Brazilian Journal of Probability and Statistics
2024, Vol. 38, No. 1, 1–31
https://doi.org/10.1214/24-BJPS598

A review of dynamic borrowing methods with applications in pharmaceutical research

Emmanuel Lesaffre[1,a], Hongchao Qi[2,3] Akalu Banbeta[4,5] and Joost van Rosmalen[2,3]

[1]*I-Biostat, KULeuven, Leuven, Belgium,* [a]*Emmanuel.Lesaffre@kuleuven.be*
[2]*Department of Biostatistics, Erasmus MC, Rotterdam, the Netherlands*
[3]*Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands*
[4]*I-Biostat, UHasselt, Hasselt, Belgium*
[5]*Department of Statistics, Jimma University, Jimma, Ethiopia*

Abstract. This non-technical review discusses the use of historical data in the design and analysis of randomized controlled trials using a Bayesian approach. The focus is on comparing the philosophy behind different approaches and practical considerations for their use. The two main approaches, that is, the power prior and the meta-analytic-predictive prior, are illustrated using fictitious and real data sets. Such methods, which are known as dynamic borrowing methods, are becoming increasingly popular in pharmaceutical research because they may imply an important reduction in costs. In some cases, e.g. in pediatric studies, they may be indispensable to address the clinical research question. In addition to the two original approaches, this review also covers various extensions and variations of the methods. The usefulness and acceptance of the approaches by regulatory agencies is also critically evaluated. Finally, references to relevant software are provided.

1 Introduction

The Bayesian approach to statistical inference enjoys the unique feature that it can incorporate external information in an explicit manner into the analysis of collected data. Given a statistical model, external information summarized in a prior distribution can be combined with the observed data using Bayes' theorem to obtain a posterior distribution for the model parameters. In case the prior distribution reflects historical data and/or expert knowledge, the gain in posterior information on the model parameters can be considerable, which is one of the main advantages of the Bayesian approach over the classical frequentist approach. Although the Bayesian framework provides the capability to incorporate valuable external information, this attractive feature often goes unused, and non-informative or vague prior distributions are used instead.

In this review, we look at incorporating external information in pharmaceutical studies. In pharmaceutical research, there is often a variety of previous data available, and due to the highly controlled setting, these so-called historical data are generally quite reliable. Nevertheless such historical data are often neglected, which is unfortunate. Due to regulatory requirements, randomized controlled trials (RCTs) have become increasingly expensive and time consuming in the last decades. Also, the increasing focus on rare diseases and precision medicine has led to more trials with small patient populations, where patient recruitment for a fully powered trial is difficult. Therefore using past data, or better formulated: *borrowing information from historical studies*, has lately received considerable interest in pharmaceutical research, typically in combination with a Bayesian approach to inference. Making explicit use of historical information may not only reduce the necessary sample size of the study

Key words and phrases. Commensurate prior, historical data, meta-analytic-predictive prior, power prior, Pocock's criteria, randomized controlled trials.
Received December 2023; accepted January 2024.

and/or increase the statistical power of the analysis, it may also be essential to render a clinical trial feasible. Incorporating historical information into a current study may be useful in the following settings:

- Often the same control treatment is used to test the efficacy and safety of different experimental treatments. One may then consider to use the historical control information to reduce the sample size of the current control arm, which was first proposed by Pocock (1976).
- The development of new drugs classically follows several development stages, from Phase I studies that verify the safety of the drug or in oncology to find the optimal dose of the drug to Phase IV trials that monitor public safety, the effectiveness of the drug in real life but also potential serious adverse events in the general patient population. Results from a previous phase are used in the design of the subsequent study, but one may also use the historical data in the analysis of the new study.
- For rare diseases, it may be difficult to recruit sufficient numbers of patients to test the performance of an experimental treatment, let alone to organize two independent RCTs with enough power. Then, making use of past data obtained in a similar setting may render the clinical trial feasible, see, for example, Schmidli, Neuenschwander and Friede (2017), Wandel et al. (2017) or Lesaffre, Baio and Boulanger (2020).
- External information can be obtained by extrapolating information collected on one patient group to another patient group. This may happen in, for example, pediatric studies. Because ethical approval is not so easily obtained for RCTs in children, many pediatric studies are (small) studies upon which it is difficult to draw clinical conclusions. It may then be a logical choice to extrapolate the efficacy/safety results from adult studies to the children population, of course depending on the disease and the similarity of the disease progression in children/adolescents and adults. A bridging study is another example of an extrapolation study, where one brings over the results obtained from one geographical area to another.
- In some situations, historical information that was not collected in a highly controlled trial could be considered in the analysis of a new trial. Such data are often referred to as *real-world data*. Relevant examples include data from patient registries and expanded access programs (i.e., patient data collected as part of a program for pre-approval access to a novel drug). Of course real-world data should only be considered if they are relevant and of high quality, and any differences in patient characteristics with the current data should be addressed.
- In medical device studies, RCTs may be hard to organize since devices are subject to rapid technical innovations and their comparative efficacy/safety may be apparent only after many years. Historical controls may then be of help to shed some light on the performance of the experimental device.
- Historical data may also be useful in platform trials, which are trials with an innovative design based on a single master protocol to evaluate in an efficient manner multiple interventions. See Liu et al. (2023) for an example where data from historical controls are combined with the current control data.

To incorporate these sources of historical information in the analysis, the Bayesian approach provides a natural and elegant framework, where the historical information is summarized using an informative prior. The use of historical information is however not without pitfalls or drawbacks. Naturally, the historical information should be chosen carefully, to make sure that it is valid and relevant, and that no considerable differences with the data of the new trial are to be expected. Even then, the analyst should take into account the possibility of a *prior-data conflict*, that is, a difference between the historical information and the data of the new trial, which could lead to a bias in the study results. For this reason, it

is typically not appropriate to assign equal weight to the historical data and the data of the new trial. Various statistical approaches have been developed to protect against the effects of prior-data conflict, typically by downweighting the information from the historical data, leading to a more robust prior. These statistical approaches are in effect different ways to construct the informative prior from the historical information. Of course we need to choose the most appropriate statistical approach (i.e., the prior) for the problem at hand.

This article is structured as follows. We first look in Section 2 which of the historical data sets are eligible to be used in the analysis of the current data. In Section 3, we start our review of the dynamic borrowing methods with the two most popular approaches, that is, the power prior approach and the meta-analytic approach. Then in Section 4 we briefly discuss some alternative approaches. In Section 5, design aspects when using dynamic borrowing methods are discussed. In Section 6, the two discussed dynamic borrowing methods are illustrated using practical data sets. The operating characteristics of the dynamic borrowing methods are critically discussed in Section 7. A short overview of available software packages is given in Section 8. Concluding remarks can be found in Section 9.

2 Choice of historical studies

The (dynamic) borrowing approaches do not automatically account for how and where the historical studies were conducted. Therefore, prior to borrowing information from historical studies, one should check whether the conditions under which the past and current studies were done are similar enough. This was first pointed out by Pocock (1976). He stipulated in a seminal paper the conditions under which a historical control group can be used in the analysis of a current study. These are:

1. The group must have been treated with exactly the same treatment as the randomized controls in the current study.
2. The group must have been part of a recent clinical study with the same requirements for patient eligibility.
3. The methods of treatment evaluation must be the same.
4. The distributions of important patient characteristics in the group should be comparable with those in the new trial.
5. The historical study must have been performed in the same organization with largely the same clinical investigators.
6. There must be no other indications leading one to expect different results between the current and the historical controls.

Simply stated, Pocock argued that the historical controls should be quite similar to the current control patients and done under basically the same conditions. However, these well-known criteria are too strict to be applied in practice because they prevent using dynamic borrowing methods for, e.g., in pediatric studies when data of adults are to be used as input for the study on children; with rare diseases where the historical controls are often taken from the real world and in bridging studies, where the historical data are inevitably taken from subjects living in another geographical region and therefore cannot be taken from the same institution as the current data.

For these reasons, Hatswell et al. (2020) suggested to relax Pocock's conditions. The main relaxation was that Hatswell et al. proposed a framework to present relevant data and study design aspects and to identify and quantify differences, whereas Pocock proposed conditions that should all be satisfied. The idea of this framework was to facilitate a more nuanced assessment of the usefulness of the historical data, so that an appropriate statistical technique may then be selected. The framework also allows for differences in patient characteristics

between data sources, because such differences can be addressed using regression modeling or propensity score methods.

3 Dynamic borrowing methods: The power prior and the meta-analytic predictive prior

A variety of borrowing methods has been suggested in the statistical literature, and not only in the context of RCTs. We will focus mainly on the power prior approach and the meta-analytic prior predictive approach. Both approaches have been explored extensively in the last two decades. Although combining past with current data can be done in any context, here we will take only illustrations from pharmaceutical research. Two settings emerge with respect to borrowing information from the past: (1) control data or (2) contrast data measuring the efficacy of the experimental arm versus the control arm. We focus here on the first case.

The two extreme cases of borrowing information from past studies is either not using the historical information at all (analyzing just the current data) or completely pooling the historical with the current data. Ignoring historical information may be the optimal strategy, especially when the historical and current data tell a different story. Pooling the past data with the current data implicitly assumes that the historical studies and the current study measure exactly the same effect and that the patients from the two studies are exchangeable. However, completely pooling is not recommended in general, because exchangeability of the two sets of patients is often a too strong assumption. Nevertheless, in a review on the use of historical data, Wadsworth, Hampson and Jaki (2018) found that 8 of the 58 selected papers completely pooled historical and current data.

The parameter of interest, denoted as θ, can basically represent anything and can be univariate or multivariate. This parameter can also be part of a vector of parameters that includes nuisance parameters such as covariate effects and variance parameters. For instance, when θ represents the effect of an intervention (e.g., an odds ratio or a relative risk or their logarithm), it is combined with a parameter expressing the baseline rate. Most of the developments on dynamic borrowing methods have focused on the case of a univariate θ, as we will also do here. In this section, we discuss the power prior approach and the meta-analytic predictive prior approach. In Section 4, we review a number of alternative methods. We start with the oldest formal approach, namely the power prior. Especially Ibrahim and Chen explored the properties of two of the three power prior approaches and illustrated their use in a great variety of clinical applications (Ibrahim and Chen, 2000, 2015).

3.1 The power prior approach

The power prior approach was first developed for the case of a single historical study that can complement current data. While the approach assumes that the historical and current study measure the same effect, the historical data are downweighted when combined with the current data. We first consider three versions of the power prior approach: (1) the conditional power approach; (2) the joint power prior approach and (3) the normalized or modified power prior approach. Then we look at variations of the basic versions.

3.1.1 *The conditional power prior approach.* Assume a single historical study set up to estimate parameter θ_0, say the true proportion of subjects that show a clinical benefit when treated. Assume that the study produced a sample $\mathcal{D}_0$ of size n_0. Given a statistical model, the likelihood $L(\theta_0|\mathcal{D}_0)$ is produced. Assume also that the current study produces a sample $\mathcal{D}$ of size n to estimate parameter θ.

For the power prior approach, one assumes that $\theta_0 = \theta$. To construct the *conditional power prior* $p_{CPP}(\theta|\mathcal{D}_0, \lambda)$ one combines an initial, typically vague, prior $p_0(\theta)$ for the historical data with the historical likelihood raised to the power λ ($0 \leq \lambda \leq 1$), that is,

$$p_{CPP}(\theta|\mathcal{D}_0, \lambda) = \frac{L(\theta|\mathcal{D}_0)^\lambda p_0(\theta)}{\int L(\theta|\mathcal{D}_0)^\lambda p_0(\theta)\, d\theta}. \tag{1}$$

In the conditional power prior (CPP) approach the power parameter is chosen by the investigator. For $\lambda = 0$ no borrowing is done, while for $\lambda = 1$ the historical and current data are pooled to draw conclusions about θ. A value of λ between 0 and 1 implies downweighting the historical data with a factor. The assumption that the historical and current study estimate the same parameter may look too stringent, but this is compensated for by discounting the historical information. The posterior based on the conditional power prior follows from Bayes' theorem:

$$\begin{aligned} p_{CPP}(\theta|\mathcal{D}, \mathcal{D}_0, \lambda) &\propto L(\theta|\mathcal{D})\, p_{CPP}(\theta|\mathcal{D}_0, \lambda), \\ &\propto L(\theta|\mathcal{D}) L(\theta|\mathcal{D}_0)^\lambda p_0(\theta). \end{aligned} \tag{2}$$

When λ is established without looking at the current data, the conditional power prior is an example of borrowing information in a static way. The conditional power approach has interesting theoretical properties (Ibrahim and Chen, 2015). We now illustrate the conditional power prior for the Gaussian and binomial case.

Gaussian case. Suppose the current data are given by $\mathcal{D} = \{y_1, \ldots, y_n\}$ with elements independently distributed according to a Gaussian distribution with expectation μ and known variance σ^2, that is, $\mathrm{N}(\mu, \sigma^2)$. Suppose also that a historical sample $\mathcal{D}_0 = \{y_{01}, \ldots, y_{0,n_0}\}$ is given with elements that are independently distributed according to $\mathrm{N}(\mu, \sigma^2)$. For the construction of the power prior, an initial Gaussian prior for μ is taken, that is, $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$.

The power prior is based on the Gaussian likelihood raised to the power λ, that is, $L(\mu|\mathcal{D}_0)^\lambda = (\frac{1}{\sqrt{2\pi\sigma^2}})^\lambda \exp[-\frac{1}{2\sigma^2/\lambda} \sum_{i=1}^{n_0} (y_{0i} - \mu)^2]$. Since λ is fixed, $L(\mu|\mathcal{D}_0)^\lambda$ is in fact proportional to a Gaussian likelihood with variance σ^2/λ. In other words, the power prior for the Gaussian case inflates the prior variance by factor $1/\lambda$. With the initial Gaussian prior for μ, the power prior for the current data becomes $\mathrm{N}(\mu|\overline{\mu}_0, \overline{\sigma}_0^2)$, with

$$\overline{\mu}_0 = \frac{\mu_0/\sigma_0^2 + n_0\lambda\overline{y}_0/\sigma^2}{1/\sigma_0^2 + n_0\lambda/\sigma^2} \quad \& \quad 1/\overline{\sigma}_0^2 = 1/\sigma_0^2 + n_0\lambda/\sigma^2. \tag{3}$$

The posterior obtained from combining this prior with the current Gaussian data is then given by $\mathrm{N}(\mu|\overline{\mu}, \overline{\sigma}^2)$, with

$$\overline{\mu} = \frac{\overline{\mu}_0/\overline{\sigma}_0^2 + n\overline{y}/\sigma^2}{1/\overline{\sigma}_0^2 + n/\sigma^2} \quad \& \quad 1/\overline{\sigma}^2 = 1/\overline{\sigma}_0^2 + n/\sigma^2. \tag{4}$$

Binomial case. In this case, the current data consist of the number of 'successes' y out of a sample of size n. We assume that the current data are given by the sample $\mathcal{D} : y \sim \mathrm{Bin}(n, \theta)$ and similarly that the historical data are given by the sample $\mathcal{D}_0 : y_0 \sim \mathrm{Bin}(n_0, \theta)$. When combined with an initial beta prior $\theta \sim \mathrm{Beta}(\alpha_0, \beta_0)$, the binomial power prior is a beta distribution given by

$$\mathrm{Beta}\big(\theta|\lambda y_0 + \alpha_0, \lambda(n_0 - y_0) + \beta_0\big),$$

and when combined with the current data results in the beta posterior

$$\mathrm{Beta}\big(\theta|\lambda y_0 + \alpha_0 + y, \lambda(n_0 - y_0) + \beta_0 + (n - y)\big).$$

From the above two examples we can conclude that λ represents the proportion of historical data used in the current study, that is, $\lambda = r/n_0$, with r the amount of historical sample used. Thus in these models, λ is easy to interpret. De Santis (2006) showed that the above interpretation of λ holds more generally. In practice, though, it may not be easy to choose an appropriate value for λ.

Alternatively, one could estimate λ from the data, depending on the discrepancy between the historical and the current data. This leads to the joint and the normalized power priors.

3.1.2 *Two dynamic power prior approaches*. In this section, we look at the (joint and modified) power prior approaches. They are called dynamic borrowing methods because the amount of borrowing is determined by contrasting the historical data with the current data. Hence, the amount of borrowing can only be determined after having seen the current data.

A Bayesian way to deal with the uncertainty of a parameter is to give it a prior. For the joint power prior both the parameter of interest θ but also the power parameter λ is given a prior distribution, say a prior $p_0(\theta)$ for θ and a Beta(1,1) or Beta(0.5,0.5) prior for λ. Thus with a prior $p(\lambda)$, the *joint power prior* is given by

$$p_{JPP}(\theta, \lambda|\mathcal{D}_0) = \frac{L(\theta|\mathcal{D}_0)^\lambda p_0(\theta) p(\lambda)}{\int_0^1 \int_\Theta L(\theta|\mathcal{D}_0)^\lambda p_0(\theta) p(\lambda)\, d\theta\, d\lambda}. \tag{5}$$

Θ refers to the support of θ. The parameters θ and λ are then estimated from the joint posterior of historical and current data.

The idea is that the estimated λ will be close to 0 for discrepant historical and current data and close to 1 when they are similar. However, the problem with $p_{JPP}(\theta, \lambda|\mathcal{D}_0)$ is that it does not satisfy the *Likelihood Principle*, which in the context of the Bayesian paradigm states that proportional likelihoods should produce the same posterior distribution. That this property does not hold for p_{JPP} can easily be seen by multiplying the original likelihood of the historical data by a constant c, then p_{JPP} will involve $c^\lambda L(\theta|\mathcal{D}_0)^\lambda$, and c^λ does not cancel out from the joint power prior nor from its posterior. The reason for this behavior is that $L(\theta|\mathcal{D}_0)^\lambda$ is not the likelihood function of θ and λ given the historical data. The Likelihood Principle is one of the basic principles in likelihood theory and hence also in the Bayesian approach, strongly advocated by Berger and Wolpert (1988). Hence, it is argued that the principle should hold also here. In addition, simulations showed that p_{JPP} tends to shrink the power to 0 and hence there will be little or even no borrowing regardless of heterogeneity between historical and current data (Neelon and O'Malley, 2010).

The *modified or normalized power prior (MPP)* is a slight but important modification of the joint power prior to make sure that it satisfies the Likelihood Principle. The MPP is given by

$$\begin{aligned} p_{MPP}(\theta, \lambda|\mathcal{D}_0) &= p_{CPP}(\theta|\mathcal{D}_0, \lambda) p(\lambda), \\ &= \frac{L(\theta|\mathcal{D}_0)^\lambda p_0(\theta)}{\int_\Theta L(\theta|\mathcal{D}_0)^\lambda p_0(\theta)\, d\theta} p(\lambda). \end{aligned} \tag{6}$$

The MPP has the disadvantage that it involves an integral in the denominator. In some cases, as will be seen below, an analytical expression of that integral can be derived but most often this is not possible, impacting the computations considerably, especially when one uses a Markov chain Monte Carlo approach to estimate the parameters. Below we illustrate these two power priors for the binomial case. For the Gaussian case, the MPP is given by

$\mathrm{N}(\mu|\overline{\mu}_0(\lambda), \overline{\sigma}_0^2(\lambda)) p(\lambda)$, where the first term is given by equation (3) but now the mean and standard deviation of the power prior depend on the unknown parameter λ.

Binomial case. The joint power prior for the binomial case $p_{JPP}(\theta, \lambda|y_0, n_0)$ is given by

$$p_{JPP}(\theta, \lambda|y_0, n_0) \propto \frac{\binom{n_0}{y_0}^{\lambda} \theta^{\lambda y_0+\alpha_0-1}(1-\theta)^{\lambda(n_0-y_0)+\beta_0-1}}{B(\alpha_0, \beta_0)} p(\lambda),$$

where $B(\bullet, \bullet)$ is the Beta function. The MPP is given by

$$\begin{aligned} p_{MPP}(\theta, \lambda|y_0, n_0) &= \frac{\theta^{\lambda y_0+\alpha_0-1}(1-\theta)^{\lambda(n_0-y_0)+\beta_0-1}}{B(\lambda y_0+\alpha_0, \lambda(n_0-y_0)+\beta_0)} p(\lambda), \\ &= \text{Beta}\big(\theta|\lambda y_0+\alpha_0, \lambda(n_0-y_0)+\beta_0\big) p(\lambda), \end{aligned}$$

whereby the denominator in equation (6), given by

$$C(\lambda) = \int_0^1 L(\theta|y_0, n_0)^{\lambda} \text{Beta}(\theta|\alpha_0, \beta_0)\, d\theta = \frac{\binom{n_0}{y_0}^{\lambda} B(\lambda y_0+\alpha_0, \lambda(n_0-y_0)+\beta_0)}{B(\alpha_0, \beta_0)},$$

can be computed analytically. The posterior distribution when y events were recorded out of n subjects and combined with the MPP is then proportional to

$$\text{Beta}\big(\theta|\lambda y_0+\alpha_0+y, \lambda(n_0-y_0)+\beta_0+(n-y)\big) p(\lambda). \tag{7}$$

An illustrative example. We now illustrate the performance of the MPP for the Gaussian case using simulated data; the binary case resulted in similar conclusions. In both cases, we evaluated the MPP when the historical and the current control data were taken from the same population (congruent case) and when they were taken from different populations (incongruent case). The MPP for the Gaussian case is proportional to

$$\text{N}\Big(\mu|\frac{\overline{\mu}_0(\lambda)/\overline{\sigma}_0^2(\lambda)+n\overline{y}/\sigma^2}{1/\overline{\sigma}_0^2(\lambda)+n/\sigma^2}, \frac{1}{1/\overline{\sigma}_0^2(\lambda)+n/\sigma^2}\Big) p(\lambda).$$

When $\sigma_0 \to \infty$ and with $\mathcal{D}_0$, $\mathcal{D}$ represented by $\overline{y}_0$ and $\overline{y}$, respectively, the posterior for μ is given by

$$p(\mu|\overline{y}, \overline{y}_0) \propto \int_0^1 N\Big(\mu|\overline{y}_0, \frac{\sigma^2}{n_0\lambda}\Big)\, d\lambda L(\mu|\overline{y}).$$

The integral part of the right-hand side of the previous expression represents the actual prior which is concentrated around $\overline{y}_0$. It is a scale mixture of normals and therefore 'looks' like a t-distribution. So, if $\overline{y}_0$ and $\overline{y}$ are distant there is only minimal overlap between the prior and the likelihood and borrowing will be limited.

A simulation study illustrates these findings with the following settings:

- Historical data $y_{hi}(i = 1, \ldots, n)$ were simulated from $\text{N}(\mu_h, 2^2)$. The current control data $y_{ci}(i = 1, \ldots, n)$ were simulated from $\text{N}(\mu_c, 2^2)$ and n current experimental data from $\text{N}(\mu_e, 2^2)$;
- Two sample sizes were considered: (i) $n = 100$ and (ii) $n = 1000$ and one data set was generated for each sample size;
- Two settings were considered for the historical and current control population: (i) congruent case, i.e. same populations with $\mu_h = \mu_c = 10$ and (ii) incongruent case, i.e. dissimilar populations with $\mu_h = 9$, $\mu_c = 10$. In both cases, $\mu_e = 12$;
- A vague prior $\text{N}(0, 1000^2)$ was taken for μ_h and μ_e, whereas for μ_c the same vague prior or the MPP, based on the historical data, was taken;

- Two priors for λ were considered: Beta(1,1) (uniform prior) and Beta(0.5,0.5) (Jeffreys prior).

A BUGS program was written to obtain the posterior estimates of μ_h, μ_c, $\Delta = \mu_e - \mu_c$ and λ. Another BUGS program estimates Δ without borrowing information from the historical data. The relative efficiency of dynamic borrowing is measured by the relative efficiency equal to releff $= 100\frac{(\text{var}_{without} - \text{var}_{with})}{\text{var}_{without}}$, where $\text{var}_{without}$, var_{with} is the posterior variance for Δ without and with borrowing information from the historical controls, respectively. For each program, three chains were initiated each with length 10,000 and a burn-in size of 1000, yielding 27,000 samples.

Regarding estimating Δ, the following results were obtained:

- The posterior means were (evidently) closer to the true value for $n = 1000$ than for $n = 100$, both for the congruent as for the incongruent case and both priors for λ.
- The precision also (again evidently) increased with the sample size.
- The relative efficiency for the congruent case was around 13% for $n = 100$, and around 7% for $n = 1000$ for both λ priors. These values decreased for the incongruent case and $n = 100$ to 6% (Beta(1,1)) and to 2% (Beta(0.5,0.5)), but was for both priors around -2% for $n = 1000$.

Clearly this simulation study was too small to draw definitive conclusions, but it can already be inferred that borrowing information can increase the precision of the estimated treatment effect, but the greatest gain is to be expected in small samples which is exactly the scenario where we wish to use dynamic borrowing.

We return to the results of this simulation study for estimating the power parameter in Section 3.2.

Note that a simple BUGS program could be written because we assumed that σ is known. When σ is estimated from the data, the results in Banbeta, Lesaffre and van Rosmalen (2022) can be used to derive the marginal posterior of λ. We will now consider further extensions of, primarily, the modified power prior.

3.1.3 *The power prior approach for multiple historical studies.* The power prior has been extended to multiple historical studies by Chen, Ibrahim and Shao (2000). Suppose that there are K historical studies with data $\mathcal{D}_k$ of size n_k estimating parameters θ_k, each yielding a likelihood $L(\theta_k|\mathcal{D}_k)$. Suppose also that the current study yields a data set $\mathcal{D}$ of size n to estimate parameter θ, yielding a likelihood $L(\theta|\mathcal{D})$. Again it is assumed that the parameters of the historical studies and the current study are equal, i.e. $\theta_1 = \cdots = \theta_K = \theta$. With a(*n* often) vague prior for the historical data $p_0(\theta)$, the MPP is given by

$$p_{MPP}(\theta, \boldsymbol{\lambda}|\mathcal{D}_k) \propto \frac{[\prod_{k=1}^{K} L(\theta|\mathcal{D}_k)^{\lambda_k}]p_0(\theta)\,p(\boldsymbol{\lambda})}{\int_\Theta [\prod_{k=1}^{K} L(\theta|\mathcal{D}_k)^{\lambda_k}]p_0(\theta)\,d\theta},$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^\top$. Further, $\lambda_k = 0$ implies no borrowing from the k^{th} historical study, and $\lambda_k = 1$ means that the k^{th} historical data is pooled with the current data.

Inspired by the MAP approach discussed in Section 3.3, Banbeta et al. suggested a hierarchical version of the power prior for multiple historical studies and illustrated this for the binomial model and for linear regression (Banbeta et al., 2019; Banbeta, Lesaffre and van Rosmalen, 2022). This version of the power prior, also called the *dependent modified power prior* and denoted by p_{DMPP}, is based on the idea that historical studies are similar and there-

fore also the powers should be similar. Hence, it is assumed that the powers λ_k $(k = 1, \ldots, K)$ have the following hierarchical distribution:

$$\lambda_k \sim \text{Beta}(\alpha_\lambda, \beta_\lambda) \quad (k = 1, \ldots, K),$$
$$(\alpha_\lambda, \beta_\lambda) \sim p(\alpha_\lambda, \beta_\lambda),$$

with $p(\alpha_\lambda, \beta_\lambda)$ a hyperprior.

Binomial case. The prior $p(\boldsymbol{\lambda})$ can be an independent prior or a dependent prior as in Banbeta et al. and $p_0(\theta) = \text{Beta}(\alpha_0, \beta_0)$ with α_0 and β_0 fixed and known. The MPP for multiple historical binomial studies then becomes

$$p_{MPP}(\theta, \boldsymbol{\lambda} | \{y_k, n_k\}) = \frac{\theta^{\sum_{k=1}^K \lambda_k y_k + \alpha_0 - 1} (1-\theta)^{\sum_{k=1}^K \lambda_k (n_k - y_k) + \beta_0 - 1}}{B(\sum_{k=1}^K \lambda_k y_k + \alpha_0, \sum_{k=1}^K \lambda_k (n_k - y_k) + \beta_0)} p(\boldsymbol{\lambda}),$$
$$= \text{Beta}\left(\theta \Big| \sum_{k=1}^K \lambda_k y_k + \alpha_0, \sum_{k=1}^K \lambda_k (n_k - y_k) + \beta_0\right) p(\boldsymbol{\lambda}).$$

While the calculation of the MPP is rather straightforward for the binomial case, the determination of the denominator may not be easy in general. A method based on path sampling has been applied to calculate the integral in the denominator with a single historical control (van Rosmalen et al., 2018). The algorithm can be adapted to the case of multiple historical studies, but it is probably too computationally intensive for three or more data sets.

3.2 The choice of the power parameter

For the MPP the power parameter λ is given a prior and it is hoped that contrasting the historical with the current data will drive its posterior value to 0 for incongruent data and to a value close to 1 for congruent data. In Figure 1, we show the marginal posterior distributions for λ in the eight scenarios we discussed in the illustrative example of Section 3.1.2. There is a clear dependence of the posterior on the sample size and (in)congruency of the setting. But

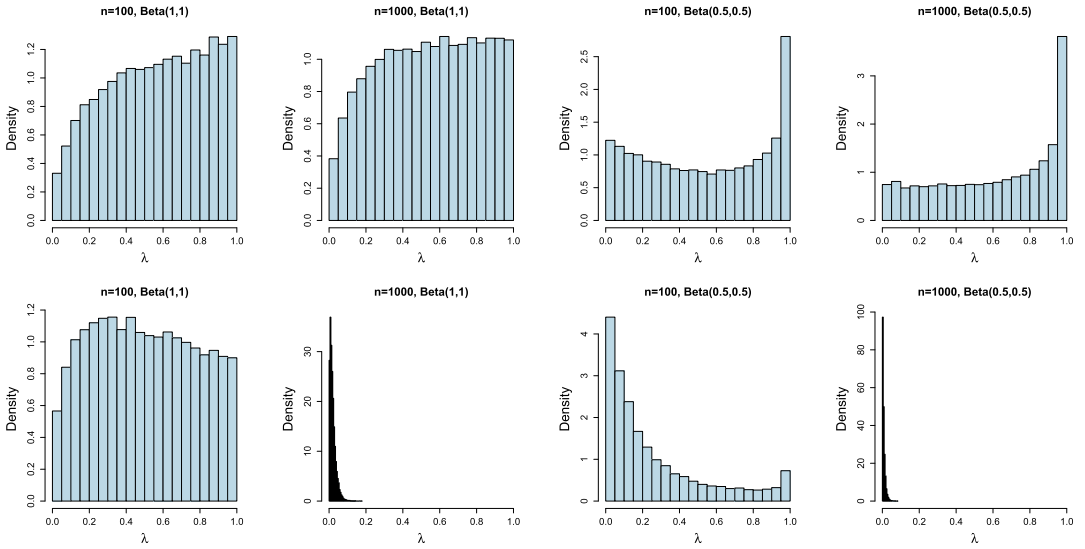

Figure 1 *Marginal posterior of λ in the Gaussian case with known standard deviation. The top row corresponds with $\mu_h = \mu_c = 10$, the bottom row corresponds with $\mu_h = 9$ and $\mu_c = 10$.*

it is striking that in the congruent case the sample size has relatively little impact on the posterior. Further irrespective of the sample size, the posterior median for λ was around 0.6 and 0.7 for Beta(1,1) and Beta(0.5,0.5), respectively. Additional simulations under the congruent case (results not shown) revealed that the uncertainty of the power parameter does not materially decrease with increasingly larger sample sizes ($n \to 10{,}000$). Neuenschwander, Branson and Spiegelhalter (2009) reached the same conclusion, which triggered recent developments to determine the power parameter in a different manner.

For Gaussian data with a known variance Pawel et al. (2023a) demonstrated that in the congruent case for an increasing sample size of the current study the marginal posterior of λ converges to a Beta($\lambda|\alpha_0 + 0.5$, β_0) distribution. For $\alpha = \beta = 1$ this implies a posterior mean for λ equal to 0.6, and equal to 0.7 for $\alpha = \beta = 0.5$, confirming the above simulation results. Thus these authors showed that in the congruent case of a Gaussian model with known variance pooling the historical and current data will never happen, and there will always be discounting. In the incongruent case the posterior mean of λ is pulled towards zero. The binomial model shows a similar behavior. We were able to confirm, by performing a small simulation study (results not shown), that the DMPP in a setting with multiple historical controls has a similar behavior.

The fact that the power prior does not borrow all historical data in the congruent case is to some extent comparable to the inability of a frequentist statistical test to prove that two populations are equal. The congruence of the observed historical and current data provides some evidence that the underlying model parameters are also congruent (or at least similar), but it does not provide a clear proof. It appears that the behavior of the MPP reflects that pooling is only appropriate when we know that both data sets have been sampled from the same population. Recently several authors suggested to work with the conditional power approach whereby the power parameter is directly determined from the discrepancy between the historical and current data. These recent approaches are reviewed now, but we start with the empirical Bayes approach of Gravestock and Held (2017).

3.2.1 *An empirical Bayesian approach.* To avoid the computation of the normalizing constant, Gravestock and Held (2017) proposed the *empirical Bayesian power prior* where the power parameter is estimated empirically. To this end, the marginal likelihood of the power parameter λ is maximized, that is,

$$\lambda(D, D_0) = \underset{\lambda \in [0,1]}{\arg\max}\, L(\lambda|D, D_0), \tag{8}$$

where $L(\lambda|D, D_0) = \int_\Theta L(\theta|D)\, p(\theta|\lambda, D_0)\, d\theta = \frac{\int_\Theta L(\theta|D) L(\theta|D_0)^\lambda p(\theta)\, d\theta}{\int_\Theta L(\theta|D_0)^\lambda p(\theta)\, d\theta}$. The empirical Bayesian power prior is then defined as a conditional power prior with the empirical estimate of λ. This method does not require to specify a prior for the power parameter and hence avoids the computational difficulties that go along with the MPP. For the other model parameters a prior still needs to be specified. However, one might argue that summarizing the power parameter using a single value likely grossly underestimates the uncertainty with which borrowing should be applied.

3.2.2 *The power prior approach using a discrepancy function.* What follows are hybrid approaches that determine the "best" power parameter based on evaluating the discrepancy between the historical and the current data in a direct way. In a second step, the conditional power prior is applied with the chosen power parameter. This was done in (Haddad et al., 2017; Haddad, 2020) for a single historical data set. The method uses a function α that measures the difference between $\overline{\theta}$ based on all data (historical and current) and $\tilde{\theta}$ based on only the current data. In their example, they defined $\alpha(\overline{\theta}, \tilde{\theta})$ as $2(1 - \Phi(Z))$ with Z a standardized

difference between the two estimates and $\Phi(\cdot)$ the cumulative normal distribution function. Then a monotonic discrepancy function $T(\alpha(\overline{\theta}, \tilde{\theta}))$, with range (0,1) computes the weight given to the historical data, with $T(\alpha) = 0$ for $\alpha = 0$. The power $\overline{\lambda}$ to downweight the historical data is estimated from $\overline{\lambda} = T(\alpha(\overline{\theta}, \tilde{\theta}))$. When $T(\alpha(\overline{\theta}, \tilde{\theta})) = 1$, $\overline{\lambda}$ is given a maximal value determined by the user.

To allow for pooling, Pan, Yuan and Xia (2017) proposed the conditional power prior approach with λ defined as a function of a discrepancy measure T (the authors called it rather a congruence measure) between $\mathcal{D}$ and $\mathcal{D}_0$. As for the approach of Haddad et al. the function that relates λ with the discrepancy measure is pre-specified. The measure T should be positive and converge to zero when $\theta = \theta_0$ ($\mathcal{D}$ and $\mathcal{D}_0$ congruent) and to ∞ when $\theta \neq \theta_0$ ($\mathcal{D}$ and $\mathcal{D}_0$ incongruent) as the sample sizes of the historical and current control increase. For a binary response one could use, for example, a scaled (for sample size) χ^2-statistic and for a continuous response, for example, a standardized Kolmogorov–Smirnov test. Further, a function $g(T)$ then maps the support of T to (0,1) in a monotonic manner such that $g(T) \to 1$ as $T \to 0$ and $g(T) \to 0$ for $T \to \infty$. To this end, Pan, Yuan and Xia (2017) proposed a logistic function $g(T) = \frac{1}{1+\exp(a+b\log T)}$. This function is then calibrated by simulations by choosing a and b such that under a congruent scenario strong borrowing will happen and little borrowing when there is incongruence. Thereby, this procedure aims to control the Type I error rate and to increase the power.

Jiang, Nie and Yuan (2023) suggested the *elastic prior*, which builds on the proposal of Pan et al. That is, Jiang et al. also make use of the above logistic function (they called it an elastic function) calibrated in the same way as done by Pan et al., but they bypassed the conditional power prior approach. Indeed, the authors simply inflated the variance of the posterior given $\mathcal{D}_0$ by a factor proportional to $g(T)^{-1}$, which suggests also a simpler definition of the effective sample size, see Section 5.2. These authors also claimed that with their approach the Type I error rate is better controlled (see Section 7) than with other dynamic borrowing methods. The extension to multiple historical controls, however, needs further research. Finally, the authors argued that their elastic power approach can better handle differential borrowing of several parameters by specifying a separate discrepancy function for each parameter.

3.2.3 *The multiparameter case.*

Correcting for covariates. The power prior approach assumes that the historical and current data sets are similar. But, even when Pocock's strict conditions are satisfied, it might be that demographics and/or other patient characteristics such as disease severity differ between the historical and current data. Further, there is also a tendency to use dynamic borrowing approaches even when Pocock's conditions are not all fulfilled as mentioned in Section 2. In that case, one might apply statistical correction before borrowing historical information. Incorporating covariates in the dynamic borrowing methods was done by, for example, by Ibrahim and Chen for the joint power prior (Ibrahim and Chen, 2000). Banbeta, Lesaffre and van Rosmalen (2022) developed MPP approaches for linear regression models, thereby extending these models to account for covariates. Here, we assume that the observed covariates capture almost all of the imbalance between the historical and the current data.

Another possibility is to summarize relevant patient characteristics using propensity scores, and then to account for measured confounding using propensity score matching, weighting or stratification. Classically, propensity score methods are used to account for confounding in non-randomized experimental settings, where the propensity score models the assignment to either treatment or control. In the context of historical data, propensity scores are however used to model the allocation between current and historical data. A detailed case study on the use of propensity score methods for including historical data is provided by Lin, Gamalo-Siebers and Tiwari (2018).

Up to now, we have focused on measured confounding, but the historical and current data can also differ by unmeasured covariates. When dynamic borrowing methods are planned at the design stage of the study one must be aware that the historical and current data may be different in an unforeseen way and one may be confronted with a prior-data conflict. Dynamic borrowing methods are designed to mitigate the effects of such differences in unmeasured covariates. But still, in Section 3.4 an approach is discussed that protects the trialist even more from such a problematic setting. Several authors have recently proposed hybrid methods which combine propensity score methods with dynamic borrowing, such as the combination of propensity score stratification with power priors (Wang et al., 2019) or a MAP approach (Liu et al., 2021). A variety of hybrid methods was considered by Wang et al. (2022), and they recommended the combination of propensity score matching or weighting with the commensurate prior (see Section 4). Our research group has recently developed a hybrid method using propensity score weighting combined with the power prior (Polak et al., 2023).

Partial borrowing power prior. Most often the model for the current data is based on more parameters than the model for the historical data. A simple example occurs when we wish to borrow information on the control rate in a future RCT where the interest lies in the beneficial effect of the experimental treatment measured for example, by a log(odds ratio). In some applications the model structure, and thus the model parameters, may differ even more between the historical data and the current data, so that the likelihood functions $L(\theta|\mathcal{D}_0)$ and $L(\theta|\mathcal{D})$ will have a different form. A simple example is when the covariates differ between data sources, or when only summary statistics are available for the historical data combined with individual-level data for the current study. There are also other situations where it is appropriate to borrow information from only a subset of the model parameters. Classically the power prior was developed to borrow information from all model parameters. To make the power prior more flexible, Ibrahim et al. (2012) and Chen et al. (2014) proposed and applied the use of *partial borrowing power priors* in several publications. In the partial borrowing power prior, some model parameters are integrated out of the informative prior, so that information on these parameters is not borrowed from the historical data. The historical data are then borrowed only through the remaining parameters, which are shared by the models for the historical data and the current data. Note that the elastic power prior approach in Section 3.2.2 provides an alternative way for partial borrowing.

3.3 The meta-analytic predictive prior

3.3.1 *The original approach.* Take the settings of Section 3.1.3, but now assume that the parameters of the historical trials and of the current trial are exchangeable, that is, $\theta_1, \ldots, \theta_K, \theta \sim G(\boldsymbol{\phi})$. Under this setting, Neuenschwander et al. (2010) suggested the *meta-analytic predictive (MAP) prior* approach. Originally they considered the MAP prior for $\theta_k = \text{logit}(\pi_k) \sim G(\boldsymbol{\phi}) \equiv \text{N}(\mu, \tau^2)$, with π_k the true proportion of events in the k^{th} historical control arm. More specifically, with the normality assumption $\theta_1, \ldots, \theta_K, \theta|\mu, \tau^2 \sim \text{N}(\mu, \tau^2)$, the MAP prior is actually the posterior predictive distribution (PPD) $\theta|\mathcal{D}_1, \ldots, \mathcal{D}_K$. Neuenschwander et al. (2010) showed that if σ_k and τ^2 are known and one takes a flat prior for μ then the MAP prior is $\text{N}\left(\theta|\frac{\sum w_k\widehat{\theta}_k}{\sum w_k}, \frac{1}{\sum w_k} + \tau^2\right)$, where $\widehat{\theta}_k$ is the estimated parameter obtained from $\mathcal{D}_k$ $(k = 1, \ldots, K)$, $w_k = \frac{1}{\sigma_k^2+\tau^2}$ are weights with $\sigma_k^2 (k = 1, \ldots, K)$ usually fixed. When K is small, the inter-study variance τ^2 is often given an informative and sensible prior. A large value of τ^2 implies that little can be learned from the past studies. Clearly, the MAP prior approach is again an example of a dynamic borrowing prior approach.

Two versions of the meta-analytic approach exist. The first version, called the MAP approach, is a two-step procedure where first the MAP prior is computed for the unknown

parameter θ and then combined with the current data. Alternatively, in the meta-analytic combined (MAC) approach, one combines the historical and current data in a hierarchical model as given above. Once the historical and current data are combined the two approaches are equivalent (Schmidli et al., 2014). But, the MAP approach requires that the historical data are available at the design stage of the study, while for the MAC approach the historical data can combined in real-time with the current data. For instance, in pediatric studies, when one wishes to extrapolate adult data to children, the MAP approach is useful when the adult data were obtained before, whereas the MAC approach offers the ability to combine the adult data with pediatric data when data on adults and children are collected at the same time.

A practical problem with the MAP approach occurs when the MAP prior is determined via Markov chain Monte Carlo methods. In that case, the MAP prior is only available from sampled values. In other words, there will be no analytical expression of this prior but rather a histogram of sampled values. Note that this problem may also occur with the joint and normalized power prior. In that case, Schmidli et al. (2014) suggested to approximate the MAP prior with a finite mixture of conjugate priors. This can be done with the R package RBesT (R Bayesian Evidence Synthesis Tools) (Weber et al., 2021). For instance, for a binomial response with θ as unknown success probability one could use a mixture of beta distributions to approximate the sampled MAP prior or alternatively the sampled MAP prior of logit(θ) could be approximated with a mixture of Gaussian distributions. Hence, the RBesT package can be used when the (uni-dimensional) posterior has no closed-form solution and one wishes to use it as a prior in a subsequent study. This is now illustrated.

Phase IV transplant trial: *The MAP prior.* We replay here the first example in Neuenschwander et al. (2010), that is, a phase IV trial in de novo transplant patients designed to compare a standard (control) treatment to an experimental treatment. The primary outcome is treatment failure. A conventional balanced design would require a total of 450 patients per treatment arm. It was therefore hoped that the information from 11 phase IV trials with essentially identical designs and amounting to 930 patients on the standard arm can be used to reduce the size of the study. The results obtained from these 11 trials were (# treatment failures/total number of patients in standard arm (%)): 6/33 (18%), 8/45 (18%), 17/74 (23%), 28/103 (27%), 26/140 (19%), 8/49 (16%), 22/83 (27%), 8/59 (14%), 6/22 (27%), 16/109 (15%) and 53/123 (23%).

The MAP approach assumes that the above proportions are estimates of true risks that are exchangeable, also with the risk in the future control arm. The MAP prior can easily be determined using, say, a simple BUGS program yielding a sample distribution as prior. Here we worked on the logit scale of the true risk. Then the RBesT package approximates the sampled MAP prior by a mixture of Gaussian distributions. This mixture can then be used as a prior in standard Bayesian software such as WinBUGS or JAGS. In Figure 2 the sampled MAP prior is shown, together with the approximating Gaussian mixture $0.59\mathrm{N}(-1.32, 0.59^2) + 0.41\mathrm{N}(-1.35, 0.38^2)$ and its components. This mixture is best out of several Gaussian mixtures (here up to 10 components) according to Akaike's information criterion as determined by the RBesT package.

3.3.2 *Correcting for covariates*. Since the MAP prior is based on a hierarchical model, it is straightforward to include covariates into the prior. The original exchangeability assumption is then replaced by conditional (on the covariates) exchangeability or partial exchangeability (Neuenschwander et al., 2010). Han et al. (2017) investigated the inclusion of patient-level covariates in this method.

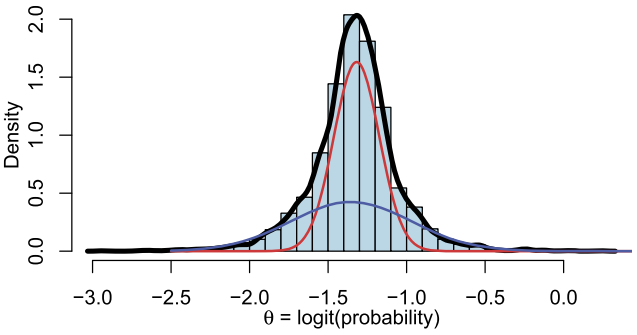

Figure 2 *MAP prior for the phase IV transplant study with the approximating mixture of Gaussian distributions and its two components.*

3.3.3 *Relaxing distributional assumptions.* The original MAP approach assumes a parametric distribution for θ. Often a Gaussian distribution is assumed for θ (or for a transformation of θ) or the more robust t-distribution with few degrees of freedom. To further relax the parametric assumptions, Hupf et al. (2021) proposed a semi-parametric MAP prior based on a non-parametric Bayesian approach. The approach handles the between-study heterogeneity in a different manner with a more flexible prior. Instead of a common between-study standard deviation as in the classical MAP approach, Hupf et al. assumed study-specific variations around the common overall mean using a Dirichlet process mixture prior for the study-specific between-study standard deviations. The Dirichlet process mixture prior is more robust and does not require the pre-specification of the amount of borrowing, i.e. the prior for the between-study standard deviation.

3.4 Protecting against prior-data conflict

Despite all precautions, the current data may be quite different from the historical data. If this happens, we speak of a prior-data conflict. Schmidli et al. (2014) suggested a variation of the MAP prior that can accommodate such a prior-data conflict by ignoring the historical information when it is too different from the current information. This is achieved by the *robustified MAP prior*, which is a mixture distribution for θ with one component the MAP prior (activated when the current control is similar to the historical controls) and a vague prior (applies when the current control is quite different from the historical controls). An example of such a robustified MAP prior is the Gaussian mixture given by

$$(1-w)\times \mathrm{N}\big(\theta|\mu,\tau^2\big)+w\times p_{0R}(\theta),$$

where $p_{0R}(\theta)$ is called the robust component and is here a Gaussian distribution with the same mean μ as the MAP prior but with a large variance and w is the mixing proportion usually taken small (0.1). We note that one must not take the robust variance too large to avoid Lindley's paradox (Mutsvari, Tytgat and Walley, 2016).

There exists also a *robustified power prior* based on the hierarchical version of the power prior for multiple historical studies, see Section 3.1.3. The robustified dependent power prior p_{RDMPP} for controls was derived from p_{DMPP} by Banbeta et al. (2019) in two ways:

- Version 1: $\lambda_k \sim (1-w)\times \mathrm{Beta}(\lambda_k|\alpha_\lambda,\beta_\lambda)+w\times p_{0R}(\lambda_k)(k=1,\ldots,K)$. In this version individual historical controls can be ignored in case of prior-data conflict.
- Version 2: $\boldsymbol{\lambda} \sim (1-w)\times \mathrm{Beta}(\boldsymbol{\lambda}|\alpha_\lambda,\beta_\lambda)+w\times p_{0R}(\boldsymbol{\lambda})$. Now either all or none of the historical controls are ignored.

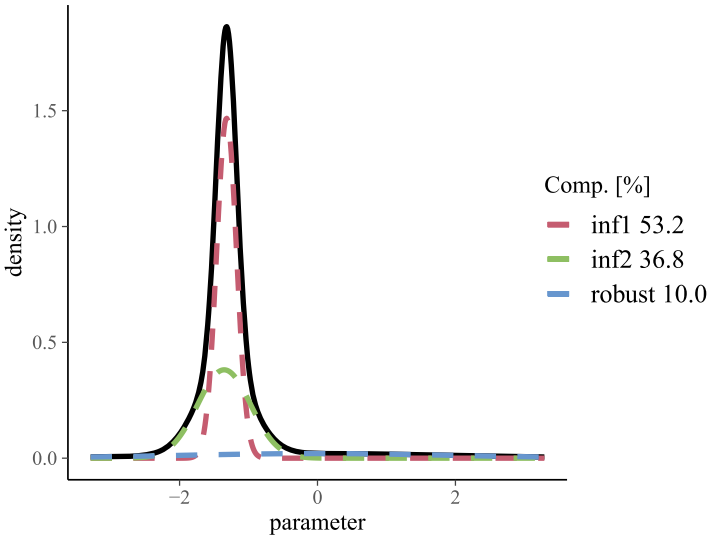

Figure 3 *Phase IV transplant study*: *Robustified MAP prior combining the MAP prior of Figure 2 with a robust component, i.e.* $N(0, 2^2)$ *as represented by the RBesT package. The two components of the MAP prior are denoted as 'inf*1*' and 'inf*2*' with weights* 53.2% *and* 36.8%*, respectively.*

Note that now the robust component in version 1, $p_{0R}(\lambda_k)$, is a Dirac function at zero implying that in case of a prior-data conflict the kth historical control is ignored. A similar result applies to the robust component in the second version. We illustrate the use of this robustified prior in Section 6 when we analyze the HOVON trials.

Phase IV transplant trial: *The robustified MAP prior.* A robustified MAP prior for the transplant study consists in augmenting the two-component Gaussian mixture for the logit probability with a vague Gaussian prior. Here, we have taken a $N(0, 2^2)$ prior as robust component. See Figure 3 for a graphical representation of the three-component robustified MAP prior produced by RBesT and based on the MAP prior of Figure 2.

3.5 Comparison of power prior and MAP prior

The power prior approach and the meta-analytic predictive prior approach have a different starting point:

- The power prior approach was first developed for a single historical study. For this approach, it is assumed that the parameter of interest is the same for the historical study and the current study.
- The MAP prior approach was developed for multiple historical studies and assumes exchangeability of the parameter of interest across the historical studies and the current study.

Chen and Ibrahim (2006) demonstrated that, for one historical study of size n_0 and i.i.d. Gaussian data with distribution $\mathrm{N}(\mu, \sigma^2)$, there is an exact relationship between the conditional power approach with a flat initial prior for μ and a hierarchical model, and hence they related the conditional power approach with the MAP prior developed by Neuenschwander and colleagues four years later (Neuenschwander et al., 2010). They thus proved that for $\lambda = \frac{1}{(1+2\tau^2 n_0/\sigma^2)}$ the p_{CPP} and the MAP prior (with chosen τ^2) match. They also proved the match for a Gaussian linear regression model with Zellner's g-prior $p(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, c(X^\top X)^{-1})$ for $\lambda = \frac{1}{(1+2cn_0/\sigma^2)}$, where X represents the design matrix. In addition they derived asymptotic relationships for multiple historical data sets and for generalized linear models. Pawel et al.

(2023b) showed that a beta prior for the MPP prior parameter λ is equivalent to a generalized F prior for the heterogeneity parameter τ^2.

Besides the above analytical results, several authors compared the performance of the two approaches via simulations. We refer to Section 7 for a review of how such comparisons are performed. A consistent finding from previous literature, based on both simulation studies and mathematical proofs (see, e.g., Pawel et al. (2023a) and Pawel et al. (2023b)), is that both the power prior and the MAP prior do not borrow all historical data. This result applies even when the historical and the current data have been drawn from the same population and the sample sizes are large.

4 Other dynamic borrowing methods

4.1 Pocock's approach and related proposals

Pocock (1976) was the first to propose a general Bayesian procedure to incorporate information from historical controls into a new trial. He assumed that the historical control parameter is a biased realization of the new control parameter, that is, $\theta = \theta_0 + \delta$. The bias δ is assumed to have a normal distribution with mean 0 and standard deviation σ_δ, i.e. $\delta \sim \mathrm{N}(0, \sigma_\delta^2)$, where σ_δ^2 is a measure of between-study heterogeneity. Given that it is difficult to estimate σ_δ^2 based on one historical control, Pocock recommended a pre-specified value for σ_δ, and a sensitivity analysis based on different values for σ_δ.

Hobbs, Sargent and Carlin (2012) suggested to use plausible priors for σ_δ to avoid the subjectivity of a pre-specified value, which led to the *commensurate prior*. The approach is closely related to the MAP approach, and there exists a one-to-one relationship between σ_δ^2 in the commensurate prior and the between-study standard deviation τ in the MAP approach with one historical control, that is, $\sigma_\delta^2 = 2\tau^2$. Using in fact the commensurate prior of Hobbs, Sargent and Carlin (2012), Röver and Friede (2020) showed how the MAP approach can effectively be used to borrow information from one historical study for the analysis of the current study and, more generally, to combine information from two different sources. The commensurate prior can be extended to multiple historical controls. Hobbs et al. (2011) also proposed a new version of the power prior, i.e. the *commensurate power prior*, which allows for different parameters for the historical and new data. The commensurate power prior directly parameterizes the commensurability of the historical and new data (i.e., the between-study variation) via a new commensurability parameter.

Another variation of Pocock's proposal was suggested by Ohigashi et al. (2022) who used a horseshoe prior when there are multiple historical control data. Suppose that the k^{th} control parameter is a possibly biased version of θ_0, that is, $\theta_k = \theta_0 + \delta_k$. As for the power prior approach it is expected that this bias is zero. The approach is then based on the following assumption

$$\delta_k \sim \mathrm{N}(0, \alpha_k^2 \tau^2),$$

with $\alpha_k \sim C^+(0, 1)$ and $\tau \sim C^+(0, 1)$, where $C^+(0, \sigma^2)$ is a half-Cauchy prior on the positive axis with scale parameter σ. When the current and historical controls are congruent, the posterior distribution of δ_k will concentrate around zero. But when the k^{th} historical control is in conflict with the current control group, α_k will be estimated far from zero. When many historical controls are in conflict, then the global shrinkage parameter τ will be estimated as large.

4.2 The test-then-pool approach

Viele et al. (2014) proposed a frequentist approach to incorporate historical information into a current study based on a *test-then-pool strategy*. For this approach, the historical data and the current data are first compared using a frequentist test to determine whether there exist a significant difference between the parameter in the historical control and the counterpart in the new control. If not statistically significant, the analysis is done by pooling the current and the historical control. Otherwise, the analysis will ignore the historical control.

5 Design aspects

5.1 Determining the necessary sample size

In an RCT, it is standard to compute the necessary sample size to show that with high probability a desired clinical effect of the experimental treatment, if it exists, can be demonstrated. For relatively simple frequentist statistical tests, analytical formulas can be used for this purpose. But, if the setting becomes a bit more complicated, as when mixed models are involved, computer simulations are needed to determine the power for a given sample size and hence such computations are also needed to determine the necessary sample size. For Bayesian inference computer simulations are inevitable when it comes to determining the power and the necessary sample size. Such calculations can show what the average effect is of borrowing information from a historical study when repeating the current study. This is done in the example below.

Planning the (fictitious) HOVON 43 study

In Section 6, we illustrate the use of dynamic borrowing methods on HOVON RCTs conducted by the HOVON organization since 1988. The trials called HOVON 4, HOVON 4A, HOVON 29, HOVON 42 and HOVON 42A had essentially the same control treatment for treating patients who suffer from acute myeloid leukemia (AML) when their bone marrow produces immature white blood cells (blasts) (van Rosmalen et al., 2018).

We now imagine that we wish to organize a future HOVON study, say HOVON 43, a successor to the HOVON 42A study that resulted in a success rate of 83.7% for the experimental arm to 82.6% for the control arm. The new experimental treatment is promising and it is believed that it will improve the success rate of the control treatment by about 5%. Suppose also that budget restrictions allow to allocate maximally 400 patients in each arm. The aim is to compute the power under these assumptions.

One can compute the Bayesian power for a given study size by pre-posterior calculations. Namely, one could sample from the future patient populations a large number of times, each time computing the posterior probability that the 95% credible interval of the difference in effect includes 0 or not. The proportion of times that this happens gives the power based on the conditional approach discussed in Section 7. Formally, let θ_c and θ_e be the true proportions in the current control and experimental arm, respectively and $\Delta = \theta_e - \theta_c$. Then the power is defined as the proportion of times the 95% CI for Δ excludes 0. This calculation was done here under two scenarios:

- θ_c and θ_e were given a uniform prior, expressing no prior information on these proportions, and
- θ_e is given a uniform prior, but for θ_c the power prior (equation (7) with an initial uniform prior $p_0(\theta)$ and a uniform prior for λ) is applied, thereby making use of the HOVON 42A study results with 214 patients out of 259 showing CR.

We also considered two settings with respect to the gain in success rate. In both cases $\theta_c = 0.83$ and $\theta_e = \theta_c + \Delta$, but in the first case we assumed a fixed $\Delta = 0.05$, while in the second case we made use of a design prior, that is, $\Delta \sim \text{N}(0.05, 0.01^2)$.

To determine the power, we sampled 1000 times from the assumed populations. Based on a self-written BUGS program, we obtained a power of 0.53 for Δ fixed at 0.05 and ignoring prior information on θ_c. This increased to a power of 0.65 when the HOVON 42A results were incorporated into the analysis. For the design prior, the powers were respectively, 0.56 and 0.77. This shows that including historical information may increase the power considerably, if the assumptions are correct.

5.2 Effective sample size

The impact of the historical data can be assessed using the above simulation-based approaches by comparing the necessary sample size under a non-informative prior for the control arm and that under the borrowing prior. This decrease in necessary sample size could be considered as the sample size represented by the borrowing prior. A less computer-intensive approach is to establish in a direct manner the equivalent number of fictitious subjects that is represented by the borrowing prior. This has been called the *prior effective sample size (PESS)* in the literature. Knowing the effective number of observations that impacts the analysis later on, allows to make sure that the historical information is not overwhelmingly large, so that the conclusions from the current analysis are not dominated by historical data. On the other hand, it is also useful to see whether it pays off to include historical information.

In the binomial and Gaussian (with σ known) case, the power in the conditional power prior represents the fraction of the historical information that will be used for the analysis of the current data. Namely, in these cases the PESS is easily calculated as λn_0, where n_0 is the size of the historical study. For the joint and the modified power prior, one might use $\widehat{\lambda} n_0$ for PESS, where $\widehat{\lambda}$ is the prior mean or median of λ.

Several proposals have been made for the PESS of a known prior. Their starting point is the notion of the equivalent amount of data in a conjugate prior. For instance, when a Beta(a, b) prior is used for a binomial likelihood, the PESS is equal to the sum $a + b$ (Morita, Thall and Müller, 2008). The two methods discussed below reduce to the known PESS when applied to the conjugate cases.

The proposal of Morita, Thall and Müller (2008) is based on the following reasoning. For a given prior $p(\theta)$, look for the sample size m such that the distance is minimized between this prior and the posterior $q_m(\theta|\boldsymbol{y})$, obtained by combining a minimally informative prior (called ϵ-information prior) $q_0(\theta)$ and the likelihood of a sample $\boldsymbol{y}$ of size m. Since the information in a distribution can be expressed as the curvature evaluated at a central value of the distribution, the authors computed the distance between the curvatures (second derivatives) of (1) the logarithm of $p(\theta)$ and (2) the logarithm of $q_m(\theta|\boldsymbol{y})$ at the mean of $p(\theta)$. Their suggestion for the PESS, PESS_{MTM}, reduces to the classical PESS when applied to conjugate priors.

The proposal of Neuenschwander et al. (2020) builds on the previous approach, but drops the ϵ-information prior from the calculation and integrates over the uncertainty of θ instead evaluating the distance at one particular point. Their so-called *expected local-information-ratio (ELIR) method* provides the PESS denoted here as PESS_{ELIR} that fulfills the predictive consistency criterion, that is, for a sample of size n, the expected *posterior ESS (PoESS)* must be the sum of PESS and n. Technically speaking, the PESS_{ELIR} is equal to the expectation of the ratio of the information of $p(\theta)$, $i(p(\theta)) = -\frac{d^2 \log p(\theta)}{d\theta^2}$, to the Fisher information of one information unit, $i_F(\theta) = -E_{y_1|\theta}\{\frac{d^2 \log f(y_1|\theta)}{d\theta^2}\}$. Namely,

$$\text{PESS}_{ELIR} = E_\theta \left\{ \frac{i(p(\theta))}{i_F(\theta)} \right\}.$$

The adjective 'effective' in PESS might be misleading since the effective impact of the borrowing prior can only be evaluated a posteriori, that is, when the prior is combined with the current data. A simple a posteriori calculation of the effective sample size was provided by Malec (2001). He proposed as PoESS

$$\text{PoESS}_M = n\frac{\text{Var}(\theta|data, \text{non-informative prior})}{\text{Var}(\theta|data, \text{informative prior})} - n.$$

The first part in the above expression expresses the total sample size when based on an informative prior compared to not using prior information. This value is subtracted with n, which is the original sample size, so with a non-informative prior. Hence, PoESS_M expresses the gain in sample size by using the informative prior.

Wiesenfarth and Calderazzo (2020) pointed out that the above PESSs ignore the impact of historical data on the new data, especially when there is prior-data conflict. Note that this is also the case with the above simulation-based approach. Hence, these authors argue that an effective sample size only makes sense when evaluating the impact of the prior on the actual results after combining the historical with the current data, as with the PoESS of Malec. They proposed the *effective current sample size (ECSS)*, which accounts for the influence of the prior on the newly observed data. The ECSS is calculated by finding the m that minimizes the distance of mean squared errors of (1) a posterior combining a sample of size k and a noninformative prior, and (2) a posterior combining a sample of size $k - m$ and the prior of interest.

The discrepancy approach of Jiang, Nie and Yuan (2023) provides an alternative for PESS. Since the variance of the dynamic borrowing prior is inflated by a factor $g(T)$, the authors defined the PESS simply as $g(T)n_0$, where n_0 is the sample size of the historical data set.

Thus, the simulation-based approaches and the proposals for the prior ESS may give us an idea of the impact of the historical data. However, the proof of the pudding lies in the eating, which in this case means that the ultimate impact of the historical data can only be seen when they are actually combined with the current data. This is a trivial conclusion, though.

The R package RBesT (Weber et al., 2021) supports the calculation of PESS_{MTM} and PESS_{ELIR} (and a third one). It has been documented and will also be illustrated below that these proposals often provide markedly different estimates for the PESS in non-conjugate settings. Note that there is no publicly available software for the other discussed proposals.

6 Applications

In this section, we will illustrate the dynamic borrowing methods using two applications. The first application examines selection criteria for historical controls and introduces the borrowing methods using a simple binomial model, based on HOVON data with a binary outcome. The second application highlights the computational aspects of the borrowing methods in a more complex negative binomial regression model, using data on incontinence episodes with a count outcome.

6.1 Application 1: HOVON data—binary outcome

6.1.1 *Selection of historical trials.* Here, we consider the five HOVON trials that were first mentioned in Section 5. All of these trials had essentially the same control treatment for AML. We consider here HOVON 42A as the current study that investigates the effect of, at that time, a possible promising new treatment. More details on these studies can be found in van Rosmalen et al. (2018). The control arms of the previous four trials provide historical information on the control arms to help estimate the efficacy of the investigational arm

in HOVON 42A. The endpoint of interest in all trials is complete remission or complete response (CR) from AML, which is a dichotomous outcome.

The HOVON trials were conducted over a long period of time. The first trial, HOVON 4, was conducted between 1988 and 1992, HOVON 4A between 1992 and 1993, HOVON 29 between 1997 and 2000, HOVON 42 between 2002 and 2004, and HOVON 42A between 2004 and 2006. Over a 20-year period, techniques in bone-marrow transplantation and supportive care for AML patients have considerably improved, for example, with new antibiotics, better antimicrobial prophylaxis and treatment, better blood transfusion support, and other medical care changes.

Pocock's criteria suggest that HOVON 4 and HOVON 4A are too old to be acceptable for the analysis of HOVON 42A and should therefore be discarded for our analysis. However, the studies HOVON 29, HOVON 42 and HOVON 42A were conducted in a relatively short period of time and the AML patients probably have received similar treatment and blood transfusion support. Hence, we assumed that the distribution of patient characteristics was similar across these three studies. For our analysis, this evaluation is important because we have first considered a simple binomial model, which does not allow for covariates to adjust for confounding.

In the investigational arm of HOVON 42A trial, 211 patients out of 252 (83.7%) showed CR, while 214 out of 259 (82.6%) in the control arm. The CR rates in the control arms of HOVON 29 and HOVON 42 trials were 86.3% (358 CR out of 437 patients) and 81.9% (598 CR out of 693 patients), respectively.

6.1.2 *Applying the borrowing methods*. We applied the power prior and the MAP approach to estimate the treatment effect (difference in CR rates between treatment arm and control arm) in HOVON 42A trial using: (a) a single historical (only HOVON 42) control arm, and (b) multiple historical control arms (HOVON 29 and HOVON 42). The two extreme cases, that is, complete pooling and ignoring the historical data, were also considered.

For the binomial case, the posterior of the MPP can be derived analytically as seen in Sections 3.1.2 and 3.1.3. For the empirical Bayesian power prior (EBPP) approach, seen in Section 3.2.1, the power parameter was first estimated and then the conditional power prior based on this power was applied. For the MAP approach, a half-normal distribution with mean zero and standard deviation 1, that is, HN(0, 1), was used for the between-trial standard deviation. The analyses were done on log odds scale, but the results in the tables below are given in differences in percentages. We used the JAGS software in combination with the R package MCMCpack (Martin, Quinn and Park, 2011). JAGS was used to obtain the posterior samples for the Current data, Pooled data, MAP, and Robust MAP methods, MCMCPack was used for the power prior (EBPP, MPP, DMPP, RDMPP1 and RDMPP2) methods.

A single historical control. The results of the different borrowing methods are summarized in Table 1. One can conclude that, taking into account the 95% CIs, the results are quite similar across the different methods, with the EBPP and MPP approaches resulting in a treatment effect estimate close to pooling, while the MAP and robust MAP approach provide an estimate basically equal to that obtained by analyzing just the current data. Thus, the MAP approaches tended to borrow less information from the historical study than the power prior approaches. Further, all borrowing methods reduced the uncertainty of the estimated treatment effect compared to using just the current data.

The posterior median of the power parameter was 0.57 with 95% equal-tail CI = [0.06, 0.98] for the MPP approach based on a Beta(1,1) prior for the power parameter. However, the estimated power was almost 1 for EBPP, which led to basically pooling the data from the historical control and current control arms.

Table 1 *HOVON data*: *Posterior summary measures of the treatment effect* (*difference in %*) *in the HOVON* 42*A trial based on different methods using the control data from the HOVON* 42 *study*

Method	Mean	Median	SD	95% CI
Current data	1.10	1.10	3.29	(−5.41, 7.54)
Pooled data	1.37	1.41	2.74	(−4.09, 6.56)
EBPP	1.36	1.38	2.74	(−4.16, 6.62)
MPP	1.31	1.41	2.91	(−4.51, 6.94)
MAP	1.10	1.10	3.10	(−5.08, 7.08)
Robust MAP	1.10	1.10	3.13	(−5.24, 7.02)

CI: equal-tail credible interval; EBPP: Empirical Bayesian power prior; MAP: meta-analytic predictive; MPP: modified power prior; SD: standard deviation.

Table 2 *HOVON data*: *Posterior summary measures of the treatment effect* (*difference in %*) *in the HOVON* 42*A trial based on different methods using the control data from the HOVON* 29 *and HOVON* 42 *studies*

Method	Mean	Median	SD	95% CI	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$
Current data	1.10	1.10	3.29	(−5.41, 7.54)	-	-
Pooled data	−0.73	−0.67	2.52	(−5.88, 4.07)	-	-
EBPP	0.70	0.76	2.63	(−4.64, 5.64)	0.12	0.99
MPP	−0.22	−0.14	2.75	(−5.96, 5.00)	0.48	0.55
DMPP	−0.28	−0.25	2.63	(−5.57, 4.75)	0.52	0.52
RDMPP1	−0.17	−0.10	2.74	(−5.69, 5.02)	0.45	0.50
RDMPP2	−0.21	−0.12	2.67	(−5.47, 4.96)	0.47	0.48
MAP	0.33	0.28	3.07	(−5.73, 6.49)	-	-
Robust MAP	0.32	0.31	3.10	(−5.65, 6.53)	-	-

The naïve estimate for the PoESS of a MPP is equal to $\widehat{\lambda} n_0$, where n_0 is the size of the control arm in the HOVON 42 study (equal to 693) and $\widehat{\lambda}$ is the posterior median of the power. Using the estimated powers, the PoESSs for EBPP and MPP were computed as 693 and 396, respectively. Using the suggested procedure of Malec, we also computed the PoESS for the different approaches, that is, pooling (115), EBPP (115), MPP (73), MAP (33) and robust MAP (28). The above calculation was done using the formulas in Section 5.2.

Multiple historical controls. Adding the HOVON 29 study to the historical data reduced the estimated treatment effect and even changed the sign for some of the methods, see Table 2. The uncertainty of the estimated treatment effect (posterior SD & 95% CI) was reduced only a bit. We conclude that the different dynamic approaches again indicate no treatment effect, but also that (again) the MAP approaches tend to borrow less information from the historical data than the power prior approaches. See Table 2 for the results. It is also of interest to look at the estimated median powers for the power prior approaches. There are now two powers to estimate: λ_1 for HOVON 29 and λ_2 for HOVON 42. As before, the estimated powers of the EBPP are extreme, putting doubt on the usefulness of this approach. For the MAP approach, the posterior median of τ (between-study standard deviation) was 0.25 on the log-odds scale, indicating a moderate variability among the HOVON trials (Neuenschwander et al., 2010).

We have again taken the naïve estimate of the PoESS for the power prior approaches. PoESSs based on $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ for EBPP, MPP, DMPP, RDMPP1, and RDMPP2 were 739, 589, 585, 541, and 539. PoESS values based on Malec's method for the above power prior approaches were 149, 114, 149, 117, and 137, respectively. The naïve PoESS estimates exaggerate considerably the impact of the historical data when compared with Malec's PoESS

values. On the other hand, the naïve estimates of the PoESS for the robustified power prior approaches were rightfully lower than the corresponding original power priors, which was not the case for Malec's estimates.

For the MAP approach, the MAP prior was approximated with a mixture of beta distributions using the RBesT package (Weber et al., 2021), which also computes the two proposals for a prior ESS. The results were $\text{PESS}_{MTM} = 120$ and $\text{PESS}_{ELIR} = 41$, which confirms that different methods can yield significantly different PESS values as mentioned in Section 5.2. PESSs for the robust MAP prior were $\text{PESS}_{MTM} = 77$, and $\text{PESS}_{ELIR} = 33$. The PoESSs based on Malec's method were 41 and 35 for the MAP and the robust MAP priors, respectively. As for the power prior approaches, the PESS value for the robustified MAP prior was lower than for the corresponding MAP prior, and this was also the case here for Malec's estimates. In addition, the PESS estimates for the MAP approaches appeared to overestimate the importance of the historical data, but to a much lesser extent than for the naïve PESS estimates of the power prior approaches. That there was quite some variability in the PESS estimates for the MAP prior was already mentioned by Neuenschwander et al. (2020). These authors argued though that only PESS_{ELIR} is appropriate. In contrast, the PoESSs based on Malec's method showed less variability.

6.2 Application 2: Incontinence episodes data—count outcome

We now apply the borrowing methods in a more complex setting. Three Phase III RCTs of treatments for patients with overactive bladder (OAB) were considered, described in Khullar et al. (2013) (RCT 1), Nitti et al. (2013) (RCT 2) and Herschorn et al. (2013) (RCT 3). The primary endpoint was the number of incontinence episodes observed three days prior to week 12. The variance of the counts showed in all trials overdispersion at baseline and at week 12. Therefore, we used a negative binomial regression model including sex, age group and the number of incontinence episodes at baseline as covariates for the analysis of the primary endpoint. In addition, we wished to incorporate the data of the control arms of RCT 1 and RCT 2 to better estimate the effect of active treatment over the control treatment in RCT 3. Baseline characteristics of the three studies are shown in Table 3.

The MAP approach can easily be applied to the negative binomial regression model. But for the MPP approaches, there is the difficulty that there is no closed-form expression for the normalizing constant. Hence, we needed to use a more sophisticated algorithm based on a path sampling algorithm (Banbeta et al., 2023) in combination with a Metropolis–Hastings algorithm using the R package MCMCpack. Note that the EBPP approach in this example

Table 3 *Incontinence episodes data*: *Descriptive statistics for the three trials*

Trial	Khullar et al.	Nitti et al.	Herschorn et al.	
Group	Control	Control	Control	Treatment
Number of subjects	291	325	262	256
Inc12, mean (SD)	4.6 (7.3)	5.3 (8.2)	4.5 (7.0)	3.4 (6.2)
Inc0, mean (SD)	8.0 (7.1)	9.0 (9.1)	7.3 (7.0)	7.5 (6.8)
Sex, *n* (%)				
Female	243 (83.5)	270 (83.1)	211 (80.5)	204 (79.7)
Male	48 (16.5)	55 (16.9)	51 (19.5)	52 (20.3)
Age group, *n* (%)				
< 65 years	177 (60.8)	191 (58.8)	165 (63.0)	148 (57.8)
≥ 65 years	114 (39.2)	134 (41.2)	97 (37.0)	108 (42.2)

Inc0, total number of incontinence events in the 3-day diary at baseline; Inc12, the number of incontinence episodes observed three days prior to week 12; SD, standard deviation.

Table 4 *Posterior summary measures of the incidence rate ratio (IRR) of 50 mg mirabegron versus placebo for patients with overactive bladder in the trial described in* Herschorn et al. (2013)*, estimated using different borrowing methods*

Methods	Mean	Median	SD	95% CI
Current data	0.731	0.724	0.105	(0.547, 0.958)
Pooled data	0.797	0.792	0.087	(0.640, 0.982)
MAP	0.755	0.751	0.096	(0.581, 0.954)
RMAP	0.754	0.749	0.095	(0.581, 0.955)
MPP Ind	0.768	0.762	0.098	(0.594, 0.975)
DMPP	0.767	0.759	0.093	(0.596, 0.960)
RDMPP1	0.764	0.757	0.095	(0.598, 0.967)
RDMPP2	0.767	0.760	0.097	(0.594, 0.980)

was not applied because we had difficulties obtaining the marginal likelihood to determine the powers.

From Table 4, we can conclude that the dynamic borrowing methods resulted in quite similar treatment effects, in-between the estimates obtained from complete pooling and by using only the current data. Also, all borrowing methods reduced the uncertainty of the estimated treatment effect compared to just using the current data. Furthermore, again estimates of the powers were obtained for the different MPP approaches, showing a similar behavior as before. We do not show the ESS values here, since they showed the same behavior as for the HOVON example, i.e. quite some variability for the prior ESSs but less for the posterior ESSs.

7 Operating characteristics of dynamic borrowing methods

Bayesian inference is based on the parameter space and not on the sample space. Unlike the frequentist approach, it does not have a built-in mechanism to control its behavior over repeated sampling. It is thus important to assess the operating characteristics (OCs), Pr(Type I error) and Pr(Type II error), of the dynamic borrowing methods. From a regulatory perspective, any increase in Pr(Type I error) above the nominal 5% level may be a cause for concern, and it is generally considered unethical to have a too high Pr(Type II error).

To calculate the OCs one must imagine what additional settings of data could have occurred besides the setting that was observed. The question is how rich these fictitious settings should be chosen. In the standard scenario, without historical data, the procedure to calculate the OCs is clear. Namely, one imagines all possible alternative outcomes of the current study (sample space) and then establishes how extreme the observed test statistic is compared to all possible settings, either analytically or via simulation.

For the calculation of the OCs of dynamic borrowing methods, the main question is whether alternative outcomes for the historical study should be considered part of the sample space. In other words, should one condition on the historical data when evaluating the OCs (i.e., a conditional approach) or should one determine the OCs by sampling simultaneously the data of the historical studies and the current study (i.e., an unconditional approach)? In both approaches the OCs can be calculated as proportions of repeatedly sampled data sets yielding a particular result.

To choose among these options, we may consider the context of how the study is planned. When the historical data are available at the design stage of the current study, the conditional approach, where the sampling space is restricted to the current study, may seem more natural. An unconditional approach appears more logical when the historical and the current

data are obtained simultaneously. Because the historical data are typically available at the design stage, there is a tendency to prefer the conditional approach for dynamic borrowing methods. However, such an approach may not be customary in other areas of statistics. For instance, in group sequential trials the OCs are determined by considering all possible sample paths. After the first interim analysis there is no intention to control the conditional type Pr(Type I error) given the data of the first stage (Quan et al., 2022). We could treat historical data in clinical trials similarly to data observed before an interim analysis, by considering historical and current data as part of a combined scientific procedure. Alternatively we could see the historical data as fully external to the current study. We thus argue that the choice between the conditional and the unconditional approach cannot be completely justified by statistical arguments: it should depend on the perspective that is adopted when performing an analysis.

To what extent can OCs illustrate the value of dynamic borrowing methods? It is natural to assume that historical data can provide a gain in statistical power while maintaining the Pr(Type I error) below the nominal significance level when the historical data are similar to the current data, and that there might be a loss in statistical power and/or an inflation of Pr(Type I error) in case of a prior-data conflict. Unfortunately this intuition does not fully hold under a conditional approach. Kopp-Schneider et al. showed that it is not possible to gain power while controlling Pr(Type I error) (Kopp-Schneider, Calderazzo and Wiesenfarth, 2020). A simplified version of their argument goes as follows. Under a conditional approach, including historical data will affect the results of hypothesis testing based on borrowing methods only through changes of the rejection region, which will be part of the current data sample space. For the current data, there are efficient statistical tests available (which cannot be improved upon in terms of the OCs), which have a rejection region defined on the current data sample space. Including historical data will lead to a different test, with a different rejection region, but this cannot lead to a substantial improvement of the OCs when these OCs are defined on the same sample space. The result of Kopp-Schneider et al. means that with a conditional approach, it is difficult to show any value of including historical data in an analysis.

In the unconditional approach, assumptions and/or scenarios are needed for the values of the model parameters of the historical data and how these may differ from the new trial. If we make the (slightly unrealistic) assumption that the model parameters of the historical data and the current data are identical, gains in power while controlling Type I error rate are possible. Such gains in power are, however, generally not possible if we insist on controlling Pr(Type I error) for all possible values of the model parameters of the historical data and the current data and the difference between these parameters. In our own research group, we have obtained useful comparisons of borrowing methods using an unconditional approach with a meta-analytic model for the population parameters of the historical trials and the current trial. OCs were calculated by sampling first population parameters for each trial, and then generating data given the sampled parameters per study (van Rosmalen et al., 2018). The variance of the meta-analytic model, which represents the amount of between-trial heterogeneity, can then be varied in scenarios to investigate the effects of differences between trials. Using this setup, we found that several borrowing methods are able to increase the statistical power for the treatment at a usually minor increase in Pr(Type I error). In this case, including historical data did provide some advantage, but the gains were limited and dependent on assumptions regarding the comparability of the historical data and the current data.

There currently is no consensus on how OCs of dynamic borrowing methods should be assessed. Some argue that, due to the possibility of a prior-data conflict, borrowing historical data always yields an increase in Pr(Type I error), which should be balanced against the increase in statistical power, see, for example, Viele et al. (2014). According to that line

of reasoning, some degree of inflation of Pr(Type I error) should be considered acceptable, although it is unclear what the upper limit (greater than the α of 0.05) should be. A more purist point of view is that any inflation of Pr(Type I error) is not acceptable, at least not for regulatory drug approval, and that therefore historical data and dynamic borrowing methods should not be used. There are some methods that can assess trade-offs between increases in statistical power and increases in Pr(Type I error), such as a calibrated power metric (Banbeta et al., 2019) and a decision-theoretic approach that maximizes a utility function after specifying costs for committing a Type I error and a Type II error (Caderazzo, Wiesenfarth and Kopp-Schneider, 2022). Nevertheless it is clear that any simulation should at least consider the possibility of a prior-data conflict, for example by evaluating various scenarios for the difference between historical and current data. The historical data should be chosen carefully, so that we may expect any prior-data conflict to be relatively small. We thus recommend that the breadth of scenarios is informed by a substantive assessment of the comparability of the historical trials and the current trial, for example, in terms of study design, patient characteristics, intervention and outcome measures. Furthermore, when reporting the results of a simulation of dynamic borrowing methods, it should be clear from the publication how the OCs are calculated and what the sample space is.

Due to the difficulties associated with OCs of dynamic borrowing methods, some researchers prefer to focus on other metrics, such as the bias in the estimated treatment effect, its mean square error and the width of the 95% CI of the treatment effect. Results of these metrics have been reported in many publications, see, for example, Viele et al. (2014). Compared to an analysis without historical data, most dynamic borrowing methods reduce the width of the posterior CIs, although the gains are sometimes limited. If the historical data are biased, then the estimates of the treatment effect of dynamic borrowing methods will generally also be biased, but typically less so than with naïve pooling of data sources. The mean square error, which combines bias and variance, may be reduced by using dynamic borrowing methods, because the reduction in variance often outweighs the bias. Viele et al. (2014) report that there is a *sweet spot* in the parameter values, where the historical data are sufficiently similar to the current data, such that the inclusion of historical data reduces the mean square error and the Pr(Type I error) and increases the power. Note however that the calculation of most of these metrics also requires choosing a sample space, and that there may thus be theoretical debate on how they should be calculated.

8 Software

Currently, there are multiple R packages or R scripts available for the implementation of the methods discussed in this paper.

8.1 Software for the power prior approach

As seen in Section 3.1.1, determining the conditional power prior and combining it with the likelihood of the current data is straightforward with standard Bayesian software such as Win/OpenBUGS, JAGS, etc. The same is true for the joint power prior, please consult the website reported at the end of the paper for some examples. More problematic is the computation of the MPP, as the normalizing constant needs to be determined at each iteration of the MCMC program. Hence, the MPP is often implemented in a Metropolis–Hastings algorithm, which can be done using the R package MCMCpack. R programs for the binomial, Poisson, negative binomial models and regression models can be found at the same website. The scaling constant of the MPP has a closed-form expression for the binomial model, Poisson and Gaussian model whereas for the negative binomial model we used a path sampling algorithm prior to posterior sampling.

In Section 3.2.2, we discussed two approaches to fix the power for the conditional power prior approach using a discrepancy function. For the approach suggested by Haddad et al. (Haddad et al., 2017; Haddad, 2020), Balcome et al. developed the R package bayesDP (Balcome et al., 2022), which makes use of MCMCpack. For the elastic power prior approach, the R scripts can be found (Jiang, Nie and Yuan, 2023) at https://onlinelibrary.wiley.com/doi/10.1111/biom.13551.

The R package psborrow https://CRAN.R-project.org/package=psborrow is a tool that combines propensity scores and the commensurate prior approach discussed in Section 4 and is based on Liu et al. (2019). The recent extension psborrow2 is the successor of psborrow. For more information we refer to https://genentech.github.io/psborrow2/develop/index.html.

8.2 Software for the MAP prior approach

The MAP approach is basically a hierarchical model, therefore its software implementation is less challenging than that of the power prior. Software packages such as Win/OpenBUGS, JAGS and Stan can be used for its implementation. The RBesT package is based on the package Stan and was developed specifically for the MAP prior. The ess function of the RBesT package was used to calculate ESS_{MTM} and PESS_{ELIR}. The RBesT package offers also to approximate any uni-dimensional posterior that has been generated by a sampling algorithm by a mixture of conjugate distributions, which is quite a nice feature.

The R package bayesmeta allows for a Bayesian random effects meta-analysis which provides shrinkage estimates that can be used for dynamic borrowing for the normal-normal hierarchical case without invoking an MCMC algorithm.

See https://rdocumentation.org/packages/bayesmeta/versions/3.3, but also Röver (2020) and Röver and Friede (2023).

9 Discussion

In this paper, we have focused on the two main approaches for dynamic borrowing information from historical studies: the power prior approach and the meta-analytic predictive prior approach. The two approaches seem quite different in their philosophy and setup. The power prior was first suggested when only a single historical study was available, while the MAP prior works best with more than one historical study. Both approaches have been generalized to a great variety of settings, such as generalized linear models, generalized linear mixed models, survival models, frailty models, multivariate models and nonlinear models. Although their assumptions and mechanics are quite different, Chen and Ibrahim (2006) showed that for the Gaussian case there is a mathematical relationship between the conditional power prior and the MAP prior developed four years later. Our experience, confirmed by findings in the literature (Gravestock and Held, 2017; Hupf et al., 2021), shows that the power prior approach tends to borrow more information from the historical studies than the MAP prior approach. In our analyses, the choice of the prior of λ had a non-negligible impact on its posterior. Finally, it appears that the empirical Bayes approach tends to produce too extreme values of the power parameter.

We have also shown in this review that there is a great variety of alternative approaches to dynamic borrowing, but they all are based on evaluating the discrepancy between the historical and current data to determine how much the historical data must be downplayed.

Having explored the different dynamic borrowing approaches, an important question pops up: ‘Which approach to choose in practice?’ The power prior approaches have been developed initially for a single historical study, but as seen here can be generalized easily, at least in principle, to multiple historical studies. The MAP prior approaches have been developed for

applications with multiple historical studies, but can work also when there is a single historical study. The calculation of the normalizing constant of the MPP becomes computationally prohibitive in general, while the inter-study variance τ^2 needs a well-chosen informative prior in order for the MAP prior approach to be effective. Apart from the aforementioned computational limitations, it is difficult to choose the best approach in general. What seems clear is that robustifying the priors is a necessity in practice to deal with the possibility of a prior-data conflict.

The dynamic borrowing methods are becoming increasingly popular in pharmaceutical research. Indeed, they allow to reduce the necessary sample size and therefore to reduce the risks for the recruited patients and to reduce costs. In this research area, regulatory authorities play a major role, since they decide on the registration of the newly developed drugs and serve as guardians to avoid accepting drugs that by chance show a positive result. This means that they are particularly worried about inflating the Type I error rate. For this reason they insist on taking frequentist criteria into account to judge the dynamic borrowing approaches. However, it has been shown that, while the borrowing methods may imply a considerable increase in power, this benefit will typically come with an increase of Pr(Type I error), depending also on what approach is used to calculate Pr(Type I error) and power.

Numerous simulation studies, conditional on the historical data or unconditional, have been performed to conclude that there is no free lunch. So one might think that this implies the end of the dynamic borrowing methods. However, a potential way out to this deadlock is not to focus primarily on the Pr(Type I error) but to use a decision-theoretic approach as suggested in Caderazzo, Wiesenfarth and Kopp-Schneider (2022). But, the question is also whether the frequentist criteria are the only beatific criteria. After all they refer to future studies, which probably will never happen in life (the classical argument for Bayesians to criticize the frequentist principles). Luckily, one can give several practical examples where the regulators have taken a less strict frequentist viewpoint. Indeed, an example of an early success of a Bayesian approach that makes use of historical information can be found in Recht et al. (2009).

In that paper, Recht et al. describe the process of demonstrating the safety of a production process of a recombinant factor VIII product called ReFacto Antihemophilic Factor (AF)®, for the treatment of hemophilia A patients. Refacto AF® is manufactured in a more sophisticated way than the previous version called Refacto®. The company wished to demonstrate safety of Refacto AF®, but the initial FDA requirements for a successful study revealed unrealistic necessary sample sizes of more than 10,000 patients. For this reason, the company suggested an alternative, Bayesian approach, to reduce the necessary sample size by borrowing information from the past and to make use of published data on pivotal registration trials for four licensed products. These data produced a beta posterior distribution from which they suggested a Bayesian criterion for the coming study to be successful. This was accepted by the FDA. Using historical data from earlier studies with Refacto® and combining it with the data of a study of 90 patients, the company was able to show the safety of the new manufacturing process, and obtained approval from FDA.

In general, the acceptance of Bayesian methods in clinical trial research goes slow. The regulators worry that the prior distributions are based on favorable data for the experimental drug, so that bias and an inflated Type I error rate are caused. As a result, the Bayesian approach is in general tolerated and accepted only when conventional clinical trial designs are impossible to implement in practice. This is the case for orphan diseases and pediatric studies, as mentioned in the Introduction. For medical devices the Bayesian approach is generally accepted by the regulatory authorities assuming, of course, the same rigorous setup and conduct as with a frequentist approach, see, for example, Haddad (2020). Moreover, the Bayesian approach often is the recommended approach to deal with small sample issues.

Price and Scott (2021) described a recent initiative of FDA to discuss the feasibility and acceptance of complex innovative trial designs, which are rarely used in new drug applications. Since conventional trial designs are becoming increasingly expensive and require often many patients, the FDA realized that in a number of cases alternative designs may be needed to accelerate product development or to make the product available at an earlier stage. To this end, the FDA has set up the Complex Innovative Trial Design Pilot Meeting Program. Sponsors may suggest in such a meeting an innovative design for their clinical trial which can then be discussed with the FDA for their acceptance.

Price and Scott (2021) described five accepted innovative designs. All of them are Bayesian in nature and make use of historical data. This initiative clearly shows the increasing interest of sponsors and regulators to embark on Bayesian clinical trial designs and to make use of the above discussed dynamic borrowing methods. For an additional reference that illustrates the usefulness and increasing popularity of Bayesian clinical trial designs, see Carlin and Nollevaux (2022).

Finally, note that most of the programs used in this paper can be found on the website of I-Biostat, that is, https://ibiostat.be/online-resources/bayesian.

Acknowledgments

A preliminary version of this review paper was presented by the first author at the Annual South African Statistical Association (SASA) meeting in 2021 in Stellenbosch, South Africa.

The authors thank Brad Carlin and David Dejardin for the interesting and helpful discussions during the preparation of the paper. The authors are indebted to Tim Friede whose comments on a previous version of the paper improved the review considerably.

Funding

The authors acknowledge the BOF bilateral cooperation of UHasselt for the financial support to the third author.

References

Balcome, S., Musgrove, D., Haddad, T. and Hickey, G. L. (2022). bayesDP: Implementation of the Bayesian discount prior approach for clinical trials R package version 1.3.6.

Banbeta, A., Lesaffre, E., Martina, R. and van Rosmalen, J. (2023). Bayesian borrowing methods for count data: Analysis of incontinence episodes in patients with overactive bladder. *Statistics in Biopharmaceutical Research* **00**, 1–15.

Banbeta, A., Lesaffre, E. and van Rosmalen, J. (2022). The power prior with multiple historical controls for the linear regression model. *Pharmaceutical Statistics* **21**, 418–438.

Banbeta, A., van Rosmalen, J., Dejardin, D. and Lesaffre, E. (2019). Modified power prior with multiple historical trials for binary endpoints. *Statistics in Medicine* **38**, 1147–1169.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. *IMS Lecture Notes—Monograph Series* **9**.

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

Carlin, B. and Nollevaux, F. (2022). Bayesian complex innovative trial designs (CIDs) and their use in drug development for rare disease. *Journal of Clinical Pharmacology* **62**, S56–S71.

Chen, M.-H. and Ibrahim, J. G. (2006). The relationship between the power prior and hierarchical models. *Bayesian Analysis* **1**, 551–574.

Chen, M.-H., Ibrahim, J. G. and Shao, Q.-M. (2000). Power distributions for generalized linear models. *Journal of Statistical Planning and Inference* **84**, 121–137.

Chen, M.-H., Ibrahim, J. G., Zeng, D., Hu, K. and Jia, C. (2014). Bayesian design of superiority clinical trials for recurrent events data with applications to bleeding and transfusion events in myelodyplastic syndrome. *Biometrics* **70**, 1003–1013.

De Santis, F. (2006). Power priors and their use in clinical trials. *American Statistician* **60**, 122–129.

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

Haddad, T. (2020). Bayesian statistical methodology in the medical device industry. In *Bayesian Methods in Pharmaceutical Research* (E. Lesaffre, G. Baio and B. Boulanger, eds.) 467–483.

Haddad, T., Himes, A., Thompson, L., Irony, T. and Nair, R. (2017). Incorporation of stochastic engineering models as prior information in Bayesian medical device trials. *Journal of Biopharmaceutical Statistics* **27**, 1089–1103.

Han, B., Zhan, J., John Zhong, Z., Liu, D. and Lindborg, S. (2017). Covariate-adjusted borrowing of historical control data in randomized clinical trials. *Pharmaceutical Statistics* **16**, 296–308.

Hatswell, A., Freemantle, N., Baio, G., Lesaffre, E. and van Rosmalen, J. (2020). Summarising salient information on historical controls: A structured assessment of validity and comparability across studies. *Clinical Trials* **17**, 607–616.

Herschorn, S., Barkin, J., Castro-Diaz, D., Frankel, J., Espuna-Pons, M., Gousse, A., Stoelzel, M., Martin, N., Gunther, A. and van Kerrebroeck, P. (2013). A phase III, randomized, double-blind, parallel-group, placebo-conrolled, multicentre study to assess the efficacy and safety of the β_3 adrenoceptor agonist, mirabegron, in patients with symptoms of overactive bladder. *Urology* **82**, 313–320.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67**, 1047–1056.

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

Hupf, B., Bunn, V., Lin, J. and Dong, C. (2021). Bayesian semiparametric meta-analytic-predictive prior for historical control borrowing in clinical trials. *Statistics in Medicine* **40**, 3385–3399.

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

Ibrahim, J. G. and Chen, M.-H. (2015). The power prior: Theory and applications. *Statistics in Medicine* **34**, 3724–3749.

Ibrahim, J. G., Chen, M.-H., Xia, H. A. and Liu, T. (2012). Bayesian meta-experimental design: Evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics* **68**, 578–586.

Jiang, L., Nie, L. and Yuan, Y. (2023). Elastic priors to dynamically borrow information from historical data in clinical trials. *Biometrics* **79**, 49–90.

Khullar, V., Amarenco, G., Angulo, J., Cambroner, J., Hoye, K., Milsom, I., Radzisewski, P., Rechberger, T., Boerigter, P., Drogendijk, T., Wooning, M. and Chapple, C. (2013). Efficacy and tolerability of mirabegron, a beta(3)-adrenoceptor agonist, in patients with overactive bladder: Results from a randomized European–Australian Phase 3 trial. *European Urology* **60**, 283–295.

Kopp-Schneider, A., Calderazzo, S. and 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**, 361–374.

Lesaffre, E., Baio, G. and Boulanger, B. (2020). *Bayesian Methods in Pharmaceutical Research*: CRC Press.

Lin, J., Gamalo-Siebers, M. and Tiwari, R. (2018). Propensity score matched augmented controls in randomized clinical trials: A case study. *Pharmaceutical Statistics* **17**, 629–647.

Liu, J., Lu, C., Jiang, Z., Alemayehu, D., Nie, L. and Chu, H. (2019). Borrowing from historical control data in cancer drug development: A cautionary tale and practical guidelines. *Statistics in Biopharmaceutical Research* **11**, 67–78.

Liu, J., Lu, C., Jiang, Z., Alemayehu, D., Nie, L. and Chu, H. (2023). Borrowing concurrent information from non-concurrent control to enhance statistical efficiency in platform trials. *Current Oncology* **30**, 3964–3973.

Liu, M., Bunn, V., Hupf, B., Lin, J. and Lin, J. (2021). Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Statistics in Medicine* **40**, 4794–4808.

Malec, D. (2001). A closer look at combining data among a small number of binomial experiments. *Statistics in Medicine* **20**, 1811–1824.

Martin, A. D., Quinn, K. M. and Park, J. (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software* **42**, 323.

Morita, S., Thall, P. F. and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics* **64**, 595–602.

Mutsvari, T., Tytgat, D. and Walley, R. (2016). Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics* **15**, 28–36.

Neelon, B. and O'Malley, A. J. (2010). Bayesian analysis using power priors with application to pediatric quality of care. *Journal of Biometrics and Biostatistics* **1**, 1–9.

Neuenschwander, B., Branson, M. and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine* **28**, 3562–3566.

Neuenschwander, B., Capkun-Niggli, G., Branson, M. and Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials* **7**, 5–18.

Neuenschwander, B., Weber, S., Schmidli, H. and O'Hagan, A. (2020). Predictively consistent prior effective sample sizes. *Biometrics* **76**, 578–587.

Nitti, V., Auerbach, S., Martin, N., Calhoun, A., Lee, M. and Hershorn, S. (2013). Results of a randomized phase III trial of mirabegron in patients with overactive bladder. *The Journal of Urology* **189**, 1388–1395.

Ohigashi, T., Maruo, K., Takashi, S. and Gosho, M. (2022). Using horseshoe prior for incorporating multiple historical control data in randomized controlled trials. *Statistical Methods in Medical Research* **31**, 1392–1404.

Pan, H., Yuan, Y. and Xia, J. (2017). A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Applied Statistics* **66**, 979–996.

Pawel, S., Aust, F., Held, L. and Wagenmakers, E.-J. (2023a). Normalized power priors always discount historical data. *Stat* **12**, e591.

Pawel, S., Aust, F., Held, L. and Wagenmakers, E.-J. (2023b). Power priors for replication studies. Test.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases* **29**, 175–188.

Polak, T. B., Labrecque, J. A., Uyl-de Groot, C. A. and van Rosmalen, J. (2023). Augmenting treatment arms with external data through propensity-score weighted power-priors: An application in expanded access. arXiv preprint. Available at arXiv:2306.01557.

Price, D. and Scott, J. (2021). The U.S. Food and Drug aDministration's complex innovative trial design pilot meeting program: Progress to date. *Clinical Trials* **18**, 706–710.

Quan, H., Chen, X., Chen, X. and Luo, X. (2022). Assessments of conditional and unconditional type I error probabilities for Bayesian hypothesis testing with historical data borrowing. *Statistics in Biosciences* **14**, 139–157.

Recht, M., Nemes, L., Matysiak, M., Manco-Johnson, M., Lusher, J., Smith, M., Mannucci, P., Hay, C., Abshire, T., O'Brien, A., Hayward, B., Udata, C., Roth, D. A. and Arkin, S. (2009). Clinical evaluation of moroctocog alfa (AF-CC), a new generation of B-domain deleted recombinant factor VIII (BDDrFVIII) for treatment of haemophilia A: Demonstration of safety, efficacy, and pharmacokinetic equivalence to full-length recombinant factor VIII. *Haemophilia* **15**, 869–880.

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software* **93**, 1–51.

Röver, C. and Friede, T. (2020). Dynamically borrowing strength from another study through shrinkage estimation. *Statistical Methods in Medical Research* **29**, 293–308.

Röver, C. and Friede, T. (2023). Using the bayesmeta R package for Bayesian random-effects meta-regression. *Computer Methods and Programs in Biomedicine* **299**, 107303.

Schmidli, H., Neuenschwandert, B. and Friede, T. (2017). Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials. *Computational Statistics & Data Analysis* **113**, 100–110.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014). Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials. *Biometrics* **70**, 1023–1032.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B. and Lesaffre, E. (2018). Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research* **27**, 3167–3182.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., et al (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* **13**, 41–54.

Wadsworth, I., Hampson, L. V. and Jaki, T. (2018). Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods. *Statistical Methods in Medical Research* **27**, 398–413.

Wandel, S., Neuenschwander, B., Röver, C. and Friede, T. (2017). Using phase II data for the analysis of phase III studies: An application in rare diseases. *Clinical Trials* **14**, 277–285.

Wang, C., Li, H., Chen, W. C., Lu, N., Tiwari, R., Xu, Y. and Yue, L. Q. (2019). Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics* **29**, 731–748.

Wang, X., Suttner, L., Jemielita, T. and Li, X. (2022). Propensity score-integrated Bayesian prior approaches for augmented control designs: A simulation study. *Journal of Biopharmaceutical Statistics* **32**, 170–190.

Weber, S., Li, Y., Seaman, J. W. III, Kakizume, T. and Schmidli, H. (2021). Applying meta-analytic-predictive priors with the R Bayesian evidence synthesis tools. *Journal of Statistical Software* **100**(19), 1–32.

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