To see this, note that both decisions are induced by a fixed nondegenerate prior distribution, and therefore they are a monotone function of the data. When $\eta = w = 1$ prior information is fully incorporated, the same data threshold is induced by both decisions, and the test has TIE rate $\beta_0(\Theta_0)$. For $w < 1$, data thresholds, and therefore test TIE rates, induced by the same values of $w$ and $\eta$ will however tend to differ, as exemplified in Section 4.

An additional well-known approach reaching a compromise between FDs and BDs is represented by a BD with respect to a prior arising from a mixture of the informative prior distribution $\pi$ and a vague or weakly informative prior distribution $\pi_r$, elicited with the aim of robustifying the analysis, that is, $\gamma = (1 - \epsilon)\pi_r + \epsilon\pi, \epsilon \in [0, 1]$. Such a prior has been referred to as the “robust mixture prior” in the literature (Berger & Berliner, 1986; Schmidli et al., 2014). Here, we denote the decision induced by such prior $d_\gamma^\nu$ as “robust mixture decision” (RMD). As the robust mixture prior is a fixed and nondegenerate prior, if the data outcome belongs to an exponential family distribution, we can again conclude that it induces UMP tests for any choice of $\pi_r$. Typically, however, as in the case of the CD and TI-RBD, the TIE rate of the test will differ for the RMD and TI-RBD for the same value of $\epsilon$ and $\eta$.

While identical test decisions can therefore be obtained under the CD and the robust mixture prior with generic robust component $\pi_r$ by appropriately tuning $w$, it is possible to analytically relate the robust mixture prior weight $\epsilon$ and $w$ only in a data-dependent way. Appendix A.2 shows how the loss ratio $\kappa$ should be modified to induce the same test decisions under the mixture and the FD.

Several additional robust borrowing approaches have been proposed in the literature, for example, power and commensurate prior formulations (Hobbs et al., 2012; Ibrahim & Chen, 2000). When the adopted prior is fixed and nondegenerate, any such approach would lead to a UMP test at some (inflated, if the prior favours $H_1$) TIE rate level.

3.3 Properties and sample size selection

Sensitivity analyses

Sensitivity analyses can be performed by exploiting the distinction between analysis and sampling prior (see, e.g., Calderazzo et al., 2022, and references therein). The sampling prior represents the prior of the data-generating process, that is, the prior under which parameters and consequently data samples are obtained. For evaluation of frequentist operating characteristics, sampling priors are point masses at different parameter values, typically $\theta_0$ as well as a $\theta$ value belonging to the alternative hypothesis support for power calculations. In the Bayesian context, sampling priors can belong to the same family as, for example, the informative prior distribution. When the sampling and analysis prior coincide, the integrated risk is minimized. Sensitivity analyses are then carried out by adopting different sampling prior choices when computing the integrated risk, while test decisions are taken according to the analysis prior which will be used to fit the data. Formally, the integrated risk reads

$$r(\pi_s, d_\kappa(\mathbf{y})) = \int_{-\infty}^{\theta_0} \beta(\theta)\pi_s(\theta)\,d\theta + \kappa \int_{\theta_0}^{\infty} [1 - \beta(\theta)]\pi_s(\theta)\,d\theta,$$

(6)

where $\pi_s$ denotes the sampling prior and $d_\kappa$ the test decision function under the 0-$\kappa$ loss induced by the analysis prior of choice. If $\pi_s = \pi$, the CD with weight $w$, $d_{\pi_0}^\nu$, is suboptimal for (6) unless $w = 1$, but more robust in case the “true” sampling prior differs from $\pi$. To perform sensitivity analyses, the decision can therefore be fixed to the CD $d_{\pi_0}^\nu$, while the sampling prior $\pi_s$ may be varied. One could in principle vary both location and variance of the sampling prior. In both cases, decisions under analysis priors that more closely match those of the sampling prior will lead to a lower integrated risk. Conflict in terms of location is generally of greater interest, as it reflects a wrong assumption concerning the mean effect. As a smaller $w$ will result in solutions closer to the frequentist ones, the integrated risk would be less influenced by a shift in location between the sampling prior and the prior obtained from historical information than if the BD $d_\kappa^\pi$ were adopted.
**Integrated risk inflation**

We can also compute the “relative saving loss” (RSL), that is, the ratio between the integrated risk inflation under the CD, and under the FD. Efron and Morris (1971) define it as

\[
\text{RSL}(\pi) = \frac{r(\pi, d_{\pi_0}^w(y)) - r(\pi, d_{\pi}^w(y))}{r(\pi, d_{\pi_0}^w(y)) - r(\pi, d_{\pi}^w(y))},
\]

where recall that \(\pi_0\) represents the prior for which the posterior probability of the alternative is equal to the p-value. Note that in this formulation the risk is computed under \(\pi_s = \pi\), and thus is minimized when fully incorporating prior information. The RSL is equal to 1 when all prior information is discarded, meaning that all advantages in terms of integrated risk minimization (provided that \(\pi_s = \pi\)) are lost, while it is equal to 0 when all prior information is incorporated. The RSL focuses on the global integrated risk, that is, it sums losses in both type I and type II error rate averaged with respect to the prior \(\pi\).

**Sample size selection**

Sample size selection can be naturally incorporated in the proposed framework, and depends on the operating characteristics of interest. Several possibilities have been explored in the literature in addition to the classical frequentist approach which targets power at a given alternative \(\theta_1\), and in particular a hybrid frequentist–Bayesian evaluation which fixes the TIE rate and targets a form of average power can be pursued. Different weightings have been proposed to perform such an averaging, with the underlying idea of leveraging the information contained in the prior distribution. A review of the approaches and terminology is provided in Kunzmann et al. (2020). Measures that incorporate uncertainty about the treatment effect often lead to larger required sample sizes as compared to the classical frequentist approach, particularly when small effects have nonnegligible probability. However, they also more realistically reflect available knowledge about the treatment effect as summarized in its prior distribution, and are thus recommended following Spiegelhalter and Freedman (1986). It is important to note that the CD allows selecting a given power function, which defines the TIE rate at \(\theta_0\), as well as the probability of rejection for any \(\theta\): sample size can then be selected to reflect any target power measure of choice. Among the available measures, if one wishes to limit the impact of the prior distribution on the average power measure, expected power can be used. This is formally computed as \(\int_{\theta_0}^{\infty} {\{1-\beta(\theta)\} \pi(\theta | \theta > \theta_0)} \, d\theta\) (see also Psioda & Ibrahim, 2019). One can elicit the sample size such that desired power (either under the informative prior, or across a set of “realistic” sampling priors) has reached a chosen level. Note that, irrespective of the chosen power definition, a larger power is always achieved when TIE rate is larger, as long as a UMP test is adopted.

**4 | EXAMPLE**

We illustrate the reviewed and proposed borrowing approaches focusing on a normal outcome. An analogous study for a binomial outcome is presented in the Supplementary Material. We consider the set of hypotheses \(H_0 : \theta \leq \theta_0\) versus \(H_1 : \theta > \theta_0\). We assume \(\pi = 0.025/0.975\), so that \(\tau = \pi/(1 + \pi) = 0.025\), which corresponds in turn to the TIE rate under \(\pi_0\).

Let \(\bar{y}\) denote the mean of independent and identically distributed observations from an \(N(\mu, \sigma^2 = 1)\) distribution. Let \(\pi\) denote the informative prior distribution, which is assumed to be \(N(\mu_x, \sigma_x^2)\), and arising from \(n_0 = 50\) historical observations with mean \(\mu_x = 0.25\) and standard deviation \(\sigma\), that is, \(\sigma_x = 1/\sqrt{50}\). Moreover, let \(\theta_0 = 0\). The FDs are in this case induced by a normal prior \(\pi_0\) with variance approaching infinity. Assuming \(\pi_0\) to be an \(N(0,100)\) provides in practice a good approximation for our purposes. Furthermore, we explore the impact of decisions under two mixture prior specifications: The RMDs-Unit and MDs-Vague. The RMDs-Unit are induced by \(\gamma^u = (1 - w)N(\mu_0, 1) + wN(\mu_0, \sigma_0^2)\), where the weakly informative component is taken to be a unit-information prior (Kass & Raftery, 1995), that is, its variance is equal to that of a single observation. For the MDs-vague, we consider decisions under \(\gamma^v = (1 - w)N(\mu_0, 100) + wN(\mu_0, \sigma_0^2)\) instead. Note that the MD-Vague prior is not a robust prior choice and is not generally recommended (see O’Hagan, 2010, for a comprehensive overview of robustness of \(\epsilon\)-contamination priors), and is only included for explanatory purposes with the aim of fully compromising between the FD and BD by tuning its weight. Let the TI-RBDs be induced by \(\nu = (1 - w)\mu_0 + wN(\mu_x, \sigma_x^2)\), where \(\mu_0 = \delta_{\theta_0}\), and the CDs by \(\pi_0\) with posterior probability threshold for rejection equal to \(\tau^w = (1 - w)\tau + w\tau_x\). We also introduce as a comparator a dynamic borrowing approach based on the power
### Table 1
List of implemented approaches.

| Decision  | Prior                      | Threshold            | TIE rate               |
|-----------|----------------------------|----------------------|------------------------|
| FD        | $\pi_0 = N(0, 100)$        | $\tau = 0.025$       | $\approx \tau$         |
| BD        | $\pi = N(0.25, \frac{1}{\sqrt{50}})$ | $\tau = 0.025$ | $\tau^\pi$             |
| CD        | $\pi_0 = N(0, 100)$        | $(1 - w)\tau + w\tau^\pi$ | $\approx (1 - w)\tau + w\tau^\pi$ |
| EBPD      | $\pi = N(0.25, \frac{1}{\sqrt{50}})$ | $\tau = 0.025$ | to be eval.             |
| CD-Adapt  | $\pi_0 = N(0, 100)$        | $1 - \hat{w}\tau + \hat{w}\tau^\pi$ | to be eval. ($\leq 0.15$) |
| RMD-Unit  | $\gamma^u = (1 - w)N(0.25,1) + wN(0.25, \frac{1}{\sqrt{50}})$ | $\tau = 0.025$ | to be eval.             |
| MD-Vague  | $\gamma^v = (1 - w)N(0.25,100) + wN(0.25, \frac{1}{\sqrt{50}})$ | $\tau = 0.025$ | to be eval.             |
| TI-RBD    | $\nu = (1 - w)\mu_0 + wN(0.25, \frac{1}{\sqrt{50}})$ | $\tau = 0.025$ | to be eval.             |

Prior specification, the Empirical Bayes (EB) power prior (Gravestock & Held, 2017), which induces the empirical Bayes power prior decision (EBPD). The EB power prior can be interpreted as a posterior arising from an historical study which collected $n_0$ observations, of which $a_0n_0$ are accounted for in the analysis, where the parameter $a_0 \in [0, 1]$ is selected by maximization of the marginal likelihood $m(\bar{y}) = \int f(\bar{y}|n, \theta)\pi(\theta|a_0, \mu, n_0) d\theta$, where $\pi(\theta|a_0, \mu, n_0)$ denotes the power prior with parameter $a_0$, and $\mu$ and $n_0$ are the historical outcome and sample size, respectively (Gravestock & Held, 2017). When $a_0 = 1$, the EBPD coincides with the BD. Note that the EB commensurate prior (Hobbs et al., 2012) for normal outcomes coincides with the EB power prior (see Appendix C.1 of Wiesenfarth & Calderazzo, 2020), and therefore all results for the EB power prior directly apply also to the EB commensurate prior in this setup. To avoid prior domination, we impose $a_0 \leq \min(1, \frac{\sigma^2}{\sigma^2/n})$, so that historical information equivalent to no more than $n$ samples is incorporated, and refer to this as the “bounded” approach. Finally, in the CD-Adapt approach, the decision is undertaken under $\pi_0$, with threshold $\tau^\pi = \min(0.15, (1 - \hat{w})\tau + \hat{w}\tau^\pi)$, with $\hat{w}$ as in (5), and the constraint on $\hat{w}$ ensures that TIE rate is always below 0.15. Note that, to facilitate comparisons, the weight assigned to historical information is kept constant and equal to $w$ across all robust approaches with the exclusion of “CD-Adapt” and “EBPD.” Note, however, that the mixture prior automatically adapts such prior weight in the resulting posterior mixture according to the Bayes Factor. Therefore, the impact of the informative component is additionally adapted based on the observed data. A summary of the implemented prior assumptions and decisions is provided in Tables 1.

Figure 2 compares relative saving loss $RSL(\pi)$ in (7), TIE rate, and expected power of the different test decisions, that is, the CD, the RMDs, the TI-RBD, EBPD, CD-Adapt, and, as reference, the BD and the FD. The weight for CD, RMDs, and TI-RBD is varied from 0, corresponding to no borrowing, to 1, corresponding to full borrowing. The results for CD-Adapt and EBPD are averaged across data outcomes (and, thus, weights $\hat{w}$). We first notice that the TI-RBD can induce TIE rate and expected power levels well below those of the FD, even for relatively large weight assigned to the informative prior component. This, in turn, induces a sharp increase in expected power and decrease in $RSL$ only when $w$ is larger than approximately 0.9. Note the TI-RBD is not in itself a proposed analysis strategy, and is just added for completeness following (2). The RMD-Unit and the CD show a strong similarity, particularly for $n = 100$. Note that the RMD-Unit does not induce exactly the same results as the FD when $w = 1$ due to the fact that the robust component is taken to be $N(\mu, 1)$. It is of interest to note, however, that the closeness of the CD and RMD-Unit depends quite heavily on the choice of the robust component. As can be observed in the graphs, under the MD-Vague, historical information is effectively only discarded for very small values of $w$. This is a known phenomenon and related to the computation of the posterior weights: a very dispersed component induces a small marginal likelihood and thus causes the informative component to be favored, irrespective of its commensurability to the data. For the same reason, vague prior components are not recommended when performing point-null hypothesis testing (Bartlett, 1957; Kass & Raftery, 1995; O’Hagan, 2010). It is worth noting, however, that similarity between the RMD-Unit and the CD approach is also related to the location and variance of the informative component itself. The EBPD lead to modest borrowing when $n = 20$, as a consequence of the prior domination avoidance bound, while it is equivalent to the BD for $n = 100$.

By construction, the TIE rate of the CD varies linearly with the weight $w$. The RMD-Unit follows this pattern very closely in this example. An approximately linear relationship between the weight assigned to prior information and the change in the operating characteristic on which a compromise is sought seems a desirable feature which may guarantee an easier interpretation of the weight itself. We believe that such a linearity can indeed be a useful measure to investigate for any borrowing mechanism, particularly when analytic relationships are not available. The CD-Adapt approach ultimately
leads to values identical to those of the CD approach with \( w = 0.75 \) for \( n = 20 \), and approximately \( w = 0.95 \) for \( n = 100 \). While this result is in general not guaranteed, it shows that in this case it avoids the need for a prespecification of \( w \) itself. Note that, for \( n = 20 \), the results are driven by enforcement of the TIE rate bound of 0.15.

The left panels of Figure 3 show how the required minimum sample size to achieve 80% expected power would be affected by the choice of \( w \) under the different borrowing approaches. Note that the maximum sample size has been truncated at 250. The minimum sample size decreases approximately linearly from the sample size required under no borrowing (\( n = 214 \)) to that required under full borrowing (\( n = 91 \)) under the RMD-Unit and the CD. Such a decrease mimics the increase in TIE rate, while expected power is generally controlled and above 0.8 by construction; the only exception is observed for the TI-RBD as sample sizes above 250 would be required under most \( w \) values to achieve the expected power target. Again, the CD-Adapt approach corresponds to a weight of approximately 0.95 under CD. Finally, MD-Vague induces again a behavior closer to the BD one for most weight choices.

To demonstrate how sensitivity analyses in the context of sample size selection can be conducted, we fix \( w = 0.5 \) for the TI-RBD, (R)MDs, and CD approaches, and again a target expected power equal to 0.8. We assume the sampling prior to be \( N(\mu_{\pi_i}, \sigma_{\pi_i}) \), where \( \sigma_{\pi_i} = \sigma_{\pi} \), and varying sampling prior mean. The top right panel of Figure 3 shows how the minimum required sample size is affected by the location of the sampling prior. Operating characteristics under the BD and FD are superimposed for comparison. Recall that the FD always controls TIE rate at 0.025, while the BD TIE rate varies according to the sample size, which is in turn determined by the expected power target. Recall also the we truncate the sample size at 250, so expected power would be below the target 0.8 when a larger sample size would be needed to achieve it. We
observe that historical information is beneficial in terms of expected power and thus, in turn, sample size savings can be achieved over the whole range of sampling prior means when prior do incorporate such information; however, a decrease in required sample size can induce a significant inflation in TIE rate. The CD behaves again closely to the RMD-Unit in terms of expected power and sample size requirements, but its TIE rate is fully determined by the choice of \( w \), and implies that, under any sample size and data-generating mechanism, the inflation in TIE rate would be 50% of the increase in TIE rate induced by the BD (for the same sample size). Finally, the CD-Adapt approach achieves sample size gains close to that of the informative and MD-Vague prior distribution while maintaining the TIE rate, as planned, below 0.15.

Sensitivity analyses for the integrated risk, for sample size equal to 20 and 100, are shown in Figure 4. For both sample sizes, the informative prior and MD-Vague again achieve very close results, with the integrated risk being minimized when the sampling prior mean coincides with the informative prior mean, but with potential significant inflation otherwise. The FD has the lowest maximum integrated risk, although it loses some integrated risk gains as compared to decisions incorporating historical information, when the sampling and informative prior are consistent, as expected.

5 | CASE STUDY

We further investigate the behavior of the CD approach in a case study reported in Travis et al. (2023) (see also Brunner et al., 2020; Food & Drug Administration, 2018). The investigation concerns pediatric patients between 5 and 17 years
old suffering from systemic lupus erythematosus (SLE). Log-odds ratios for the binary primary endpoint, SLE Responder Index (SRI4) response rate at Week 52 after treatment with Belimumab as compared to Placebo, are available for the pediatric patients, but the study lacks power due to recruitment feasibility limitations. Borrowing of information from adult studies was then considered to aid reaching a definite conclusion. Under approximate normality, adult information is robustly incorporated through the adoption of a robust meta-analytic predictive mixture prior (RMD-Unit) of the form \((1 - 0.55)N(0, \sqrt{8.27}) + 0.55N(0.51, \sqrt{0.016})\), where the weight \(w = 0.55\) is chosen via a tipping point analysis and considered to convey an acceptable level of trust in adult information. The BD is taken under the informative component of the mixture \(N(0.51, \sqrt{0.016})\). The pediatric log-odds ratio distribution is approximated as \(\bar{y} \sim N(0.4, \sqrt{0.19})\). Table 2 shows the results of the analysis under different dynamic borrowing mechanisms. For the CD-Adapt and the EBPD both bounded and unbounded results are reported. All dynamic borrowing approaches detect substantial homogeneity between adult and pediatric data, leading in almost all cases to full borrowing when no further constraints are applied. Figure S5 provides additional insight into the adaptive weight choice of the CD-Adapt in the unbounded situation, and in particular its sensitivity with respect to different potential data outcomes. Note that the RMD-Unit approach has been calibrated to have the smallest prior weight leading to rejection. Note also that, for CD-Adapt, the threshold for rejection is varied, rather than the posterior probability of the null. In essence, the estimated trust on adult information is so strong that the cost of a TIE can be modified to be approximately 0. This shows an intrinsic limitation of any data-dependent estimation of heterogeneity: it still requires some trust to be placed on historical information. Indeed, if no trust is placed on adult information, the unbounded CD-Adapt leads to a TIE rate of approximately 0.5 (Figure S4).
TABLE 2  Hypothesis testing for the case study described in Section 5 on under different approaches: posterior probabilities of the null hypothesis, rejection threshold, weight assigned to adult information and test decision.

|               | Test decision | $P(H_0|y)$ | Adaptive (post.) weight | Threshold |
|---------------|--------------|-----------|-------------------------|-----------|
| FD            | 0.000        | 0.182     | 0.000                   | 0.025     |
| RMD-Unit ($w = 0.55$) | 1.000        | 0.021     | 0.884                   | 0.025     |
| BD            | 1.000        | 0.000     | 1.000                   | 0.025     |
| Bounded       |              |           |                         |           |
| CD-Adapt     | 0.000        | 0.182     | 0.128                   | 0.150     |
| EBPD          | 0.000        | 0.072     | 0.083                   | 0.025     |
| Unbounded     |              |           |                         |           |
| CD-Adapt     | 1.000        | 0.182     | 1.000                   | 1.000     |
| EBPD          | 1.000        | 0.000     | 1.000                   | 0.025     |

Abbreviations: BD, Bayes decision; CD, compromise decision; EBPD, xxx; FD, frequentist decision; RMD, robust mixture decision.

rate is capped at 0.15, leading to $\hat{w} = 0.13$. Note that this corresponds to the same TIE rate (and therefore expected power) of the RMD-Unit with prior weight approximately equal to 0.45 (see Figure 5). Note, also, that TIE rate of RMD-Unit increases in a nonlinear way with the prior weight in this example, and is in general more conservative than the CD.

6 | CONCLUSIONS

In this work, we have approached the problem of building a principled frequentist–Bayesian CD to testing under a $0-\kappa$ loss for location parameters of exponential family distributions in one-arm studies. In the CD, the role of historical information in the TIE rate inflation is made explicit through the modification of the cost ratio between type I and type II errors. In other words, we envisage a procedure in which an initial cost ratio $\kappa$ unrelated to prior information is chosen; $\kappa$ can be either a standard value such as 0.025/0.975, or rather related to test situation at hand (e.g., disease severity or rarity). The second step is the choice or automatic selection of the weight $w$, which specifies the trust in historical information and allows a gradual tuning of the initial $\kappa$. Finally, if desired and in case of a dynamic choice of $w$, a cap on the maximum TIE rate $\tau_{\text{bound}}$ can be applied. The latter serves as an additional layer of robustness and guarantees against excessive borrowing in a data-independent way.

Note that the approach does not require a Bayesian analysis strategy in itself. However, it does require a certain amount of trust in historical information. Under a fixed borrowing mechanism, the CD has been shown to be optimal in terms of minimization of the integrated risk, subjected to TIE rate constraint. While this analytical property may be difficult to generalize to more complex situations, we believe that a compromise explicitly targeting the inevitable TIE rate inflation may carry some advantages, particularly when complex dynamic borrowing mechanisms do not allow its direct quantification. If the Bayesian paradigm is fully embraced, no such compromise is necessary. However, full commensurability is a strong assumption and sensitivity analyses are often necessary to evaluate benefits and losses associated with borrowing of historical information under various heterogeneity scenarios (Viele et al., 2014). Frequentist error rates can still serve as well-known and easily interpretable measures to perform such evaluations as they do not make any assumption about the reliability of prior information. A situation where such an assessment is particularly important is when a very large historical dataset is available for borrowing, as in the case study considered. If fixed borrowing is adopted, TIE rate can quickly reach very high levels, which is typically undesirable (see Kopp-Schneider et al., 2023, for more extensive discussion of this point).

The amount of TIE rate inflation is tuned by the choice of a specific parameter, in our case the weight $w$, which represents the allowed proportion of TIE rate inflation, as compared to an analysis fully incorporating historical information. We have outlined possibilities for the choice of such a weight, and proposed a novel adaptive approach for its elicitation which explicitly focuses on testing, and in particular on the impact of historical information on tail probabilities. It is here important to stress that neither the CD-Adapt nor the CD leads to the choice of a specific prior distribution, but rather to a set of test decisions that incorporate prior information to some, possibly adaptive, degree. When the weight is fixed, equivalence with the results of fixed prior distributions is achieved. When the weight is adapted based on the observed data, the procedure more closely resembles that of “Empirical Bayes” approaches: if monotonicity of the test decision in
the data is preserved, frequentist test error rates are equivalent to an approach with a specific fixed weight choice, but this is not in general guaranteed.

Estimation has not been investigated in the current work. Our view is that, when different operating characteristic targets are of interest, the same prior may induce satisfactory decisions (i.e., having small observed conflict) for one target, but not for the other. In other words, unless prior information is fully trusted (or discounted based only on a priori considerations), an approach which focuses on decisions and their impact can be better tuned to the specific aims. For example, if the prior and the observed data are both quite far into the alternative but far apart from each other, estimation bias may be an issue, but there would be no controversy with respect to the test decision. Our approach can be combined with a CD for estimation, based, for example, on mean square error (MSE) considerations (see Efron & Morris, 1971).

The effective sample size (ESS) is often advocated and requested by regulators as a measure of prior informativeness (Travis et al., 2023). The traditional ESS concept, which equates prior informativeness to a certain number of samples,
not applicable to our approach. However, we believe that also such a measure is best evaluated based on the impact of borrowed information on the target measure of choice. We have introduced in previous work (Wiesenfarth & Calderazzo, 2020) the effective current sample size (ECSS) measure to equate the impact of an informative prior to a certain number of samples consistent with the current observed data in the context of estimation. Currently, we are extending this approach to the framework of testing so that it would be directly applicable to the CD and CD-Adapt.

We have compared the operating characteristics obtained when assigning the same prior weight to historical information in the static CD and the mixture prior approaches. We observe a general lack of robustness of the mixture prior approach when the baseline component is chosen to be very flat, which leads to a rather sharp increase in TIE rates also at small prior weight values. Lack of robustness of such a prior is not unexpected and has been discussed before (O'Hagan, 2010). The CD approach, on the other hand, leads by construction to a linear increase. More generally, linearity between the weight assigned to historical information and the operating characteristic on which compromise is sought seems to us a desirable property of any borrowing mechanism, as it implies a good interpretability of the weight itself. From the Bayesian perspective, a related concept is that of “local sensitivity” (O'Hagan, 2010): in this case, the speed of change in some posterior inferences of interest is assessed, conditional on the observed data, and as the borrowing parameter goes to 1. We thus believe that such assessments could be of broader interest and a potential topic for further research.

We have provided tools for sensitivity analyses and sample size selection. We have primarily focused on reaching a target expected power, defined as average power with respect to the sampling distribution truncated and $\theta_0$. This is not the only option. The lower truncation boundary $\theta_0$ can be replaced, for example, by a relevance threshold, or integration can be performed over the whole parameter range (see Kunzmann et al., 2020, for a review).

We have considered normal and binomial outcomes in one-arm studies. The approach is more generally applicable to any situation in which approximate normality can be assumed, including common estimators for time-to-event outcomes, such as log-hazard ratios. Finally, extension to two-arm situations poses no difficulties if a single prior is elicited on the difference between the treatment and control mean: When normal outcomes are considered, reduction to a one-arm design is straightforward and the method can be directly applied. A situation requiring further study is when separate priors are elicited for the treatment and the control arm, and will be the focus of future research.

**ACKNOWLEDGMENTS**
The authors wish to thank the Associate Editor and two anonymous reviewers for their valuable comments and suggestions.

Open access funding enabled and organized by Projekt DEAL.

**SOFTWARE**
R code for reproduction of all results is available at https://github.com/BDTTrialDesigns/RobustBorrowingTest.

**CONFLICT OF INTEREST STATEMENT**
The authors declare no conflicts of interest.

**DATA AVAILABILITY STATEMENT**
The data that support the findings of this study are available in the supplementary material of this article.

**OPEN RESEARCH BADGES**
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

**ORCID**
Silvia Calderazzo https://orcid.org/0000-0003-3106-9089
Annette Kopp-Schneider https://orcid.org/0000-0002-1810-0267

**REFERENCES**
Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika, 44*(3-4), 533–534.
Bayram, S., & Gezici, S. (2011). On the restricted Neyman-Pearson approach for composite hypothesis-testing in presence of prior distribution uncertainty. *IEEE Transactions on Signal Processing*, 59(10), 5056–5065.

Berger, J., & Berliner, L. M. (1986). Robust Bayes and empirical Bayes analysis with $\epsilon$-contaminated priors. *Annals of Statistics*, 14(2), 461–486.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer Series in Statistics, Springer.

Brunner, H. I., Abud-Mendoza, C., Viola, D. O., Penades, I. C., Levy, D., Anton, J., Calderon, J. E., Chasnyk, V. G., Fernandiz, M. A., Keltsev, V., Gastanaga, M. E. P., Shishov, M., Boteanu, A. L., Henrickson, M., Bass, D., Clark, K., Hammer, A., Ji, B. N., Nino, A., ... Ruperto, N. (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.

Calderazzo, S., Wiesenfarth, M., & Kopp-Schneider, A. (2022). A decision-theoretic approach to Bayesian clinical trial design and evaluation of robustness to prior-data conflict. *Biostatistics*, 23(1), 328–344.

Efron, B., & Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators—Part I: The Bayes case. *Journal of the American Statistical Association*, 66(336), 807–815.

Food and Drug Administration. (2010). *Guidance for Industry and FDA staff: Guidance for the use of Bayesian statistics in Medical Device Clinical Trials*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials

Food and Drug Administration. (2018). FDA review of Benlysta (belimumab) for intravenous infusion in children 5 to 17 years of age with SLE. https://www.fda.gov/media/127912/download

Gravestock, I., & Held, L. (2017). Adaptive power priors with empirical Bayes for clinical trials. *Pharmaceutical Statistics*, 16(5), 349–360.

Hobbs, B. P., Sargent, D. J., & Carlin, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7(3), 639–674.

Hodges, J. L., & Lehmann, E. L. (1952). The use of previous experience in reaching statistical decisions. *Annals of Mathematical Statistics*, 23(3), 396–407.

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1), 46–60.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kopp-Schneider, A., Calderazzo, S., & Wiesenfarth, M. (2020). Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. *Biometrical Journal*, 62(2), 361–374.

Kopp-Schneider, A., Wiesenfarth, M., Held, L., & Calderazzo, S. (2023). Simulating and reporting frequentist operating characteristics of clinical trials that borrow external information. *Pharmaceutical Statistics*, 1–16, https://doi.org/10.1002/pst.2334

Kopp-Schneider, A., Wiesenfarth, M., Witt, R., Edelman, D., Witt, O., & Abel, U. (2019). Monitoring futility and efficacy in phase II trials with Bayesian posterior distributions—A calibration approach. *Biometrical Journal*, 61(3), 488–502.

Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Ruibach, K., & Wason, J. M. S. (2021). A review of Bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, 75(4), 424–432.

Lecoutre, B. (2007). The Bayesian approach to experimental data analysis. In C. Rao, J. Miller, & D. Rao, (Eds.), *Epidemiology and medical statistics* (Vol. 27, pp. 775–812). Handbook of Statistics. Elsevier.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). Wiley Series in Probability and Statistics. Wiley.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2), 380–391.

Nikolakopoulos, S., Tweel, I., & Roes, K. C. B. (2018). Dynamic borrowing through empirical power priors that control type I error. *Biometrics*, 74(3), 874–880.

O’Hagan, A. (2010). *Kendall’s advanced theory of statistic 2B*. Wiley.

Psioda, M. A., & Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20(3), 400–415.

Robert, C. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Schmich, H., Gsteiger, S., Roychoudhury, S., O’Hagan, A., Spiegelhalter, D., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4), 1023–1032.

Spiegelhalter, D. J., & Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1), 1–13.

Travis, J., Rothmann, M., & Thomson, A. (2023). Perspectives on informative Bayesian methods in pediatrics. *Journal of Biopharmaceutical Statistics*, 33(6), 830–843. PMID: 36710384.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., & Thompsonl, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Biostatistics*, 15(1), 46–60.

Whitt, W. (1979). A note on the influence of the sample on the posterior distribution. *Journal of the American Statistical Association*, 74(366), 424–426.

Wiesenfarth, M., & Calderazzo, S. (2020). Quantification of prior impact in terms of effective current sample size. *Biometrics*, 76(1), 326–336.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.
APPENDIX A

A.1 Restricted Bayes solution with constraint on maximum type I error (TIE)

The following theorem and proof follows Hodges and Lehmann (1952), with minor adaptations.

**Theorem A.1.** Let \( \nu = \eta \pi + (1 - \eta) \mu_0 \) and \( d' \) the Bayes solution with respect to \( \nu \). If \( d' \) is such that \( \int_{\Theta} R(\theta, d') \mu_0(\theta) d\theta = R(\theta_0, d') \), then it also minimizes

\[
\eta \int_{\Theta} R(\theta, d) \pi(\theta) d\theta + (1 - \eta) R(\theta_0, d).
\]

**Proof.** Let \( d \) be any decision, then

\[
\eta \int_{\Theta} R(\theta, d) \pi(\theta) d\theta + (1 - \eta) R(\theta_0, d) \geq \eta \int_{\Theta} R(\theta, d') \pi(\theta) d\theta + (1 - \eta) \int_{\Theta} R(\theta, d) \mu_0(\theta) d\theta
\]

\[
\geq \eta \int_{\Theta} R(\theta, d') \pi(\theta) d\theta + (1 - \eta) \int_{\Theta} R(\theta, d') \mu_0(\theta) d\theta = \eta \int_{\Theta} R(\theta, d') \pi(\theta) d\theta + (1 - \eta) R(\theta_0, d').
\]

It also follows that \( \mu_0 \) is the point–mass density at \( \theta_0 \).

A.2 Mixture prior test decision and cost elicitation

Let \( \gamma = (1 - \epsilon) \pi' + \epsilon \pi, \epsilon \in [0, 1] \) be a mixture prior, where \( \pi \) represents an informative prior distribution and \( \pi' \) be a robust component. For simplicity, we assume \( \pi' = \pi_0 \), but analogous calculations can be carried out to obtain the updated costs under a generic \( \pi' \). Denote \( \frac{\pi_0}{\pi} = p_{\pi_0}(\theta \leq \theta_0) \), \( \frac{\pi_1}{\pi} = p_{\pi_1}(\theta > \theta_0) \), and analogously for \( \pi \). Let also \( p_{\pi_0}(y|\Theta_1) = \int_{\theta_0}^{\infty} f(y|\theta) \pi_0(\theta|\theta \in (\theta_0, \infty)) d\theta, p_{\pi_0}(y|\Theta_0) = \int_{-\infty}^{\theta_0} f(y|\theta) \pi_0(\theta|\theta \in (-\infty, \theta_0]) d\theta \), and analogously for \( \pi \). Moreover, let a new cost ratio be

\[
k' = \frac{(1 - \epsilon) + \epsilon \frac{p_{\pi_0}(y|\Theta_1)}{p_{\pi_0}(y|\Theta_1)}}{(1 - \epsilon) + \epsilon \frac{p_{\pi_1}(y|\Theta_1)}{p_{\pi_1}(y|\Theta_1)}}.
\]

Under \( \pi_0 \), \( H_0 \) is kept if

\[
\frac{p_{\pi_0}(\Theta_0|y)}{p_{\pi_0}(\Theta_1|y)} \geq k',
\]

which implies

\[
k \leq \frac{(1 - \epsilon) + \epsilon \frac{p_{\pi_0}(y|\Theta_1)}{p_{\pi_0}(y|\Theta_1)} \frac{p_{\pi_0}(y|\Theta_0)}{p_{\pi_0}(y|\Theta_0)} \frac{p_{\pi_0}(y|\Theta_1)}{p_{\pi_0}(y|\Theta_1)}}{(1 - \epsilon) + \epsilon \frac{p_{\pi_1}(y|\Theta_1)}{p_{\pi_1}(y|\Theta_1)} \frac{p_{\pi_0}(y|\Theta_0)}{p_{\pi_0}(y|\Theta_0)} \frac{p_{\pi_1}(y|\Theta_1)}{p_{\pi_1}(y|\Theta_1)}} = \frac{p_{\gamma}(\Theta_0|y)}{p_{\gamma}(\Theta_1|y)}.
\]

Therefore, adopting the cost ratio \( k' \) under \( \pi_0 \) would lead to the same test decision as the test performed under the robust mixture prior \( \gamma \) with cost ratio \( k \).