



Eliciting prior information from clinical trials via calibrated Bayes factor

Roberto Macrì Demartino ^{a, *}, Leonardo Egidi ^a, Nicola Torelli ^a, Ioannis Ntzoufras ^b

^a Department of Economics, Business, Mathematics and Statistics, University of Trieste, Via A. Valerio 4/1, Trieste, 34127, Friuli-Venezia Giulia, Italy

^b Department of Statistics, Athens University of Economics and Business, Trias 2, Athens, 11362, Attica, Greece

ARTICLE INFO

Dataset link: <https://github.com/RoMaD-96/CBFpp>

Keywords:

Dynamic borrowing
Historical data
Model selection
Power prior
Prior elicitation

ABSTRACT

In the Bayesian framework power prior distributions are increasingly adopted in clinical trials and similar studies to incorporate external and past information, typically to inform the parameter associated with a treatment effect. Their use is particularly effective in scenarios with small sample sizes and where robust prior information is available. A crucial component of this methodology is represented by its weight parameter, which controls the volume of historical information incorporated into the current analysis. Although this parameter can be modeled as either fixed or random, eliciting its prior distribution via a full Bayesian approach remains challenging. In general, this parameter should be carefully selected to accurately reflect the available historical information without dominating the posterior inferential conclusions. A novel simulation-based calibrated Bayes factor procedure is proposed to elicit the prior distribution of the weight parameter, allowing it to be updated according to the strength of the evidence in the data. The goal is to facilitate the integration of historical data when there is agreement with current information and to limit it when discrepancies arise in terms, for instance, of prior-data conflicts. The performance of the proposed method is tested through simulation studies and applied to real data from clinical trials.

1. Introduction

In recent years, biostatistical applications are commonly characterized by insufficient sample sizes, which are crucial for accurate parameter estimation. Meanwhile, in the clinical framework, a large amount of historical or related data is often available. This has led to a growing interest in the use of historical data to enhance the design and analysis of new studies, particularly in clinical trials where recruiting patients can be ethically challenging. Notably, the sequential nature of information updating has made Bayesian approaches with informative priors particularly popular in this context (Chen and Ibrahim, 2000; Spiegelhalter et al., 2004; Neuenschwander et al., 2010; Hobbs et al., 2011, 2012; Viele et al., 2014; Schmidli et al., 2014; Yang et al., 2023; Alt et al., 2024a, among others). These methods facilitate the incorporation of historical data into the analysis by eliciting informative priors on the model parameters, thereby improving the robustness and efficiency of statistical inference. However, the elicitation of informative priors is widely recognized as a complex process because of the challenges in quantifying and synthesizing prior information into suitable prior distributions. For an in-depth and comprehensive analysis of the topic, see Spiegelhalter et al. (2004) and Neuenschwander and Schmidli (2020).

* Corresponding author.

E-mail address: roberto.macridemartino@deams.units.it (R. Macrì Demartino).

<https://doi.org/10.1016/j.csda.2025.108180>

Received 29 June 2024; Received in revised form 27 March 2025; Accepted 27 March 2025

Available online 31 March 2025

0167-9473/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Consequently, there is a pressing need to develop more efficient methods for synthesizing and quantifying prior information (Ibrahim et al., 2015). Specifically, there is a growing concern about the adaptive incorporation of historical data, especially in the presence of data heterogeneity and rapid changes in initial trial conditions (Ollier et al., 2020).

In this framework, the power prior (Chen and Ibrahim, 2000) is a popular method that allows historical data to influence the prior distribution in a flexible and controlled way. Ibrahim et al. (2003) provided a formal justification for power priors, demonstrating their effectiveness as a valuable class of informative priors. This effectiveness stems from their ability to minimize a convex sum of Kullback-Leibler (KL) divergences between two distinct posterior densities, in which one does not include any historical data, whereas the other fully integrates this information into the current analysis. Furthermore, De Santis (2006) proposed additional operational justifications, linking them to the so-called geometric priors. Notably, power priors have been employed across a broad spectrum of models, including generalized linear models (GLMs), generalized linear mixed models (GLMMs), and survival models (Chen et al., 2000; Chen and Ibrahim, 2006; Ibrahim et al., 2015). At its core, the idea is to raise the likelihood of the historical data to a weight parameter δ , usually defined between zero and one. This scalar parameter plays a crucial role in the power prior methodology as it determines the degree to which historical data influence the prior distribution. Specifically, when δ is set to zero, the power prior completely discounts historical information; conversely, setting δ to one fully integrate historical information into the prior.

As is intuitive, the role that the weight parameter δ plays in the final inferential conclusions is not negligible. Thus, several strategies have been developed to specify the weight parameter δ , treating it either as a fixed or a random quantity. If fixed in advance, the weight parameter δ can be set based on prior knowledge or through sensitivity analysis, considering specific criteria for borrowing information based on the prior-data conflict (Evans and Moshonov, 2006; Egidi et al., 2022). If treated as random, an initial prior distribution – typically a Beta distribution – is assigned to δ , and the use of the joint normalized power prior (Duan et al., 2006; Neuenschwander et al., 2009) is recommended. Notably, Nikolakopoulos et al. (2018) introduced a method to estimate this parameter using predictive distributions, aiming to control type I error by calibrating to the degree of similarity between current and historical data. Furthermore, Liu (2018) recommended setting δ through a dynamic p -value, assessing the compatibility of current and historical data based on the test-then-pool methodology. Gravestock and Held (2017, 2019) proposed an empirical Bayes-type approach to estimate the weight parameter by maximizing the marginal likelihood. Bennett et al. (2021) suggested two novel approaches for binary data, focusing on equivalence probability and tail area probabilities. Mariani et al. (2024) explored the use of the Hellinger distance to compare the posterior distributions of the control parameter from current and historical data, respectively. These techniques provide valuable insights on determining a specific fixed value for δ . However, in a fully Bayesian context, only Shen et al. (2023) has developed methods to specify the shape parameters of a Beta initial prior for δ , using two minimization criteria: Kullback-Leibler (KL) divergence and mean squared error (MSE). Therefore, in a fully Bayesian framework, eliciting the initial distribution of the weight parameter controlling the amount of historical information remains a challenging and underexplored area.

This paper aims to propose a novel Bayesian algorithmic approach for eliciting the initial prior distribution for δ in a somehow optimal way. This involves the use of a simulation-based calibrated Bayes factor, employing hypothetical replications generated from the posterior predictive distribution, to compare competing prior distributions for δ . Following the approach in Shen et al. (2023), a well-balanced prior should promote the integration of historical data when there is agreement with the current information and limit this integration when discrepancies arise between the two datasets.

The paper is organized as follows. Section 2 provides a review of the power prior methodology, discussing the use of the weight parameter as both a fixed and a random quantity. Furthermore, Section 3 illustrates the proposed calibrated Bayes factor to elicit a well-balanced initial prior distribution for δ . Sections 4 and 5 explore the proposed methodology through simulation studies and real data analysis from two clinical trials, E2696 and E1694, which investigated the effectiveness of interferon in melanoma treatment (Kirkwood et al., 2001a,b). Finally, Section 6 provides concluding remarks about the discussed method, highlighting its strengths and limitations, and provides insights into future developments.

2. Power priors

Power priors have become increasingly popular for the development of informative priors, especially in clinical trials, where past information is often available. These priors effectively integrate knowledge from historical data into the specification of informative priors. Let consider θ as the vector parameter of interest in the model, and let y_0 represent the historical data with their corresponding likelihood function denoted by $L(\theta | y_0)$. The basic formulation of the power prior (Chen and Ibrahim, 2000) is given by

$$\pi(\theta | y_0, \delta) \propto L(\theta | y_0)^\delta \pi_0(\theta), \quad (1)$$

where $\delta \in [0, 1]$ is the scalar weight parameter, and $\pi_0(\theta)$ is the initial – often non-informative – prior for θ . This model can be seen as a generalization of the classical Bayesian updating of $\pi_0(\theta)$. Furthermore, as noted by Ibrahim et al. (2015), the parameter δ plays a crucial role in determining the shape of the prior distribution for θ . Furthermore, updating the power prior in (1) with the current data likelihood $L(\theta | y)$ yields the following posterior distribution of θ

$$\pi(\theta | y, y_0, \delta) \propto L(\theta | y) L(\theta | y_0)^\delta \pi_0(\theta).$$

The formulation in (1) is conditional on the weight parameter and requires a predetermined and known value for δ . Therefore, to ensure an appropriate level of historical information borrowing while managing prior-data conflict, sensitivity analysis is recommended. In addition to the dynamic methods outlined in Section 1, Ibrahim et al. (2015) also proposed several other statis-

tical methods, including the Penalized Likelihood-type Criterion (PLC), Marginal Likelihood Criterion (MLC), Deviance Information Criterion (DIC), and the Logarithm of the Pseudo-Marginal Likelihood (LPML) Criterion.

2.1. Hierarchical power priors

A natural extension of the power prior in (1) can be achieved by accounting for uncertainty about the weight parameter δ . This involves adopting a hierarchical power prior where δ is treated as a random variable. To achieve this, a Beta prior distribution is assigned to δ , leading to the so-called joint unnormalized power prior (Chen and Ibrahim, 2000) for both θ and δ

$$\begin{aligned} \pi(\theta, \delta | y_0) &= \pi(\theta | y_0, \delta)\pi_0(\delta) \\ &\propto L(\theta | y_0)^\delta \pi_0(\theta)\pi_0(\delta), \end{aligned} \tag{2}$$

where $\pi_0(\theta)$ and $\pi_0(\delta)$ represent the initial priors for θ and δ , respectively.

However, as noted by Duan et al. (2006) and Neuenschwander et al. (2009), the formulation in (2) lacks a normalizing constant, leading to potential inconsistencies in the joint posterior distributions for (θ, δ) derived from different forms of likelihood functions, such as those based on raw data versus those based on the distribution of sufficient statistics (Ye et al., 2022). Consequently, Duan et al. (2006) proposed a normalized power prior (NPP) which involves first setting a conditional prior on θ given δ , followed by an initial prior distribution for δ . The resulting joint normalized power prior for (θ, δ) is

$$\begin{aligned} \pi(\theta, \delta | y_0) &= \pi(\theta | y_0, \delta)\pi_0(\delta) \\ &= \frac{L(\theta | y_0)^\delta \pi_0(\theta)}{C(\delta)} \pi_0(\delta), \end{aligned} \tag{3}$$

where the normalizing constant $C(\delta)$ is

$$C(\delta) = \int_{\Theta} L(\theta | y_0)^\delta \pi_0(\theta) d\theta. \tag{4}$$

In light of the current data, the joint posterior distribution is

$$\begin{aligned} \pi(\theta, \delta | y, y_0) &= \frac{L(\theta | y)\pi(\theta, \delta | y_0)}{\int_0^1 \int_{\Theta} L(\theta | y)\pi(\theta, \delta | y_0) d\theta d\delta} \\ &\propto L(\theta | y) \frac{L(\theta | y_0)^\delta \pi_0(\theta)}{C(\delta)} \pi_0(\delta). \end{aligned} \tag{5}$$

The marginal posterior distribution of δ is

$$\pi(\delta | y, y_0) \propto \frac{\pi_0(\delta)}{C(\delta)} \int_{\Theta} L(\theta | y)L(\theta | y_0)^\delta \pi_0(\theta) d\theta.$$

Integrating δ out in (5), the marginal posterior of θ can be written as

$$\pi(\theta | y, y_0) \propto \pi_0(\theta)L(\theta | y) \int_0^1 \frac{L(\theta | y_0)^\delta \pi_0(\delta)}{C(\delta)} d\delta.$$

The joint power prior framework offers the advantage of incorporating uncertainty regarding the weight parameter δ into the power prior formulation. This approach allows the data to determine the appropriate weight for historical information based on its compatibility with current observations. Furthermore, explicitly accounting for this uncertainty increases the flexibility in modeling the agreement between historical and current data.

In addition, a crucial theoretical advantage of the joint normalized power prior with respect to the formulation in (2) is its adherence to the likelihood principle. This ensures that the posterior distributions in (5) accurately reflect the compatibility between the current and historical data. Furthermore, this approach has further theoretical justification, as the power prior formulation in (3) is shown to minimize the weighted KL divergence (Ye et al., 2022).

From a computational perspective, the additional effort required for the normalized power prior compared to the unnormalized power prior involves computing the normalizing constant in (4). For certain models, this integral can be solved analytically, resulting in a closed-form expression for the joint posterior as specified in (5). However, for more complex models, the normalizing constant $C(\delta)$ must be determined numerically. Consequently, the posterior distribution in (5) falls into the category of the doubly intractable distributions (Carvalho and Ibrahim, 2021), and numerical methods such as Markov Chain Monte Carlo (MCMC) (Robert and Casella, 2004) are required.

Table 1
Scale of evidence for the BF proposed by Jeffreys (1961).

$BF_{0,1}$	$\log_{10}(BF_{0,1})$	Evidence Category
1 - 3.16	0 - 0.5	Barely worth mentioning
3.16 - 10	0.5 - 1	Substantial evidence for \mathcal{H}_1
10 - 31.62	1 - 1.5	Strong evidence for \mathcal{H}_1
31.62 - 100	1.5 - 2	Very strong evidence for \mathcal{H}_1
> 100	> 2	Decisive evidence for \mathcal{H}_1

3. The calibrated Bayes factor

Eliciting a well-balanced initial prior distribution for δ has proven to be challenging. Intuitively, this prior should encourage borrowing when the data are compatible and limit borrowing when they are in conflict. In this section, we propose a calibrated Bayes factor, hereafter CBF, which is a simulation-based algorithmic technique designed to effectively discriminate between some competing initial Beta prior distributions for δ . The proposed CBF aims to provide more robust decisions about which initial Beta prior for δ may be used. Specifically, this involves analyzing the behavior of the Bayes factor (Jeffreys, 1961; Kass, 1993; Kass and Raftery, 1995), henceforth BF, using different hypothetical replications generated from the posterior predictive distributions, while assessing how surprising the observed value of the BF is.

3.1. The Bayes factor

The BF provides a general Bayesian method to assess the relative evidence in support of competing hypotheses based on their compatibility with the observed data. Furthermore, the BF represents the ratio between the posterior and the prior odds when comparing two distinct point hypotheses. Specifically, let consider two competing hypotheses about the initial prior distribution of the weight parameter

$$\mathcal{H}_0 : \delta \sim \text{Beta}(\eta_0, \nu_0) \text{ vs. } \mathcal{H}_1 : \delta \sim \text{Beta}(\eta_1, \nu_1),$$

where η_i and ν_i represent the strictly positive Beta shape parameters under the hypothesis \mathcal{H}_i , for $i = \{0, 1\}$. The corresponding BF can be expressed as the ratio of the two marginal likelihoods

$$BF_{0,1}(y) = \frac{m(y | \mathcal{H}_1)}{m(y | \mathcal{H}_0)}, \tag{6}$$

where the marginal likelihood is

$$m(y | \mathcal{H}_i) = \int_0^1 \frac{\int_{\Theta} L(\theta | y) L(\theta | y_0)^\delta \pi_0(\theta) \pi_0(\delta | \mathcal{H}_i) d\theta}{\int_{\Theta} L(\theta | y_0)^\delta \pi_0(\theta) d\theta} d\delta,$$

with $\pi_0(\delta | \mathcal{H}_i)$ representing the initial Beta prior distribution for δ under the hypothesis \mathcal{H}_i , for $i = \{0, 1\}$. Assuming equal prior probabilities for both models, a BF exceeding one suggests stronger evidence in favor of \mathcal{H}_1 . Conversely, a BF below one denotes stronger evidence for \mathcal{H}_0 . A BF close to one indicates no clear preference for either hypothesis, reflecting a similar degree of empirical evidence for both \mathcal{H}_0 and \mathcal{H}_1 . Several approaches have been developed to summarize and classify the strength of evidence according to the observed BF. Firstly, Jeffreys (1961) introduced a categorization, as illustrated in Table 1. Subsequently, Kass and Raftery (1995) simplified this scale by omitting one category and redefining the thresholds. Lastly, Lee and Wagenmakers (2014, Table 7.1, p.105) further refined the Jeffreys scale with additional modifications.

For most complex models, the BF computation is challenging, as the marginal likelihood is not analytically tractable. Therefore, numerical approximation methods become essential. A widely used algorithm is the so-called bridge sampling (Meng and Wong, 1996). This method employs a Monte Carlo technique, generating samples from an auxiliary distribution that bridges the model’s prior and posterior distribution. The generated samples are then used to calculate bridge sampling weights, which correct the bias introduced by the auxiliary distribution, providing an unbiased estimate of the marginal likelihood.

Another notable method is the Savage–Dickey algorithm (Dickey and Lientz, 1970). This method approximates the BF by calculating the ratio of the posterior and prior densities at a model parameter value of zero. However, its use is limited to nested models and may be unstable if the posterior density significantly deviates from zero.

3.2. Simulation-based calibration

The BF has an inherent dependence on the observed data when used for decision making. Consequently, decisions based solely on current observations may lead to potentially misleading conclusions due to the possible fluctuations and noise present in the data. Furthermore, Schad et al. (2023) emphasized two crucial issues regarding BF computations: the instability of BF estimates in complex statistical models and the potential bias within these estimates. Therefore, to effectively and responsibly employ the BF, it is crucial to adjust and calibrate it, ensuring that the conclusions drawn are more robust and reliable.

The concept of simulation-based calibration, hereafter SBC, was originally developed to validate the computational correctness of applied Bayesian inference methods (Geweke, 2004; Cook et al., 2006; Talts et al., 2018; Gelman et al., 2020; Schad et al., 2021; Modrák et al., 2023, among others). In addition, Schad et al. (2023) proposed a structured approach based on SBC to verify the accuracy of the BF calculations. Their calibration method involves simulating multiple artificial datasets to assess whether a BF estimated in a given analysis is accurate or biased. This type of calibration approach is intuitive and logical, as it mirrors the classical approach of hypothesis testing, where the decision-making criterion is determined by the sampling distribution under the hypothesis of repeated sampling.

The SBC-inspired method explored in this paper is motivated by the insights of Garcia-Donato and Chen (2005), who posited that BFs should be considered as a random variable before observing the sample. This perspective emphasizes the necessity of calibrating the BF to accurately account for the inherent randomness in the data. However, the analytic form of the BF distribution is frequently not available. In such cases, a simulation-intensive approach becomes a valuable tool to approximate this distribution (Vlachos and Gelfand, 2003). This involves generating replicated datasets from some type of predictive distribution or other data-generating processes, estimating the marginal likelihoods with these synthetic datasets, and then computing the BFs. By iterating this process multiple times, an approximation of the BF distribution is obtained. Subsequently, once the data have been collected, the observed BF can be used as a measure of agreement between the observed data and the underlying statistical model.

Both Garcia-Donato and Chen (2005) and Schad et al. (2023) suggested a calibration method based on the prior predictive distribution. However, this approach is not suitable in our context due to the inherent bias in the replicated data from the prior predictive distribution toward the historical data y_0 , as highlighted by the structure of Equation (3). This can potentially yield replicated samples that are much more in agreement with the historical data than the historical data are with the current data. Consequently, we explored the use of hypothetical replications generated from the posterior predictive distribution to approximate the distribution of the BF; where the predictive distribution is given by

$$p(y^{\text{rep}} | y) = \int_{\Theta} L(y^{\text{rep}} | \theta) \pi(\theta | y, y_0, \delta) d\theta.$$

A key advantage of this approach is the use of the information in the current data y through the posterior distribution, focusing on a relevant region within the parameter space (Robert, 2022). Furthermore, the BF computed using the replications from the posterior predictive distribution $\text{BF}_{0,1}(y^{\text{rep}})$, henceforth the replicated BF, mimics the behavior of the BF using the original data y when a specific model is the true data-generating mechanism. Our CBF approach aims to define a decision criterion that not only assesses the inherent variability of the BF but also incorporates the observed data. Thus, to effectively ensure a more comprehensive and balanced decision rule, it is essential to define a criterion that incorporates the observed BF, denoted by $\text{BF}_{0,1}(y)$, as a measure of surprise, favoring scenarios where it is less unexpected. We propose a criterion based on

- The survival function of the BF distribution, denoted by $S_{\text{BF}_{0,1}}(\cdot)$.
- The inclusion of the observed BF within a defined Highest Posterior Density Interval (HPDI).

Specifically, the decision criterion focuses on selecting alternative hypotheses that provide stronger evidence against the null hypothesis. This is achieved by giving preference to the BF distribution that yields more values in favor of the alternative hypothesis. This implies that the survival function of the BF distribution, calculated at the value where the BF indicates equal support for both hypotheses, is greater than 0.5. Furthermore, the inclusion of the observed BF within a defined HPDI assesses its coherence with respect to the underlying distribution. This dual approach ensures a balanced and comprehensive decision rule, accounting for both the BF distribution and the surprise measure associated with the observed BF.

3.3. The procedure

To streamline our method, we focus on the logarithmic transformation of the BF, referred to as Log-BF. This transformation is advantageous because values less than zero suggest evidence for the null hypothesis \mathcal{H}_0 , while values greater than zero provide evidence for the alternative hypothesis \mathcal{H}_1 . To further simplify comparisons and improve interpretability, we use a reference null hypothesis \mathcal{H}_0 computing the Log-BF between each alternative hypothesis and the reference, reducing the number of comparisons from $\binom{M}{2}$ to $(M - 1)$ Log-BFs. The chosen reference null hypothesis is the Beta(1, 1) prior for δ , a commonly used non-informative prior for the weight parameter (Ibrahim et al., 2015). Notably, this choice is motivated by our goal of eliciting a more informative prior distribution for the weight parameter compared to the standard uniform prior.

The initial step involves defining a reasonable grid of potential alternative hypotheses related to the Beta initial prior for the weight parameter, denoted by $\mathcal{H} = \{\mathcal{H}_1, \dots, \mathcal{H}_i, \dots, \mathcal{H}_M\}$, with

$$\mathcal{H}_i : \delta \sim \text{Beta}(\eta_i, \nu_i), \tag{9}$$

where η_i and ν_i represent the shape parameters under \mathcal{H}_i , for $i = 1, \dots, M$. In particular, the grid should explore parameter space regions ranging from scenarios with minimal borrowing of historical information to those with extensive borrowing. We suggest defining a grid that spans from 0.5 to 6, covering a wide range of Beta priors. For instance, a Beta(0.5, 6) prior assigns minimal weight to historical data, while a Beta(6, 0.5) prior incorporates a considerable amount of information from the historical study.

After computing the observed Log-BF between hypothesis \mathcal{H}_i and \mathcal{H}_0 , the next step involves generating K hypothetical samples from the posterior predictive distribution under each alternative hypothesis \mathcal{H}_i , that is, $\mathbf{y}_{\mathcal{H}_i}^{\text{rep}} = (y_{\mathcal{H}_i,1}^{\text{rep}}, \dots, y_{\mathcal{H}_i,K}^{\text{rep}})$, for i, \dots, M . Then, the

Algorithm 1 Calibrated Bayes Factor Procedure.

- 1: Define a grid of size M of potential competing hypotheses regarding the Beta prior for the weight parameter $\mathbf{H} = \{\mathcal{H}_1, \dots, \mathcal{H}_i, \dots, \mathcal{H}_M\}$, with $\mathcal{H}_i : \delta \sim \text{Beta}(\eta_i, \nu_i)$, for $i = 1, \dots, M$.
- 2: For $i = 1, \dots, M$ compute the Observed Log-BF between hypothesis \mathcal{H}_i and \mathcal{H}_0 as in (6).
- 3: For $i = 1, \dots, M$ generate K hypothetical samples from the posterior predictive distribution under the alternative hypothesis \mathcal{H}_i , that is $\mathbf{y}_{\mathcal{H}_i}^{\text{rep}} = (y_{\mathcal{H}_i,1}^{\text{rep}}, \dots, y_{\mathcal{H}_i,k}^{\text{rep}})$.
- 4: For $i = 1, \dots, M$ and for $k = 1, \dots, K$ compute the replicated Log-BF, $\log \text{BF}_{0,i}(\mathbf{y}_{\mathcal{H}_i,k}^{\text{rep}})$, between hypothesis \mathcal{H}_i and \mathcal{H}_0 as in (6) using the hypothetical posterior predictive samples generated in Step 3.
- 5: Select the Beta prior for δ using the criterion in (8).

replicated Log-BF is computed between the alternative hypothesis \mathcal{H}_i and the null hypothesis \mathcal{H}_0 , for all combinations of $i = 1, \dots, M$ and $k = 1, \dots, K$.

Subsequently, it is essential to define a criterion based on the distribution of the Log-BF, obtained using the replicated Log-BFs, and that incorporates the observed Log-BF as a measure of surprise, favoring scenarios where it is less unexpected. Our criterion aims to identify the BF distributions in which the alternative hypothesis \mathcal{H}_i , as given in (7), is more likely than the null hypothesis $\mathcal{H}_0 : \delta \sim \text{Beta}(1, 1)$. This is obtained when the cumulative distribution function (CDF) of the Log-BF at zero is below 0.5 – or when the survival function at zero exceeds 0.5 – suggesting stronger theoretical evidence in favor of the alternative hypothesis. The observed Log-BF is also incorporated in our decision criterion by considering the hypotheses associated with values greater than zero, reflecting stronger empirical evidence relative to the null hypothesis. The robustness of the observed Log-BF is evaluated by assessing its position within the approximated distribution. Ideally, the observed Log-BF should be within a specific HPDI, indicating that it is not an outlier but rather a value consistent with the underlying Log-BF distribution. Consequently, the well-balanced prior is determined using the following criterion

$$\mathcal{H}_{\text{Opt}} = \begin{cases} \mathcal{H}^* & \text{if } \mathcal{H}^* > 0 \\ \mathcal{H}_0 & \text{otherwise,} \end{cases} \tag{8}$$

with

$$\mathcal{H}^* = \arg \max_{i=1, \dots, M} \left\{ \mathbb{1}_{\{S_{\log \text{BF}_{0,i}}(0) > 0.5\}} S_{\log \text{BF}_{0,i}}(0) \times \mathbb{1}_{\{\log \text{BF}_{0,i}^{\text{obs}} \in (\cdot)\% \text{HPDI}\}} \log \text{BF}_{0,i}^{\text{obs}} \right\},$$

where $S_{\log \text{BF}_{0,i}}(0)$ represents the survival function of the Log-BF distribution evaluated at zero, and $\log \text{BF}_{0,i}^{\text{obs}}$ is the observed Log-BF between the alternative hypothesis \mathcal{H}_i and the null hypothesis \mathcal{H}_0 , for $i = 1, \dots, M$. The first indicator function focuses on distributions where the alternative hypothesis is more probable than the null hypothesis. That is, selecting distributions where the survival function of the Log-BF at zero is greater than 0.5, indicating higher theoretical evidence for the alternative hypothesis. The second indicator function evaluates the presence of the observed Log-BF within a specific HPDI, working from a Bayesian perspective as a measure of surprise. The computational steps of the CBF procedure are summarized in Algorithm 1.

Fig. 1 provides an illustrative example of the CBF procedure for selecting a well-balanced initial Beta prior for δ . Let consider a null hypothesis \mathcal{H}_0 and three alternative hypotheses \mathcal{H}_i , for $i = 1, 2, 3$, regarding the initial Beta prior of δ . The Log-BF distribution for \mathcal{H}_1 , represented by the purple curve, shows a higher probability of negative Log-BF values, suggesting stronger theoretical evidence in favor of the null hypothesis. Conversely, the approximated Log-BF distributions for \mathcal{H}_2 and \mathcal{H}_3 , represented by the orange and green curves, respectively, provide stronger evidence in favor of the associated alternative hypotheses. However, although \mathcal{H}_2 shows an observed Log-BF within the selected HPDI, denoted by the dashed lines, the corresponding observed Log-BF value is negative, suggesting empirical evidence in favor of \mathcal{H}_0 . Only \mathcal{H}_3 demonstrates a positive observed Log-BF, which suggests stronger empirical evidence for the alternative hypothesis, and also falls within the respective HPDI. Consequently, based on the selection criterion in (8), a well balanced prior for δ is the one associated with the alternative hypothesis \mathcal{H}_3 .

4. Simulation studies

In this section, we assess the efficacy and applicability of the CBF approach through simulation studies. Specifically, we consider three distinct scenarios, each involving different statistical distributions and the corresponding parameters. The goal of each simulated study is to evaluate the method’s ability to identify a well-balanced prior that effectively integrates extensive historical information when current data closely align with the historical data, while minimizing such integration in presence of disagreement. To assess this, we analyze a historical dataset and a series of current datasets that progressively diverge from it. We compare the performance of the proposed CBF approach against other dynamic information borrowing methods, including the NPP with a uniform prior on δ , the robust meta-analytic predictive (RMAP) prior (Schmidli et al., 2014), and the self-adapting mixture (SAM) prior (Yang et al., 2023). Notably, the RMAP prior employed assigns equal weight ω to both the non-informative mixture component and the informative mixture component based on the historical data.

As described in Section 3.3, an expanded grid for the Beta prior parameters is used, ranging from 0.5 to 6 in increments of 0.5. According to some sensitivity checks, each simulation study considers a 75% HPDI threshold for the selection criterion in (8).

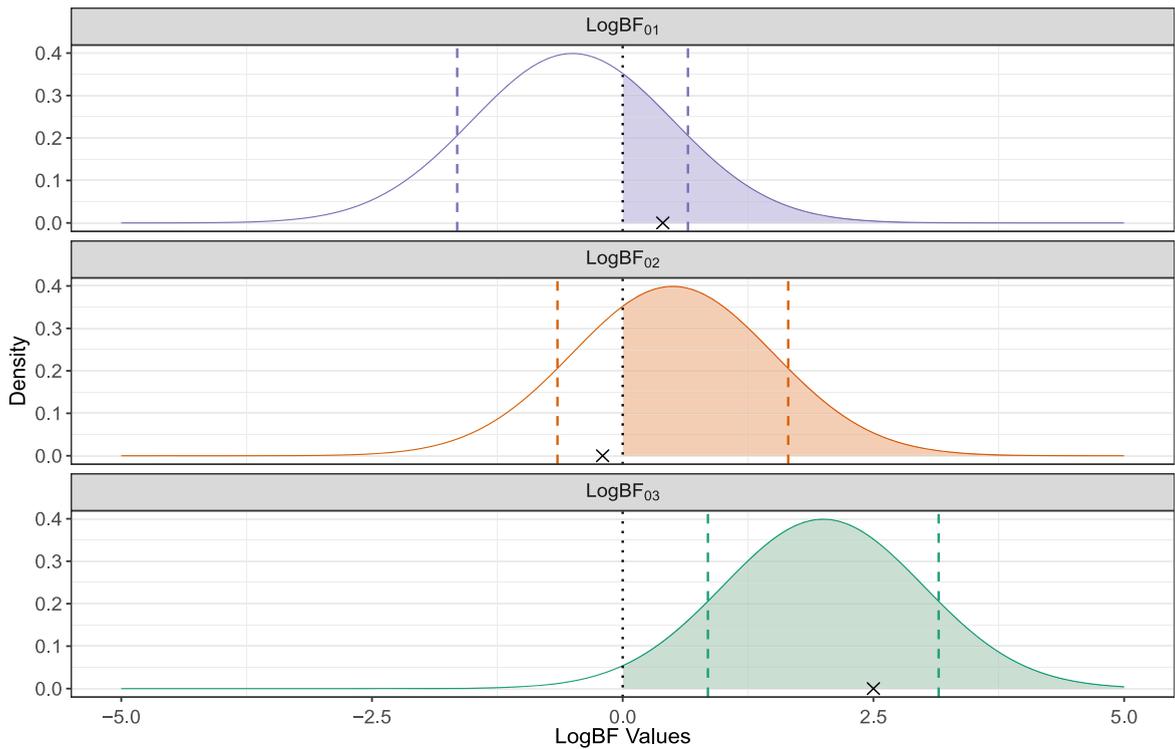


Fig. 1. Illustrative example of the CBF procedure. The Log-BF distributions (purple: \mathcal{H}_1 vs. \mathcal{H}_0 , orange: \mathcal{H}_2 vs. \mathcal{H}_0 , green: \mathcal{H}_3 vs. \mathcal{H}_0), the observed Log-BFs (crossed points) and HPDIs (dashed lines) are shown. The filled area after the zero threshold (dotted line) in each distribution represents the portion where the evidence for the alternative hypothesis is stronger than for the null hypothesis. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

4.1. Poisson log-linear model

Let $y_0 = (y_{0,1}, \dots, y_{0,N_0})$ and $y = (y_1, \dots, y_N)$ be the count outcome of a historical and a current study, respectively. Let denote with $\mathbf{x}_{0,k} = (x_{0,k1}, \dots, x_{0,kp})$, for $k = 1, \dots, N_0$, and $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})$, for $j = 1, \dots, N$, the corresponding covariate vector. The Poisson log-linear model is

$$y_j | \boldsymbol{\beta} \sim \text{Poisson}(\lambda_j), \quad \text{where } \lambda_j = \exp(\mathbf{x}_j^\top \boldsymbol{\beta}),$$

with $\boldsymbol{\beta}$ being the p -dimensional vector of the regression coefficients. Let $\pi_0(\boldsymbol{\beta})$ be the initial multivariate normal prior on $\boldsymbol{\beta}$, the BF is given by

$$\text{BF}_{0,i}(y) = \frac{\int_0^1 \int L(\boldsymbol{\beta} | y, X) [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) \text{Beta}(\delta | \eta_i, \nu_i) d\boldsymbol{\beta} d\delta}{\int_0^1 \int L(\boldsymbol{\beta} | y, X) [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) \text{Beta}(\delta | 1, 1) d\boldsymbol{\beta} d\delta},$$

where $[L(\boldsymbol{\beta} | y_0, X_0)]^\delta = \exp\left(\sum_{k=1}^{N_0} \left[\delta(y_{0,k} \mathbf{x}_{0,k}^\top \boldsymbol{\beta} - \exp(\mathbf{x}_{0,k}^\top \boldsymbol{\beta}))\right]\right) \prod_{k=1}^{N_0} (y_{0,k}!)^{-\delta}$. For further details, see Appendix A.

In this simulation study, we first generate a historical dataset that includes three distinct covariates and an intercept. We then generate a series of current datasets that gradually diverge from the historical data by increasing the regression coefficient $\beta_{c,1}$ associated with the first covariate. Fig. 2 illustrates the evolution of the prior distribution selected for δ in relation to the level of disagreement between current and historical studies. This disagreement is quantified by the difference between the historical and current regression coefficients. The plot shows the median values of the selected prior for δ as points, with error bars representing the first and third quartiles. The values in brackets correspond to the chosen Beta parameters selected from the predefined grid. The plot highlights the procedure’s ability to select an appropriate prior according to the level of disagreement between the datasets. As the disagreement grows, the chosen prior shifts from a left-skewed Beta distribution – a Beta(5.5, 0.5), indicating substantial incorporation of historical data (equal weight to the actual data), to a right-skewed Beta distribution – a Beta(1.5, 5.5), reflecting more conservative integration of historical information.

Fig. 3 shows the standard deviation (SD) and mean of the marginal posterior distribution for the regression coefficient β_1 . The left panel of Fig. 3 illustrates that the marginal posterior distributions using the NPP with the CBF selected prior for δ are less diffuse than those derived from the NPP with the standard uniform prior, the RMAP prior and the SAM prior, leading to more precise results for β_1 . Since the SAM prior does not allow the incorporation of covariate information (Yang et al., 2023), we compute the mixing

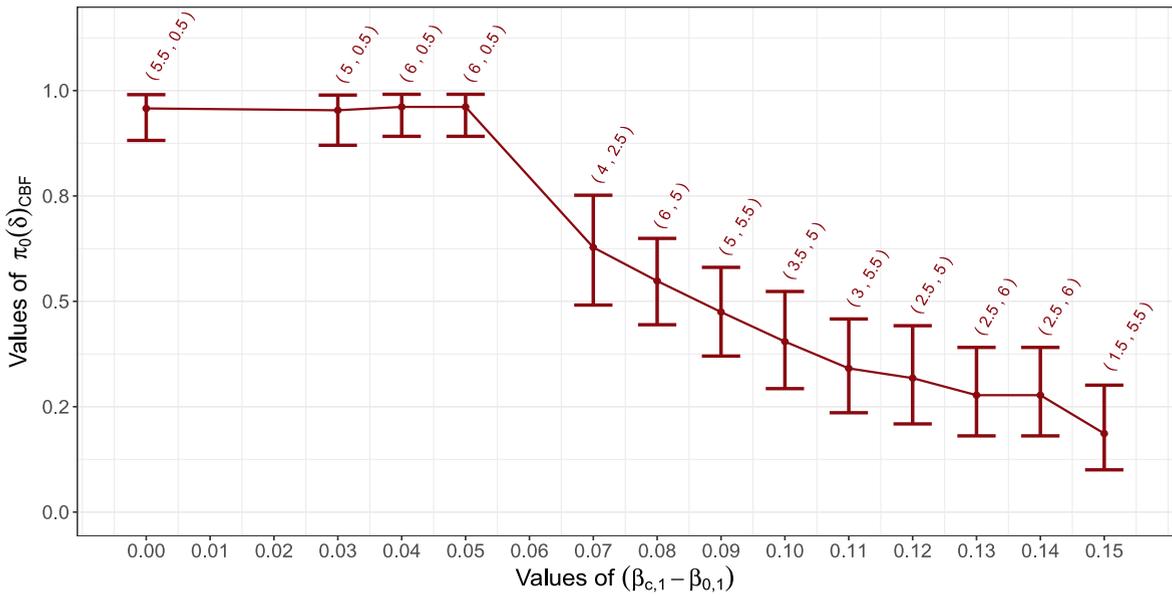


Fig. 2. Poisson log-linear model. Shifts in the prior median (point) for the CBF derived prior for δ , considering the 25th and 75th Percentiles (bar), in function of the difference between the current regression coefficient $\beta_{c,1}$ and the historical regression coefficient $\beta_{0,1}$. Above each interval, the corresponding selected Beta parameter are displayed.

weight for it using a Poisson model on the count outcome. Specifically, when there is minimal disagreement between current and historical data, the posterior standard deviation for β_1 is lower when using the CBF selected prior. As disagreement increases, the difference between the standard deviations tends to reduce. Furthermore, the right panel of Fig. 3 highlights that the posterior mean in all scenarios shows a similar increasing trend. This can be attributed to the progressive increase in the current regression coefficient $\beta_{c,1}$ which results in a greater discrepancy between current and historical data, leading to a higher discount of historical data.

Fig. 4 shows the posterior distribution of the weight parameter δ . When there is a low level of disagreement between current and historical data, the CBF prior leads to posterior distributions that incorporate more historical information compared to the uniform prior for δ . Conversely, as the disagreement increases, the CBF prior becomes more conservative, discounting the historical data to a greater extent. Furthermore, the CBF prior consistently leads to more precise estimates with lower variability in the posterior distribution.

4.2. Binomial model

A similar simulation study is conducted for the binomial model, which is frequently applied in medical contexts involving power priors. Let N_0 and N denote the number of Bernoulli trials in the historical and current studies, respectively. Furthermore, let y_0 and y represent the number of successes in these studies. Assuming a binomial likelihood with success probability θ for each study, and an initial Beta prior for both θ and the weight parameter δ , the BF is

$$BF_{0,i}(y) = \frac{\int_0^1 \text{BBin}(y | N, \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | \eta_i, \nu_i) d\delta}{\int_0^1 \text{BBin}(y | N, \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | 1, 1) d\delta}, \quad i = 1, \dots, M,$$

where $\text{BBin}(\cdot | N, \alpha, \beta)$ is the beta-binomial discrete distribution. For further details, see Appendix B.

Fig. 5 shows the evolution of the selected prior for δ when analyzing a historical dataset followed by a series of current datasets. A $\text{Beta}(1, 1)$ is used as the initial prior distribution for θ , and the historical binomial likelihood presents a success probability of 0.2. This simulation study demonstrates the ability of the proposed method to dynamically adapt the amount of historical information borrowed, based on the agreement between current and historical data. Specifically, when there is almost perfect agreement, the selected prior for δ is $\text{Beta}(6, 0.5)$, indicating a higher level of historical information borrowing. As the level of agreement decreases, the prior for δ progressively shifts to Beta distributions that reduce the incorporation of historical data, reaching a $\text{Beta}(0.5, 6)$ distribution in cases of high disagreement.

Fig. 6 presents the marginal posterior standard deviations and means for θ comparing the NPP with the CBF selected initial prior for the weight parameter with the other dynamic information borrowing methods. Specifically, the left panel of Fig. 6 highlights that the CBF prior for δ yields marginal posterior distributions for θ that are generally more concentrated. This is particularly evident when there is either a low or high level of disagreement between current and historical data, compared to the NPP with a uniform prior on δ . Conversely, when compared to the RMAP prior and the SAM prior, the NPP using the CBF procedure produces less diffuse posterior distributions when there is a moderate level of disagreement between current and historical data. Furthermore, the right

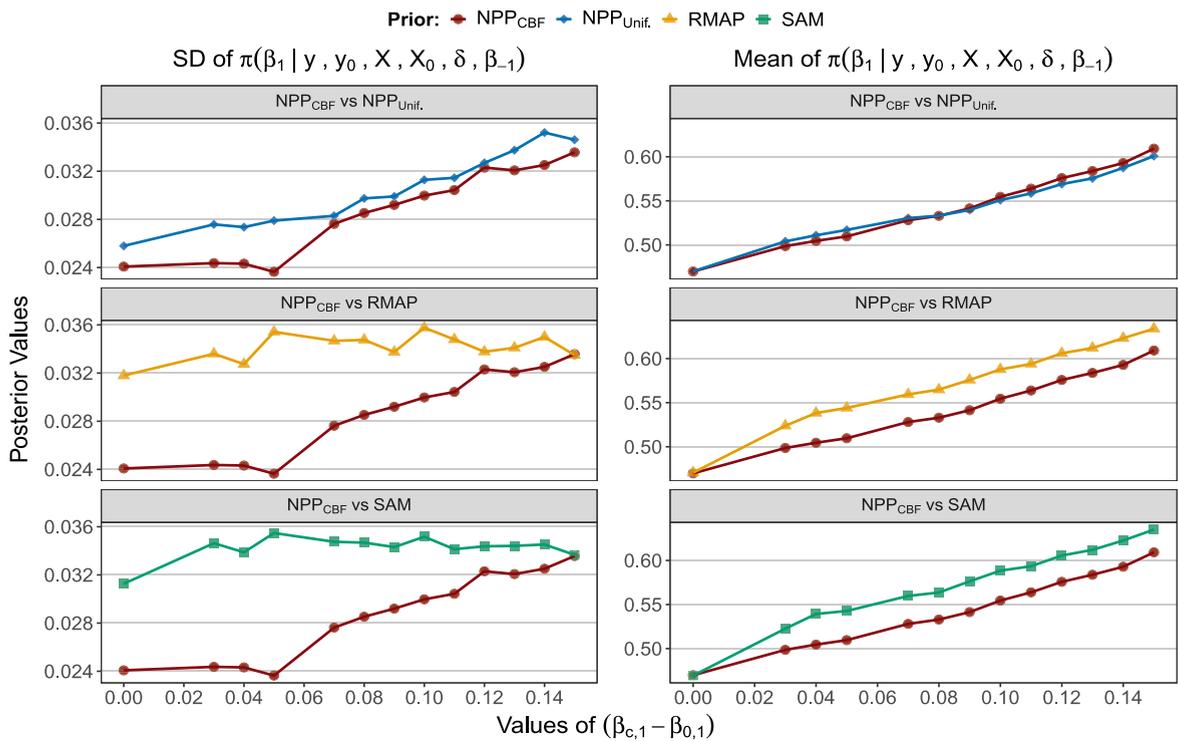


Fig. 3. Poisson log-linear model. Standard deviation (SD) and mean of the marginal posterior for β_1 , comparing four different priors: The normalized power prior (NPP) using the CBF derived Beta prior on δ (red dotted lines), the NPP using the standard uniform prior (blue diamond lines), the robust meta-analytic predictive prior (RMAP) (yellow triangular lines), and the self-adapting mixture prior (SAM) (green squared lines).

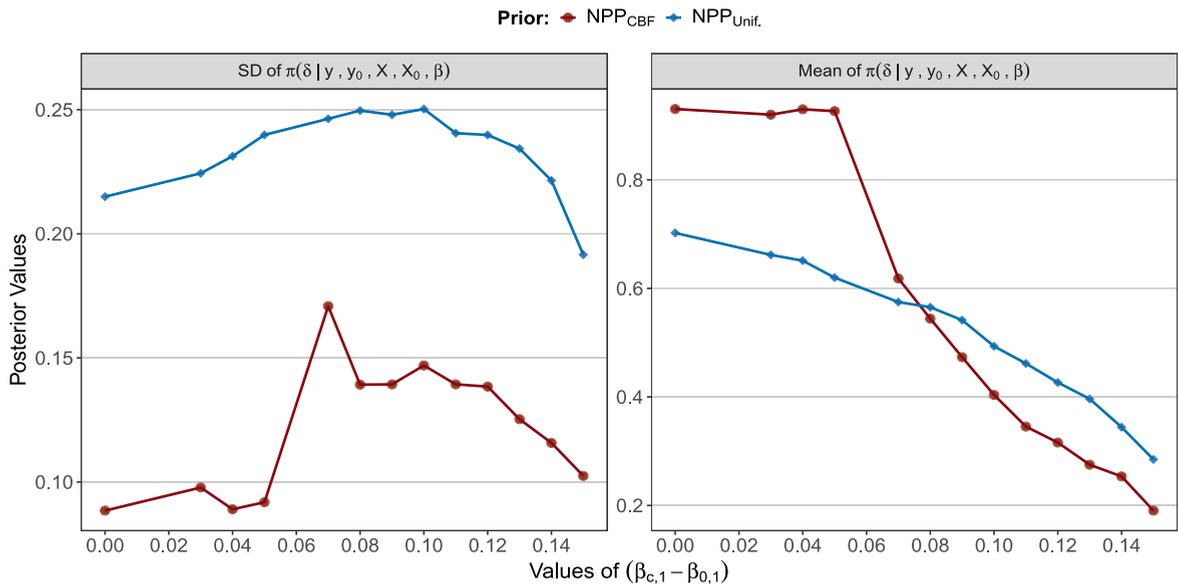


Fig. 4. Poisson log-linear model. Standard deviation (SD) and mean of the marginal posterior for δ using the standard uniform prior (blue squared lines) and the CBF derived Beta prior (red dotted lines).

panel of Fig. 6 shows that the posterior mean of θ follows a trend consistently similar to that obtained using the competing information borrowing methods. Therefore, using the CBF prior leads to more accurate inferential conclusions in general.

Fig. 7 shows that when the discrepancy between the historical and current data is minimal, the CBF prior for δ produces posterior distributions that integrate more historical information. As the discrepancy increases, the posterior distributions become more conservative, progressively discounting the historical data. Furthermore, the CBF prior for δ generates less diffuse posterior distributions compared to those obtained from a uniform prior.

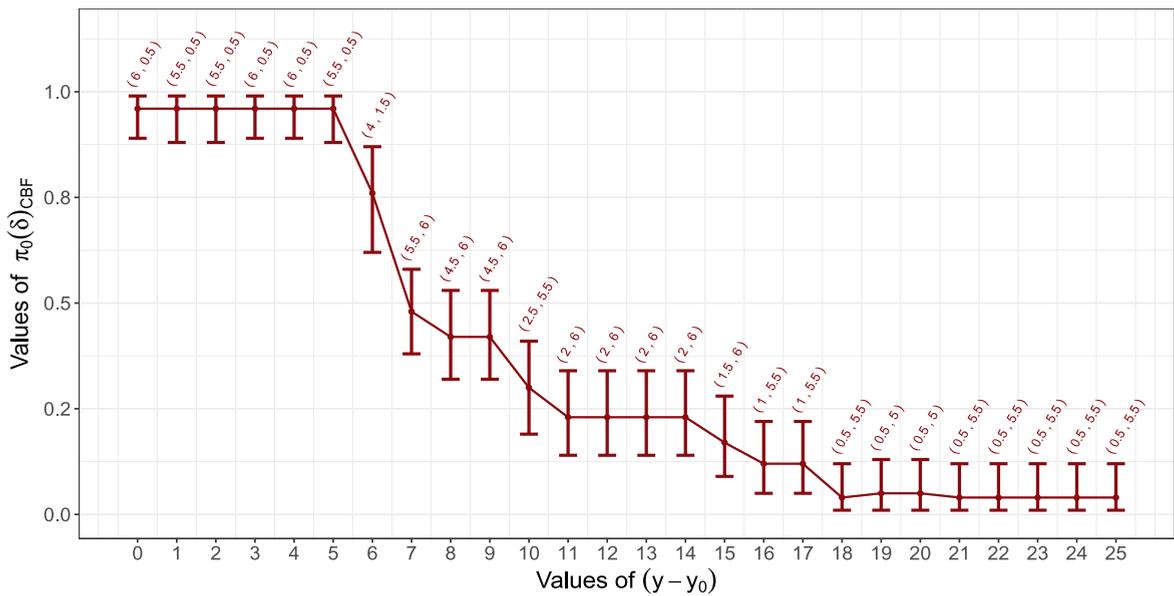


Fig. 5. Binomial model. Shifts in the prior median (point) for the CBF derived prior for δ , considering the 25th and 75th Percentiles (bar), in function of the difference between the current successes y and the historical successes y_0 . Above each interval, the corresponding selected Beta parameter are displayed.

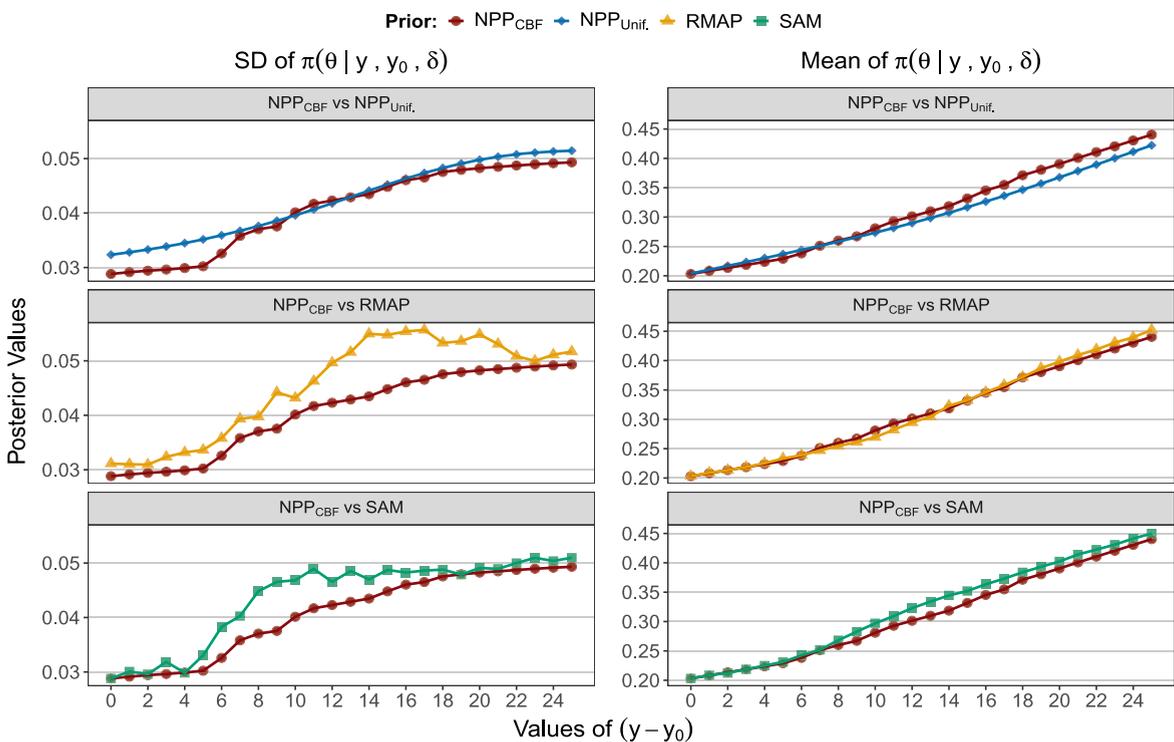


Fig. 6. Binomial model. Standard deviation (SD) and mean of the marginal posterior for θ , comparing four different priors: The normalized power prior (NPP) using the CBF derived Beta prior on δ (red dotted lines), the NPP using the standard uniform prior (blue diamond lines), the robust meta-analytic predictive prior (RMAP) (yellow triangular lines), and the self-adapting mixture prior (SAM) (green squared lines).

4.3. Gaussian model

We examine a Gaussian model commonly observed in replication studies (Pawel et al., 2024). A crucial question in this context is how effectively a replication (current) study has reproduced the results of an original (historical) study. Let μ denote the unknown true

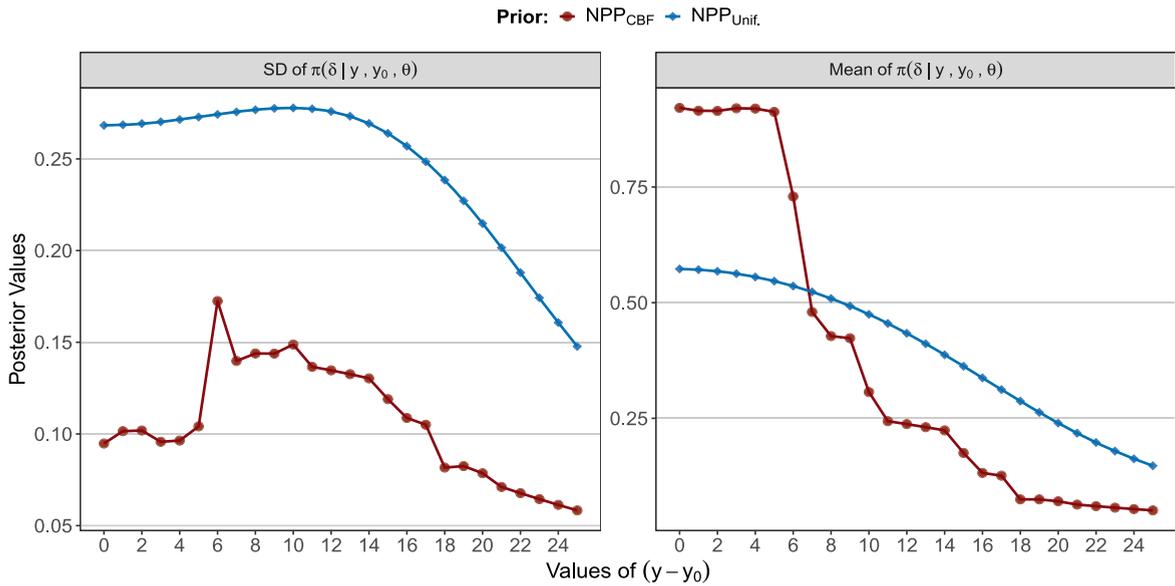


Fig. 7. Binomial model. Standard deviation (SD) and mean of the marginal posterior for δ using the standard uniform prior (blue squared lines) and the CBF derived Beta prior (red dotted lines).

effect size, and let $\hat{\mu}_s$ represent the estimated effect size from study s , where $s \in \{o, r\}$ corresponds to the “original” and “replication” studies, respectively. Furthermore, the effect size estimates are assumed to be normally distributed.

$$\hat{\mu}_s | \mu \sim N(\mu, \sigma_s^2),$$

where σ_s^2 represents the variance of the estimated effect size $\hat{\mu}_s$, assumed to be known. The BF is defined as

$$BF_{0,i}(y) = \frac{\int_0^1 N(\hat{\mu} | \hat{\mu}_o, \sigma^2 + \sigma_o^2/\delta) \text{Beta}(\delta | \eta_i, \nu_i) d\delta}{\int_0^1 N(\hat{\mu} | \hat{\mu}_o, \sigma^2 + \sigma_o^2/\delta) \text{Beta}(\delta | 1, 1) d\delta}, \quad i = 1, \dots, M.$$

Further details are provided in Appendix C.

In Fig. 8, it is assumed that the effect size in the original study follows a normal distribution with mean $\mu_o = 0$ and variance $\sigma_o^2 = 1$. We incrementally varied the true effect size of the replicated study in steps of 0.2, starting from a scenario of perfect agreement, where the replicated study’s effect size is $\mu_r = 0$, and extending to scenarios of progressively greater disagreement, reaching a point where $\mu_r = 6$. Our method effectively selected a well-balanced prior to address the plausible level of agreement between the original and replicated studies. Similarly to the binomial case, a Beta(6,0.5) prior is chosen when the agreement is high. As disagreement increases, the amount of information borrowed is progressively reduced, selecting the prior that minimizes the incorporation of historical information – a Beta(0.5, 6) – in cases of high disagreement.

Fig. 9 shows that the CBF procedure effectively selects a prior for the weight parameter, reducing the standard deviation of the marginal posterior for μ compared to the NPP with a uniform prior on δ and the RMAP prior. The NPP with the CBF prior also achieves a greater reduction in the standard deviation of the marginal posterior for μ than the SAM prior. However, when the level of disagreement is high or low, the behavior of the NPP with the CBF prior and the SAM prior becomes more comparable. As in previous simulation studies, the posterior mean for the effect size μ remains consistent in all the methods evaluated.

Furthermore, Fig. 10 shows results similar to those observed in previous simulation studies for the posterior distribution of δ .

5. Melanoma clinical trial

The efficacy of the CBF procedure on real data is evaluated by analyzing two clinical trials. This analysis incorporates recent data from a new trial along with historical data from a previous study to assess earlier findings. Two melanoma trials conducted by the Eastern Cooperative Oncology Group (ECOG), specifically E2696 and E1694, are examined, involving 105 and 200 patients, respectively. For further details, see Kirkwood et al. (2001a,b). These trials investigate the effects of interferon alfa-2b (IFN) treatment on patient survival rates. The E2696 trial evaluates the efficacy of combining the GM2-KLH/QS-21 (GMK) vaccine with high-dose IFN therapy compared to the GMK vaccine alone in patients with resected high-risk melanoma. Furthermore, the E1694 trial evaluates the effectiveness of the GMK vaccine versus high-dose IFN therapy in a comparable group of patients. In conclusion, the findings of the E1694 trial corroborate earlier results of E2696, demonstrating that intravenous and subcutaneous IFN can significantly reduce the relapse rate in patients with melanoma. Fig. 11 presents the survival curves for both trials, highlighting the beneficial impact of interferon treatment on patient survival.

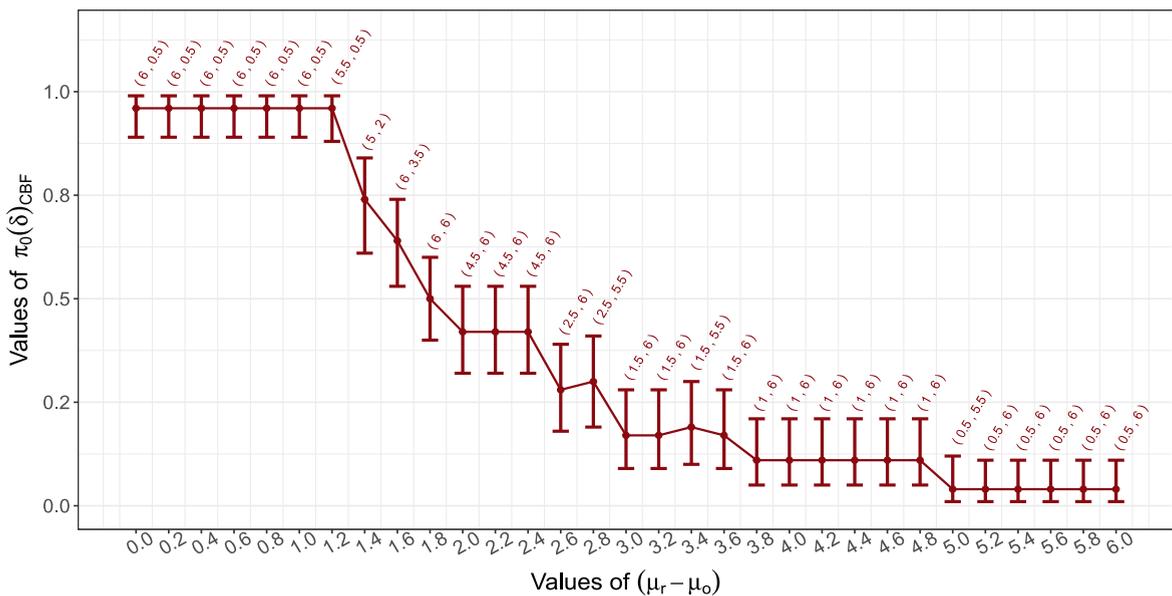


Fig. 8. Gaussian model. Shifts in the prior median (point) for the CBF derived prior for δ , considering the 25th and 75th Percentiles (bar), as a function of the difference between the replicated study's effect size μ_r and the original study's effect size μ_0 . Above each interval, the corresponding selected Beta parameter are displayed.

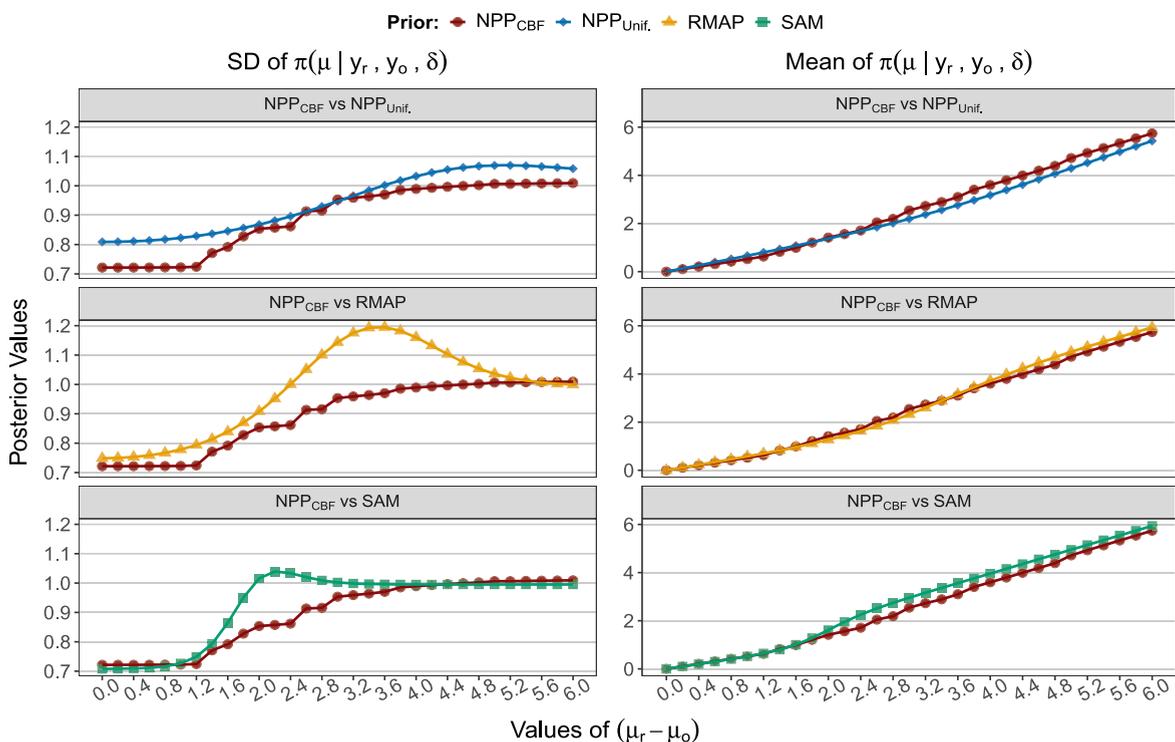


Fig. 9. Gaussian model. Standard deviation (SD) and mean of the marginal posterior for μ , comparing four different priors: The normalized power prior (NPP) using the CBF derived Beta prior on δ (red dotted lines), the NPP using the standard uniform prior (blue diamond lines), the robust meta-analytic predictive prior (RMAP) (yellow triangular lines), and the self-adapting mixture prior (SAM) (green squared lines).

A Bayesian logistic regression model is applied to the data from the E1694 trial and additional historical information from the E2696 trial is integrated using a NPP as in (3). The analysis includes four additional covariates: age, sex, performance status, and treatment indicator. Parameter estimation is conducted using the probabilistic programming language Stan (Carpenter et al., 2017) to perform Markov Chain Monte Carlo (MCMC) sampling via the `hdbayes` R package (Alt et al., 2024b). This involves four independent

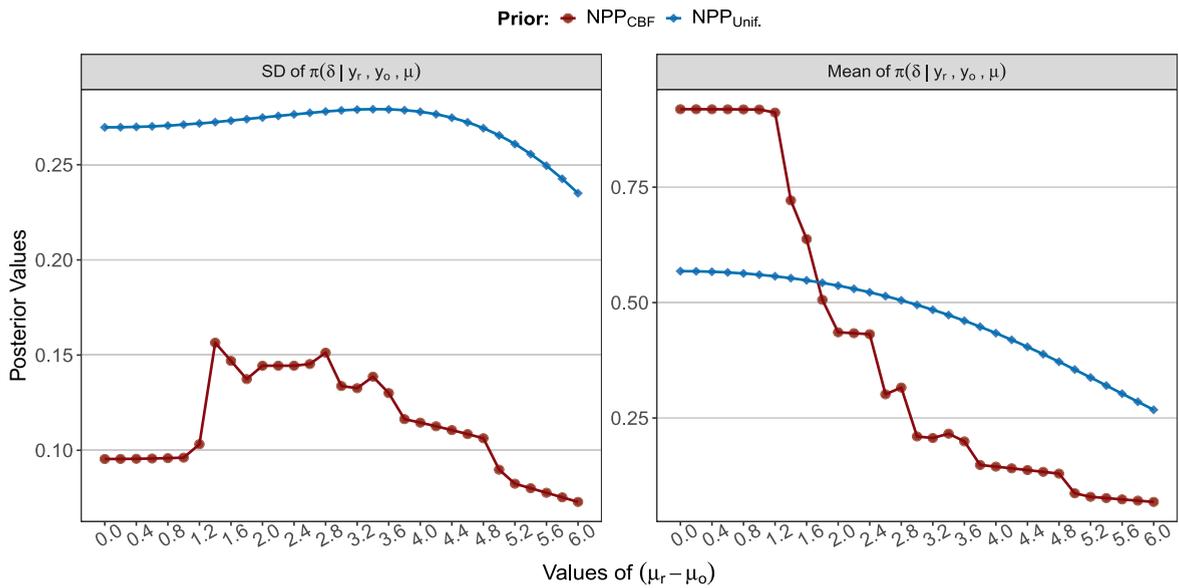


Fig. 10. Gaussian model. Standard deviation (SD) and mean of the marginal posterior for δ using the standard uniform prior (blue squared lines) and the CBF derived Beta prior (red dotted lines).

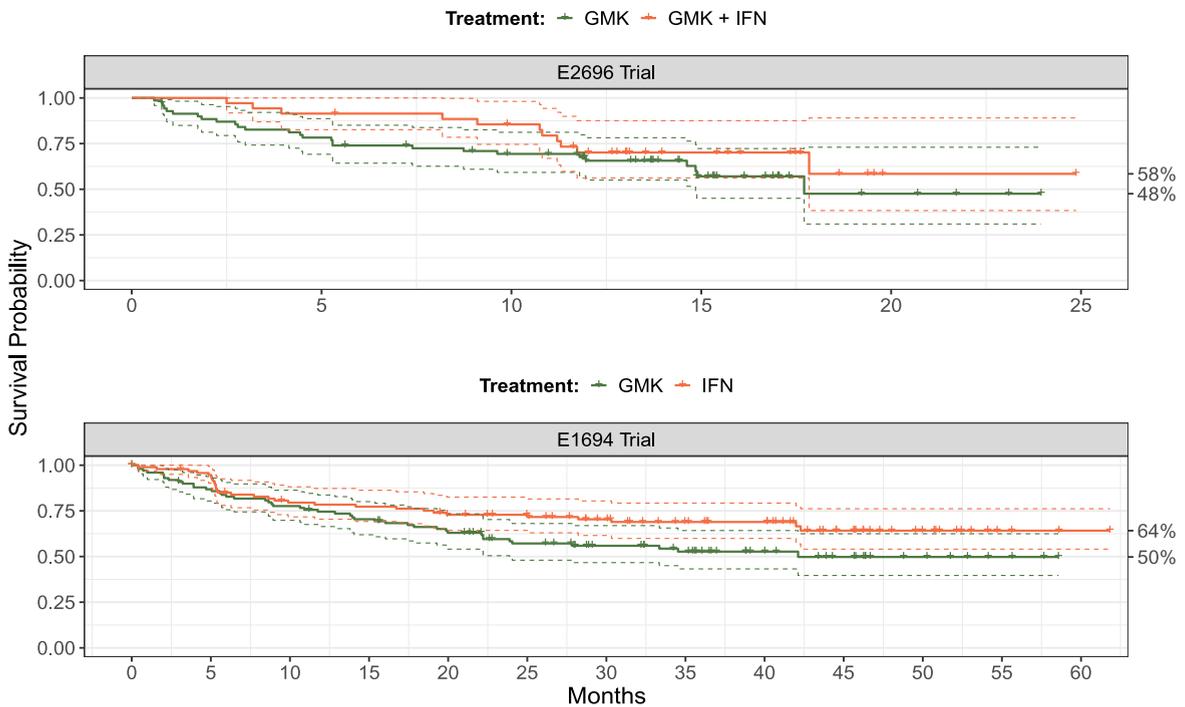


Fig. 11. Melanoma clinical trial. Survival curves for E2696 and E1694 trials. On the right, the end time survival probability for each treatment.

chains, each with 2000 iterations, discarding the first 1000 iterations as burn-in. To determine a well-balanced initial prior for the weight parameter using the CBF procedure, as described in Section 3.3, the bridge sampling approximation of the BF is used.

The initial priors for the coefficients of the four covariates are set using a weakly informative approach as outlined by Gelman et al. (2008). Specifically, a normal distribution with mean 0 and standard deviation 10 is assigned to each coefficient. Furthermore, for the initial prior of the weight parameter, a Beta(5,0.5) is chosen based on the CBF procedure. This process involves a comprehensive evaluation of competing initial priors, considering a range of different Beta parameters from 0.5 to 6 in increments of 0.5. Furthermore, a parallel processing strategy is employed to efficiently manage the computational effort required for this extensive parameter exploration, ensuring a streamlined and effective computational execution of the methodology.

Table 2
 Melanoma clinical trial. Posterior mean, standard deviation (SD), and 95% HPDI for the age, treatment, sex, and performance status comparing different initial prior for the weight parameter of the normalized power prior (NPP), the robust meta-analytic-predictive (RMAP) prior, self-adapting mixture (SAM) prior, and the commensurate prior (CP).

Prior	Parameter	Mean	SD	95% HPDI
NPP				
Beta(5, 0.5) ⁽¹⁾	Age	0.015	0.010	(-0.004, 0.033)
	Treat.	-0.507	0.243	(-0.999, -0.055)
	Sex	-0.110	0.253	(-0.600, 0.386)
	Perf.	-0.462	0.326	(-1.106, 0.156)
Beta(1, 1) ⁽²⁾	Age	0.015	0.010	(-0.006, 0.033)
	Treat.	-0.525	0.267	(-1.035, -0.013)
	Sex	-0.138	0.276	(-0.679, 0.374)
	Perf.	-0.490	0.346	(-1.135, 0.199)
Beta(0.5, 0.5) ⁽³⁾	Age	0.015	0.010	(-0.005, 0.034)
	Treat.	-0.525	0.257	(-1.028, -0.016)
	Sex	-0.124	0.264	(-0.634, 0.385)
	Perf.	-0.479	0.350	(-1.125, 0.252)
0.5× Beta(0.5, 6) + 0.5× Beta(6, 0.5)	Age	0.015	0.010	(-0.003, 0.036)
	Treat.	-0.517	0.250	(-1.03, -0.042)
	Sex	-0.110	0.254	(-0.606, 0.396)
	Perf.	-0.473	0.358	(-1.181, 0.211)
RMAP				
$\omega = 0.5$	Age	0.015	0.010	(-0.005, 0.034)
	Treat.	-0.554	0.284	(-1.093, -0.065)
	Sex	-0.183	0.268	(-0.673, 0.368)
	Perf.	-0.590	0.394	(-1.296, 0.199)
SAM				
$\omega = 0.96$	Age	0.014	0.010	(-0.006, 0.033)
	Treat.	-0.552	0.283	(-1.086, -0.046)
	Sex	-0.186	0.273	(-0.745, 0.304)
	Perf.	-0.589	0.408	(-1.256, 0.279)
CP				
Spike & Slab	Age	0.015	0.011	(-0.004, 0.038)
	Treat.	-0.546	0.270	(-1.063, -0.023)
	Sex	-0.166	0.287	(-0.724, 0.389)
	Perf.	-0.502	0.395	(-1.282, 0.251)

⁽¹⁾ CBF.
⁽²⁾ Uniform.
⁽³⁾ Jeffreys.

Table 2 compares the posterior estimates for the regression parameters obtained using the NPP with different initial Beta priors for δ , as well as other dynamic information borrowing methods. These estimates include posterior means, standard deviations, and 95% HPDIs. The NPPs considered include the one based on the CBF criterion described in (8), the standard uniform prior, Jeffreys’ prior, and a mixture of two Beta priors for δ akin to a model averaging approach (Röver et al., 2019; Best et al., 2021; Weru et al., 2024). In addition, we consider the RMAP prior, which places equal weight ω on both the non-informative and informative components derived from historical data; the SAM prior for a binary endpoint; and a fully Bayesian commensurate prior (CP) (Hobbs et al., 2011, 2012), which accounts for uncertainty in the commensurability parameter and employs a spike-and-slab formulation (Mitchell and Beauchamp, 1988), specified as a mixture of two half-normal priors (Hobbs et al., 2012). Notably, the well-balanced prior identified using the CBF procedure – a Beta(5,0.5) – results in consistently smaller posterior standard deviations for the treatment, sex, and performance status parameters compared to the other evaluated priors, indicating more precise inferential conclusions. Furthermore, the HPDIs for all the parameters of interest are narrower under the CBF selected prior, suggesting that, by incorporating substantial historical information, it enhances the precision of the posterior parameter estimation.

Fig. 12 presents a horizontal bar plot that compares the standard deviations of the examined prior distributions with that of the NPP using a uniform prior on δ . Among the evaluated priors, the NPP with the CBF selected prior on δ demonstrates the most significant overall reduction in standard deviation for all parameters of interest.

6. Discussion

The power prior method presents a flexible way to construct an informative prior by combining a prior distribution with the weighted likelihood of previous data. This combined posterior then serves as the prior for new studies. However, determining the

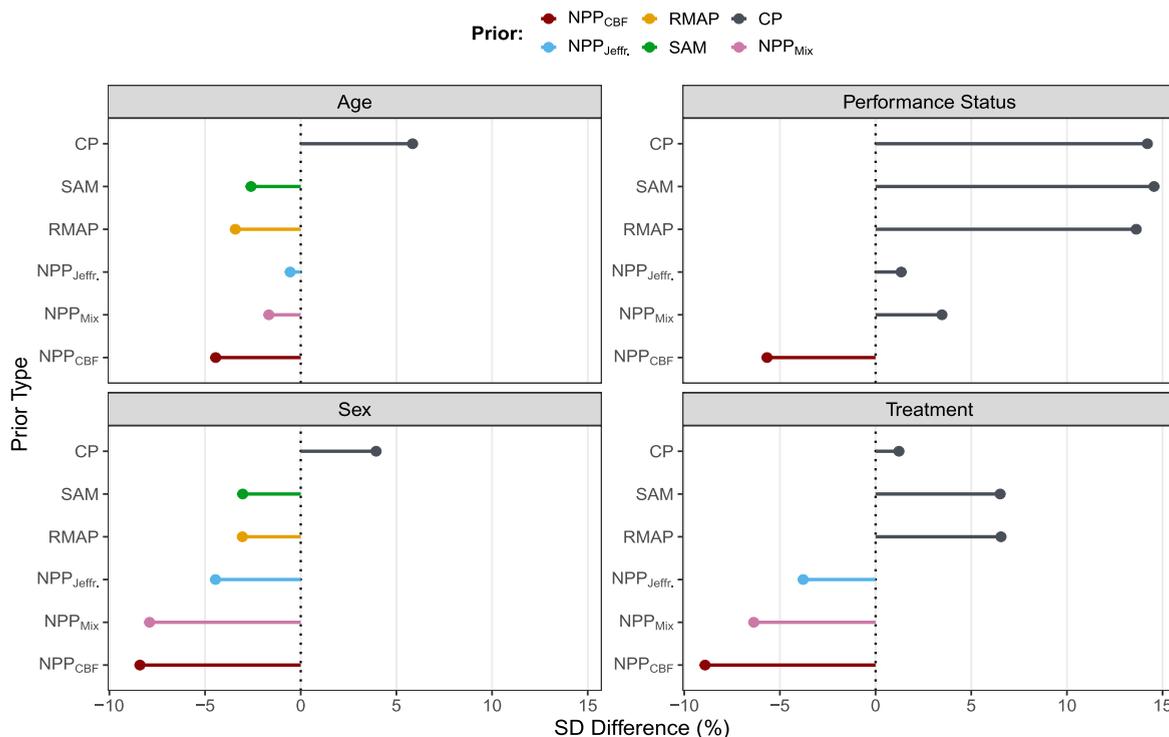


Fig. 12. Melanoma clinical trial. Horizontal barplot for the age, performance status, sex, and treatment indicator representing the difference in percentage of the standard deviation of the evaluated priors in comparison with the normalized power prior with a uniform prior for δ .

appropriate weight parameter presents a significant challenge, whether it is fixed or its prior distribution is being evaluated. Although several methods exist for setting a fixed weight parameter, fully Bayesian approaches for eliciting more informative priors are usually not addressed in the related literature.

Gravestock and Held (2017) highlighted that while the fully Bayesian approach is inherently flexible, it may not effectively address the disagreement between historical and current data. This issue frequently stems from the default use of a non-informative prior, which might not be sensitive enough to detect significant divergences. Consequently, we advocate for the use of a more informative prior specifically designed to detect potential conflict between historical and current datasets, improving the robustness of the resulting statistical inferences.

Our proposed CBF procedure is a novel response to this challenge. It seeks to select a more informative prior than the conventional non-informative one by using hypothetical replications derived from the posterior predictive distribution. The selected prior has minimal influence on the posterior central summary statistics while simultaneously achieving a smaller posterior variance for the parameters of interest. The efficacy of this approach is demonstrated through both simulation studies and the application to melanoma data, proving its robustness and effectiveness in distinguishing between different prior specifications. The ability of this method to select a well-balanced prior based on the agreement between historical and current data, as evidenced in the melanoma study, emphasizes its practical relevance in real-world applications. Furthermore, the flexibility of our proposal suggests that it can be easily extended to other types of endpoint, such as survival outcomes, by adopting, for instance, the cure rate model proposed by Ibrahim et al. (2015).

Our approach requires a higher computational cost to specify the weight parameter distribution, compared to simply assuming a standard uniform prior for δ . This cost depends mainly on the formulation of the model, the number of hypothetical replications used for the log-BF distribution, and the grid search method. However, we believe that this additional effort is justified by more precise inferential conclusions about the parameters of interest. In practice, users can adapt the computational load to the formulation of their models by reducing the number of replications or using parallel computing, while still benefiting from the improved accuracy and robustness offered by the CBF procedure.

A crucial aspect of our CBF procedure is the choice of the HPDI to assess the location of the observed Log-BF within the distribution of replicated Log-BFs. This decision is crucial because it directly influences the interpretation of empirical evidence relative to the modeled hypotheses. A narrower HPDI is recommended when the goal is to limit the range of acceptable values, thereby enhancing the strength and reliability of empirical findings. Future research will focus on developing quantitative methods to determine the appropriate HPDI width.

The methodology presented in this paper offers several areas for potential improvement. Firstly, the selection criteria outlined in (8) could be refined to more effectively identify well-calibrated priors, particularly in cases of moderate agreement between historical and current data. Additionally, it is advisable to consider alternatives to the grid search method used in this study. Instead

of exhaustively exploring all the parameters within the grid, methods that target a relevant subset of the parameter space should be explored. Furthermore, future work should focus on providing a more comprehensive analysis of the theoretical properties of the CBF method.

Finally, a thorough comparison with the optimal prior proposed by Shen et al. (2023) and methods that provide an estimate for δ , possibly in terms of MSE or other measures, is a primary goal for future research.

Software and data availability

All analyses were conducted in the R programming language version 4.4.1 (R Core Team, 2024). The code and data to reproduce this manuscript are openly available at <https://github.com/RoMaD-96/CBFpp>.

Acknowledgements

The authors sincerely thank the Editor, the Associate Editor, and the two anonymous reviewers for constructive and valuable comments that improved the paper substantially.

Appendix A. Poisson log-linear model

Let $y_0 = (y_{0,1}, \dots, y_{0,N_0})$ and $y = (y_1, \dots, y_N)$ be the count outcome of an historical and a current study, respectively. Let denote with $\mathbf{x}_{0,k} = (x_{0,k1}, \dots, x_{0,kp})$, for $k = 1, \dots, N_0$, and $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})$, for $j = 1, \dots, N$, the corresponding covariate vector. The Poisson log-linear model is

$$y_j | \boldsymbol{\beta} \sim \text{Poisson}(\lambda_j), \quad \text{where } \lambda_j = \exp(\mathbf{x}_j^\top \boldsymbol{\beta}),$$

with $\boldsymbol{\beta}$ being the p -dimensional vector of the regression coefficients. Let $\pi_0(\boldsymbol{\beta})$ be an initial multivariate normal prior on $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim N_p(\mathbf{m}, V),$$

where \mathbf{m} is a p -dimensional mean vector and V is a $p \times p$ covariance matrix. The NPP for $(\boldsymbol{\beta}, \delta)$ is then

$$\pi(\boldsymbol{\beta}, \delta | y_0, X_0) = \frac{[L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta})}{\int [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) d\boldsymbol{\beta}} \times \text{Beta}(\delta | \eta, \nu),$$

where $[L(\boldsymbol{\beta} | y_0, X_0)]^\delta = \exp\left(\sum_{k=1}^{N_0} \left[\delta(y_{0,k} \mathbf{x}_{0,k}^\top \boldsymbol{\beta} - \exp(\mathbf{x}_{0,k}^\top \boldsymbol{\beta}))\right]\right) \prod_{k=1}^{N_0} (y_{0,k}!)^{-\delta}$. Given current data, the posterior distribution is

$$\begin{aligned} \pi(\boldsymbol{\beta}, \delta | y, y_0, X, X_0) &= \frac{L(\boldsymbol{\beta} | y, X) \pi(\boldsymbol{\beta}, \delta | y_0, X_0)}{\int_0^1 \int L(\boldsymbol{\beta} | y, X) \pi(\boldsymbol{\beta}, \delta | y_0, X_0) d\boldsymbol{\beta} d\delta} \\ &= \frac{L(\boldsymbol{\beta} | y, X) [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) \text{Beta}(\delta | \eta, \nu)}{\int_0^1 \int L(\boldsymbol{\beta} | y, X) [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) \text{Beta}(\delta | \eta, \nu) d\boldsymbol{\beta} d\delta}, \end{aligned}$$

where $L(\boldsymbol{\beta} | y, X) = \exp\left(\sum_{j=1}^N \left[(y_j \mathbf{x}_j^\top \boldsymbol{\beta} - \exp(\mathbf{x}_j^\top \boldsymbol{\beta}))\right]\right) \prod_{j=1}^N y_j!$. Consequently, the BF is

$$\text{BF}_{0,i}(y) = \frac{\int_0^1 \int L(\boldsymbol{\beta} | y, X) [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) \text{Beta}(\delta | \eta_i, \nu_i) d\boldsymbol{\beta} d\delta}{\int_0^1 \int L(\boldsymbol{\beta} | y, X) [L(\boldsymbol{\beta} | y_0, X_0)]^\delta \pi_0(\boldsymbol{\beta}) \text{Beta}(\delta | 1, 1) d\boldsymbol{\beta} d\delta}.$$

Appendix B. Binomial with unknown success probability θ

Let N_0 and N denote the number of Bernoulli trials in the historical and current studies, respectively. The terms y_0 and y represent the successes in these studies. Assuming a binomial likelihood with a success probability θ for each study and an initial Beta prior for both θ and the weight parameter δ , the NPP is

$$\begin{aligned} \pi(\theta, \delta | y_0) &= \frac{[\text{Bin}(y_0 | \theta, N_0)]^\delta \text{Beta}(\theta | p, q)}{\int_0^1 [\text{Bin}(y_0 | \theta, N_0)]^\delta \text{Beta}(\theta | p, q) d\theta} \times \text{Beta}(\delta | \eta, \nu) \\ &= \text{Beta}(\theta | \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | \eta, \nu). \end{aligned}$$

In light of the current data, the posterior distribution is

$$\begin{aligned} \pi(\theta, \delta | y, y_0) &= \frac{L(\theta | y) \pi(\theta, \delta | y_0)}{\int_0^1 \int_0^1 L(\theta | y) \pi(\theta, \delta | y_0) d\theta d\delta} \\ &= \frac{\text{Bin}(y | \theta, N) \text{Beta}(\theta | \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | \eta, \nu)}{\int_0^1 \text{BBin}(y | N, \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | \eta, \nu) d\delta}, \end{aligned}$$

where $\text{BBin}(\cdot | N, \alpha, \beta)$ is the beta-binomial discrete distribution. Therefore, the BF is

$$\text{BF}_{0,i}(y) = \frac{\int_0^1 \text{BBin}(y | N, \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | \eta_i, \nu_i) d\delta}{\int_0^1 \text{BBin}(y | N, \delta y_0 + p, \delta(N_0 - y_0) + q) \text{Beta}(\delta | 1, 1) d\delta}, \quad i = 1, \dots, M.$$

Appendix C. Gaussian with unknown mean μ

Let μ be the unknown true effect size, with $\hat{\mu}_s$ representing the estimated effect size from study s , where $s \in \{o, r\}$ denotes “original” and “replication” studies, respectively. Furthermore, we assume that the effect size estimates are normally distributed.

$$\hat{\mu}_s | \mu \sim N(\mu, \sigma_s^2),$$

where σ_s^2 represents the variance of the estimated effect size $\hat{\mu}_s$, assumed to be known. Let consider an initial improper prior for the effect size parameter $\pi_0(\mu) \propto 1$ and a Beta prior for the weight parameter δ then the NPP is

$$\pi(\mu, \delta | y_o) = N\left(\mu | \hat{\mu}_o, \frac{\sigma_o^2}{\delta}\right) \text{Beta}(\delta | \eta, \nu).$$

Updating the previous prior with the likelihood of the replicated data yields the following posterior distribution

$$\begin{aligned} \pi(\mu, \delta | y, y_o) &= \frac{L(\mu | y) \pi(\mu, \delta | y_o)}{\int_0^1 \int_{-\infty}^{\infty} L(\mu | y) \pi(\mu, \delta | y_o) d\mu d\delta} \\ &= \frac{N(\hat{\mu} | \mu, \sigma^2) N(\mu | \hat{\mu}_o, \sigma_o^2 / \delta) \text{Beta}(\delta | \eta, \nu)}{\int_0^1 N(\hat{\mu} | \hat{\mu}_o, \sigma^2 + \sigma_o^2 / \delta) \text{Beta}(\delta | \eta, \nu) d\delta}. \end{aligned}$$

Furthermore, the BF is

$$\text{BF}_{0,i}(y) = \frac{\int_0^1 N(\hat{\mu} | \hat{\mu}_o, \sigma^2 + \sigma_o^2 / \delta) \text{Beta}(\delta | \eta_i, \nu_i) d\delta}{\int_0^1 N(\hat{\mu} | \hat{\mu}_o, \sigma^2 + \sigma_o^2 / \delta) \text{Beta}(\delta | 1, 1) d\delta}, \quad i = 1, \dots, M.$$

References

- Alt, E.M., Chang, X., Jiang, X., Liu, Q., Mo, M., Xia, H.A., Ibrahim, J.G., 2024a. Leap: the latent exchangeability prior for borrowing information from historical data. *Biometrics* 80 (3), ujae083.
- Alt, E.M., Chen, X., Carvalho, L.M., Ibrahim, J.G., 2024b. hdbayes: Bayesian Analysis of Generalized Linear Models with Historical Data. R package version 0.1.1.9000, commit c7d330bdd3caf7dc0da303d1cdd1f17abd0a48d6.
- Bennett, M., White, S., Best, N., Mander, A., 2021. A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: a comparison to standard methods. *Pharm. Stat.* 20 (3), 462–484.
- Best, N., Price, R.G., Pouliquen, I.J., Keene, O.N., 2021. Assessing efficacy in important subgroups in confirmatory trials: an example using Bayesian dynamic borrowing. *Pharm. Stat.* 20 (3), 551–562.
- Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. *J. Stat. Softw.* 76 (1), 1–32.
- Carvalho, L.M., Ibrahim, J.G., 2021. On the normalized power prior. *Stat. Med.* 40 (24), 5251–5275.
- Chen, M.-H., Ibrahim, J.G., 2000. Power prior distributions for regression models. *Stat. Sci.* 15 (1), 46–60.
- Chen, M.-H., Ibrahim, J.G., 2006. The relationship between the power prior and hierarchical models. *Bayesian Anal.* 1 (3), 551–574.
- Chen, M.-H., Ibrahim, J.G., Shao, Q.-M., 2000. Power prior distributions for generalized linear models. *J. Stat. Plan. Inference* 84 (1–2), 121–137.
- Cook, S., Gelman, A., Rubin, D.B., 2006. Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.* 15 (3), 675–692.
- De Santis, F., 2006. Power priors and their use in clinical trials. *Am. Stat.* 60 (2), 122–129.
- Dickey, J.M., Lientz, B.P., 1970. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Stat.* 41 (1), 214–226.
- Duan, Y., Ye, K., Smith, E.P., 2006. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 17 (1), 95–106.
- Egidi, L., Pauli, F., Torelli, N., 2022. Avoiding prior–data conflict in regression models via mixture priors. *Can. J. Stat.* 50 (2), 491–510.
- Evans, M., Moshonov, H., 2006. Checking for prior–data conflict. *Bayesian Anal.* 1 (4), 893–914.
- Garcia-Donato, G., Chen, M.-H., 2005. Calibrating Bayes factor under prior predictive distributions. *Stat. Sin.* 15 (2), 359–380.
- Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* 2 (4), 1360–1383.
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., Modrák, M., 2020. Bayesian workflow. arXiv preprint. arXiv:2011.01808.
- Geweke, J., 2004. Getting it right: joint distribution tests of posterior simulators. *J. Am. Stat. Assoc.* 99 (467), 799–804.
- Gravestock, I., Held, L., 2017. Adaptive power priors with empirical Bayes for clinical trials. *Pharm. Stat.* 16 (5), 349–360.
- Gravestock, I., Held, L., 2019. Power priors based on multiple historical studies for binary outcomes. *Biom. J.* 61 (5), 1201–1218.
- Hobbs, B.P., Carlin, B.P., Mandrekar, S.J., Sargent, D.J., 2011. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 67 (3), 1047–1056.
- 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 Anal.* 7 (3), 639–674.
- Ibrahim, J.G., Chen, M.-H., Gwon, Y., Chen, F., 2015. The power prior: theory and applications. *Stat. Med.* 34 (28), 3724–3749.
- Ibrahim, J.G., Chen, M.-H., Sinha, D., 2003. On optimality properties of the power prior. *J. Am. Stat. Assoc.* 98 (461), 204–213.
- Jeffreys, H., 1961. *The Theory of Probability*. Oxford University Press.
- Kass, R.E., 1993. Bayes factors in practice. *J. R. Stat. Soc., Ser. D, Stat.* 42 (5), 551–560.

- Kass, R.E., Raftery, A.E., 1995. Bayes factors. *J. Am. Stat. Assoc.* 90 (430), 773–795.
- Kirkwood, J.M., Ibrahim, J., Lawson, D.H., Atkins, M.B., Agarwala, S.S., Collins, K., Mascari, R., Morrissey, D.M., Chapman, P.B., 2001a. High-dose interferon alfa-2b does not diminish antibody response to GM2 vaccination in patients with resected melanoma: results of the multicenter eastern cooperative oncology group phase II trial E2696. *J. Clin. Oncol.* 19 (5), 1430–1436.
- Kirkwood, J.M., Ibrahim, J.G., Sosman, J.A., Sondak, V.K., Agarwala, S.S., Ernstoff, M.S., Rao, U., 2001b. High-dose interferon alfa-2b significantly prolongs relapse-free and overall survival compared with the GM2-KLH/QS-21 vaccine in patients with resected stage IIB-III melanoma: results of intergroup trial E1694/S9512/C509801. *J. Clin. Oncol.* 19 (9), 2370–2380.
- Lee, M.D., Wagenmakers, E.-J., 2014. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.
- Liu, G.F., 2018. A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. *Pharm. Stat.* 17 (1), 61–73.
- Mariani, F., De Santis, F., Gubbiotti, S., 2024. A dynamic power prior approach to non-inferiority trials for normal means. *Pharm. Stat.* 23 (2), 242–256.
- Meng, X.-L., Wong, W.H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.*, 831–860.
- Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* 83 (404), 1023–1032.
- Modrák, M., Moon, A.H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., Vehtari, A., 2023. Simulation-based calibration checking for Bayesian computation: the choice of test quantities shapes sensitivity. *Bayesian Anal.* 1 (1), 1–28.
- Neuenschwander, B., Branson, M., Spiegelhalter, D.J., 2009. A note on the power prior. *Stat. Med.* 28 (28), 3562–3566.
- Neuenschwander, B., Capkun-Niggli, G., Branson, M., Spiegelhalter, D.J., 2010. Summarizing historical information on controls in clinical trials. *Clin. Trials* 7 (1), 5–18.
- Neuenschwander, B., Schmidli, H., 2020. Use of historical data. In: *Bayesian Methods in Pharmaceutical Research*, 1st edition. Chapman and Hall/CRC, p. 27.
- Nikolakopoulos, S., van der Tweel, I., Roes, K.C.B., 2018. Dynamic borrowing through empirical power priors that control type I error. *Biometrics* 74 (3), 874–880.
- Ollier, A., Morita, S., Ursino, M., Zohar, S., 2020. An adaptive power prior for sequential clinical trials—application to bridging studies. *Stat. Methods Med. Res.* 29 (8), 2282–2294. PMID: 31729275.
- Pawel, S., Aust, F., Held, L., Wagenmakers, E.-J., 2024. Power priors for replication studies. *Test* 33 (1), 127–154.
- R Core Team, 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Robert, C.P., 2022. 50 shades of Bayesian testing of hypotheses. In: Srinivasa Rao, A.S., Young, G.A., Rao, C. (Eds.), *Advancements in Bayesian Methods and Implementation*. In: *Handbook of Statistics*, vol. 47. Elsevier, pp. 103–120.
- Robert, C.P., Casella, G., 2004. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York.
- Röver, C., Wandel, S., Friede, T., 2019. Model averaging for robust extrapolation in evidence synthesis. *Stat. Med.* 38 (4), 674–694.
- Schad, D.J., Betancourt, M., Vasishth, S., 2021. Toward a principled Bayesian workflow in cognitive science. *Psychol. Methods* 26 (1), 103–126.
- Schad, D.J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., Vasishth, S., 2023. Workflow techniques for the robust use of Bayes factors. *Psychol. Methods* 28 (6), 1404–1426.
- Schmidli, 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.
- Shen, Y., Carvalho, L.M., Psioda, M.A., Ibrahim, J.G., 2023. Optimal priors for the discounting parameter of the normalized power prior. *arXiv preprint*. arXiv: 2302.14230.
- Spiegelhalter, D.J., Abrams, K.R., Myles, J.P., 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., Gelman, A., 2018. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint*. arXiv:1804.06788.
- Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J.G., Kinnorsley, N., Lindborg, S., et al., 2014. Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Stat.* 13 (1), 41–54.
- Vlachos, P.K., Gelfand, A.E., 2003. On the calibration of Bayesian model choice criteria. In: *Special issue I: Model Selection, Model Diagnostics, Empirical Bayes and Hierarchical Bayes*. *J. Stat. Plan. Inference* 111 (1), 223–234.
- Weru, V., Kopp-Schneider, A., Wiesenfarth, M., Weber, S., Calderazzo, S., 2024. Information borrowing in Bayesian clinical trials: choice of tuning parameters for the robust mixture prior. *arXiv preprint*. arXiv:2412.03185.
- Yang, P., Zhao, Y., Nie, L., Vallejo, J., Yuan, Y., 2023. Sam: self-adapting mixture prior to dynamically borrow information from historical data in clinical trials. *Biometrics* 79 (4), 2857–2868.
- Ye, K., Han, Z., Duan, Y., Bai, T., 2022. Normalized power prior Bayesian analysis. *J. Stat. Plan. Inference* 216, 29–50.