Received: 23 January 2023 | Revised: 31 July 2023 | Accepted: 23 October 2023

DOI: 10.1002/pst.2349

WILEY

MAIN PAPER

A dynamic power prior approach to non-inferiority trials for normal means

Francesco Mariani | **Fulvio De Santis** | **Stefania Gubbiotti**

Dipartimento di Scienze Statistiche, Sapienza University of Rome, Rome, Italy

Correspondence
Francesco Mariani, Dipartimento di Scienze Statistiche, Sapienza University of Rome, Piazzale Aldo Moro n.5, 00185 Rome, Italy.
Email: f.mariani@uniroma1.it

Abstract

Non-inferiority trials compare new experimental therapies to standard ones (active control). In these experiments, historical information on the control treatment is often available. This makes Bayesian methodology appealing since it allows a natural way to exploit information from past studies. In the present paper, we suggest the use of previous data for constructing the prior distribution of the control effect parameter. Specifically, we consider a dynamic power prior that possibly allows to discount the level of borrowing in the presence of heterogeneity between past and current control data. The discount parameter of the prior is based on the Hellinger distance between the posterior distributions of the control parameter based, respectively, on historical and current data. We develop the methodology for comparing normal means and we handle the unknown variance assumption using MCMC. We also provide a simulation study to analyze the proposed test in terms of frequentist size and power, as it is usually requested by regulatory agencies. Finally, we investigate comparisons with some existing methods and we illustrate an application to a real case study.

KEYWORDS

Bayesian clinical trials, borrowing historical information, fixed-margin approach, Hellinger distance, normal endpoints, unknown variance

1 | INTRODUCTION

Non-inferiority (NI) clinical trials aim to establish whether a new experimental treatment is not worse than a standard one (active control), which is known to be effective. Much controversy exists regarding the ethic of placebo-controlled trials[1,2]: NI tests are quite popular because they do not require a placebo arm. NI trials are largely used in several medical and pharmaceutical contexts, such as vaccine experimentation[3–6]; oncology studies[7]; anti-infective product trials.[8] Statistical methodologies for NI trials have been developed mainly under a frequentist perspective. However, past data providing information on the control arm are often available: this makes Bayesian inference an ideal framework for these trials. In fact, the Bayesian approach allows inclusion of historical knowledge for the control treatment parameter of the current study by means of the prior distribution.[9–15] Borrowing historical data implies in general that more trial resources can be devoted to the novel treatment while retaining accurate estimates of the current control arm parameter. Nevertheless, the potential lack of full homogeneity between current and previous information may recommend

downweighting of historical data. Several methods for discounting prior information have been proposed in the literature: see, among others, Neuenschwander et al.,[16] Hobbs et al.[17] and, for a comprehensive review on historical data borrowing, Viele et al.[15] Among several alternative borrowing methodologies, we here focus on power priors, originally defined by Ibrahim and Chen.[18] See also De Santis,[19] De Santis,[20] Ibrahim et al.[21] The power prior is proportional to the product of an initial prior (often non informative) and the likelihood based on previous data raised to a coefficient a that takes values in $[0,1]$: the larger a, the stronger the borrowing from historical data. The choice of a allows one to take into account both heterogeneity and differences in sample sizes of historical and current data.

In its original definition a is either a fixed coefficient – unrelated to current and historical data – or a random variable. More recently, several contributions follow a dynamic approach where a is based on a measure of similarity between historical and current data: see, for instance, Pan et al.,[22] Gravestock and Held,[23] Liu[13] Nikolakopoulos et al.,[24] Ollier et al.,[25] Bennet et al.,[26] Shi et al.[27] The fundamental differences between all these approaches rely on the choice of the measure of similarity used to define the discount parameter a. In particular, Pan et al.[22] define a as a measure of congruence between historical and current data based on the Kolmogorov–Smirnov statistic. Gravestock and Held[23] introduce an empirical Bayes approach to estimate a by maximizing its marginal likelihood based on historical and current data; this approach has then been extended to multiple historical studies.[28] Liu[13] proposes a p-value based dynamic borrowing approach, where a is a function of the p-value of a test on the difference between parameters of historical and current control treatment. Nikolakopoulos et al.[24] suggest the use of the Box prior predictive p-value[29] that quantifies the conflict between the priors based on historical and current data. In order to tune the commensurability parameter a in the context of phase I bridging studies, Ollier et al.[25] introduce the idea of using the Hellinger distance between the normalized likelihoods of the control parameter given the historical and current control data. Bennet et al.[26] propose to define a on the basis of two alternative measures of agreement between current control and historical data. These quantities are both functions of the posterior distributions of the respective parameters. The first measure (*probability weight*) is a tail posterior probability computed for the difference between historical and current control parameters. The second approach (*equivalence probability weight*) exploits the posterior probability that the difference between historical and current control parameters lies within a pre-speficied interval of equivalence. A similar measure of agreement is used in De Santis and Gubbiotti.[30] Finally, Shi et al.[27] construct the agreement parameter a using the overlap area under the sampling distributions of the current control and historical data: the stronger the congruence, the closer the area to one. Note that, unlike the previous ones, this method does not depend on observed data. As a data-dependent variant of this basic idea, the authors suggest to penalize the borrowing coefficient taking into account the p-value of a two-sided test on the equality of the control and historical parameters.

Of all the concepts contained in the above proposals, in this article we select two main ideas for defining a: (i) the use of posterior distributions of the parameters, as in Bennet et al.[26]; (ii) the use of a formal distance between these distributions, as in Ollier et al.[25] who, however, consider normalized likelihoods rather than posteriors. The proposed approach, previously considered in different contexts by De Santis and Gubbiotti,[31,32] is here implemented in the framework of NI trials for normal outcomes with both known and unknown variance. Specifically we follow the analysis of Gamalo et al.,[10] who, however, incorporate all historical information (full borrowing), thus ignoring potential heterogeneity between past and current studies in the elicitation of the control parameter prior distribution.

The outline of the paper is as follows. In Section 2.1, we describe the NI setting[10] and we formalize the problem for normal random variables. In Section 2.2, we describe the construction of the dynamic power prior. In Section 2.3, we address the unknown variance case and we provide a MCMC algorithm for implementation of posterior analysis. In Section 3.1, we explore the frequentist properties (size and power) of the proposed methodology through a simulation study; in Section 3.2, we compare our proposal with other dynamic power prior approaches. In Section 4, we illustrate an application to drug data. All the analyses presented in Sections 3 and 4 are performed using R[33]; code is available as supplementary material S1. Finally, Section 5 contains some concluding remarks.

2 | METHODOLOGY

2.1 | NI trial setting

Let us consider a NI trial in which an experimental therapy E is compared to an active control C. We assume that the latter is effective since in past studies the same drug C_0 has been tested against a placebo P_0. We are interested in studying drug average effects by assuming that treatment responses in historical and current trials are distributed as mutually

independent normal random variables with unknown means μ_i, for $i = P_0, C_0, C, E$. Specifically, let $\mathbf{X}_i = (X_{i1}, ..., X_{ij}, ..., X_{in_i})$ be n_i i.i.d. responses under treatment i, with $X_{ij} \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2)$ and let $\overline{X}_i$ be the corresponding sample means, with $\overline{X}_i \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2/n_i)$, $i = P_0, C_0, C, E$. Let us also assume independence between the $\mathbf{X}_i$'s and, for the time being, known variances. Non-inferiority of the experimental therapy with respect to the active control is declared by rejecting the null hypothesis of the test:

$$H_0 : \theta \leq -\delta \quad \text{vs} \quad H_1 : \theta > -\delta \tag{1}$$

where $\theta = \mu_E - \mu_C$ and $\delta > 0$ is the so-called NI margin.

In order to define a Bayesian methodology for NI clinical trials, one should specify:

a. assumptions over the parameters;
b. choice of the NI margin δ, that may depend on past performances of the active control and that may be modified according to medical judgement;
c. rejection rule of the test.

This problem has been recently addressed by Gamalo et al.[10] as follows.

i. Non informative priors are assumed for μ_{P_0}, μ_{C_0} and μ_E.
ii. The prior distribution of μ_C is based on full borrowing of historical control data.
iii. The NI margin is chosen to be

$$\delta = (1 - \lambda) \times L_{C_0 - P_0}, \quad \lambda \in [0, 1] \tag{2}$$

where $L_{C_0 - P_0}$ is the lower bound of the $(1 - \alpha)$ credible interval for $\mu_{C_0} - \mu_{P_0}$ and where λ allows to take into account potential clinical judgements.

iv. Non-inferiority of the experimental treatment (rejection of H_0) is declared if

$$L_{E-C} > -\delta, \tag{3}$$

where L_{E-C} is the lower bound of the $(1 - \alpha)$ credible interval for $\mu_E - \mu_C$.

In the following section we describe how we modify point (*ii*) by adjusting the amount of borrowing using a dynamic power prior for μ_C.

2.2 | Priors and posteriors

Implementation of Bayesian NI test requires posterior distributions of the parameters of interest μ_i, $i = P_0, C_0, C, E$. As in Gamalo et al.[10] we consider non informative priors $\pi(\mu_i) \propto 1$ for $i = P_0, C_0, E$. Thus, the posterior density of μ_i is

$$\mu_i \mid \mathbf{X}_i \sim N\left(\overline{X}_i, \frac{\sigma_i^2}{n_i}\right), \quad i = C_0, P_0, E. \tag{4}$$

Past data regarding the active control C_0 can be exploited to define a power prior distribution for μ_C, that is proportional to a starting prior $\pi_0(\mu_C)$ times the likelihood associated to $\mathbf{X}_{C_0}$, $L(\mu_C; \mathbf{X}_{C_0})$, raised to a power $a \in [0, 1]$. This quantity tunes the amount of the borrowed historical information: if $a = 0$ it is totally neglected; if $a = 1$ it is fully incorporated in the current experiment; intermediate values of a represent partial borrowing. Therefore, the power prior for μ_C is

$$\pi^P(\mu_C|\mathbf{X}_{C_0}) \propto L(\mu_C;C_0)^a \times \pi_0(\mu_C), \quad a \in [0,1] \tag{5}$$

Assuming a non-informative starting prior, that is, $\pi_0(\mu_C) \propto 1$, the power prior of μ_C is

$$\pi^P(\mu_C|\mathbf{X}_{C_0}) = N\left(\mu_C|\overline{X}_{C_0}, \frac{\sigma^2_{C_0}}{an_{C_0}}\right). \tag{6}$$

where $N(\cdot|m,v^2)$ denotes the normal density function of parameters m and v^2. Hence, the resulting posterior distribution of μ_C is

$$\pi^P(\mu_C|\mathbf{X}_{C_0},\mathbf{X}_C) = N\left(\mu_C|\widetilde{\mu}_C,\widetilde{\sigma}^2_C\right) \tag{7}$$

where

$$\widetilde{\mu}_C = \widetilde{\sigma}^2\left(\frac{n_C\overline{X}_C}{\sigma^2_C} + \frac{an_{C_0}\overline{X}_{C_0}}{\sigma^2_{C_0}}\right) \quad \text{and} \quad \widetilde{\sigma}^2_C = \left(\frac{n_C}{\sigma^2_C} + \frac{an_{C_0}}{\sigma^2_{C_0}}\right)^{-1}. \tag{8}$$

The posterior distribution of the parameter of interest $\theta = \mu_E - \mu_C$ is

$$\mu_E - \mu_C \mid \mathbf{X}_E,\mathbf{X}_C,\mathbf{X}_{C_0} \sim N\left(\overline{X}_E - \widetilde{\mu}_C, \frac{\sigma^2_E}{n_E} + \widetilde{\sigma}^2_C\right). \tag{9}$$

To take into account potential lack of homogeneity between historical and current studies, instead of setting a prefixed value of a, following the dynamic approach proposed by Ollier et al.[25] and De Santis and Gubbiotti,[31] we define a as

$$a = \kappa \cdot (1 - d_H[\pi(\mu_C|\mathbf{X}_C), \pi(\mu_C|\mathbf{X}_{C_0})]), \tag{10}$$

where $\kappa \in [0,1]$ is a static coefficient which provides an upper limit to the quantity of information that we borrow and d_H is the Hellinger distance between the posterior distributions of μ_C, respectively, obtained by updating the non-informative $\pi_0(\mu_C)$ with $\mathbf{X}_C$ and $\mathbf{X}_{C_0}$. In other words κ represents the maximum proportion of historical information one is willing to incorporate into the current analysis in case of perfect compatibility ($d_H = 0$). For a given value of κ, the more compatible information provided by $\pi(\mu_C|\mathbf{X}_C)$ and $\pi(\mu_C|\mathbf{X}_{C_0})$, the larger a. Since both d_H and κ range in $[0,1]$, then a takes on values in the unit interval as well. In the normal case here considered, the Hellinger distance is

$$d_H[\pi(\mu_C|\mathbf{X}_C), \pi_0(\mu_C|\mathbf{X}_{C_0})] = \left(1 - \sqrt{\frac{2\frac{\sigma^2_C\sigma^2_{C_0}}{\sqrt{n_C n_{C_0}}}}{\frac{\sigma^2_C}{n_C} + \frac{\sigma^2_{C_0}}{n_{C_0}}}} \exp\left\{-\frac{1}{4}\left[\frac{\left(\overline{X}_C - \overline{X}_{C_0}\right)^2}{\frac{\sigma^2_C}{n_C} + \frac{\sigma^2_{C_0}}{n_{C_0}}}\right]\right\}\right)^{1/2}. \tag{11}$$

See Appendix A for details. The null hypothesis (inferiority) of (1) is rejected if $L_{E-C} > -\delta$, where L_{E-C} is the lower bound of the $(1-\alpha)$ credible interval for $\mu_E - \mu_C$, namely the $\alpha/2$ level quantile of the posterior distribution given by Equation (9) and δ is given by Equation (2). Note that

$$L_{E-C} > -\delta \;\Leftrightarrow\; \overline{\mathbb{P}}(H_1) > 1 - \frac{\alpha}{2}, \tag{12}$$

where $\overline{\mathbb{P}}(H_1)=\mathbb{P}(\theta>-\delta|\mathbf{X}_E,\mathbf{X}_C,\mathbf{X}_{C_0})$ is the posterior probability computed with respect to (9) and represents a measure of evidence in favour of H_1.

2.3 | Unknown variance case

Let us now assume unknown variances. Following Gamalo et al.,[10] we suppose that σ_i^2's have independent Jeffrey's non informative priors, that is, $\pi(\sigma_i^2)\propto\sigma_i^{-2}$, for $i=P_0,C_0,E$. As regards σ_C^2, the Authors propose to use as prior distribution the posterior density based on historical data with no discount. Conversely, we prefer a non-informative prior for σ_C^2 that in this set up is a nuisance parameter, in order to restrict the use of historical data only to the construction of the prior for the effect parameter of the control group μ_C. Therefore, we consider the Jeffrey's prior $\pi(\sigma_C^2)\propto\sigma_C^{-2}$ for σ_C^2 as well. From standard conjugate analysis, the corresponding posterior distributions, conditional on μ_i, are independent inverse gamma, that is,

$$\sigma_i^2\,|\,\mathbf{X}_i,\mu_i\sim IG\left(\frac{n_i}{2},\frac{n_i\overline{S}_i^2}{2}\right), \tag{13}$$

where $\overline{S}_i^2=\frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij}-\mu_i)^2$ and $i=P_0,C_0,C,E$. In implementing the NI test, recall that L_{E-C} is the lower bound of the $(1-\alpha)$ credible interval of the *marginal* posterior of $\mu_E-\mu_C$. For standard conjugate analysis results the marginal posterior distribution of μ_i, $i=E,C$, are Student t densities. As in Gamalo et al.,[10] the value of L_{E-C} can be easily obtained numerically as the $\alpha/2$ empirical quantile of an MCMC draw from the marginal posterior distribution of $\mu_E-\mu_C$, where values of μ_E are sampled from $\mu_E\,|\,\mathbf{x}_E$ and values of μ_C are sampled from $\mu_C\,|\,\mathbf{x}_C,\mathbf{x}_{C_0},\sigma_C^2,\sigma_{C_0}^2$ (power posterior). The simulation steps are detailed below.

1. Given the observed historical data $\mathbf{x}_i$, compute sample means $\overline{x}_i$, sample variances s_i^2, for $i=P_0,C_0$, and determine $L_{C_0-P_0}=(\overline{x}_{C_0}-\overline{x}_{P_0})-z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_{C_0}^2}{n_{C_0}}+\frac{s_{P_0}^2}{n_{P_0}}}$.

2. Fix $\lambda\in[0,1]$ and set $\delta=(1-\lambda)\times L_{C_0-P_0}$ according to (2).
3. Consider non informative priors for μ_i and σ_i^2, $i=C_0,C,E$.
4. Let the observed data be $\mathbf{x}_i$, $i=C,E$.
5. Gibbs sampling from the marginal posterior $\mu_E\,|\,\mathbf{x}_E$.

   Set $M=10000$ (number of MCMC draws) and $B=1000$ (burn-in).
   Initialize σ_E^2.
   For $m=1,\ldots,M$, draw

- $\mu_E^{(m)}$ from $\mu_E\,|\,\mathbf{x_E},\sigma_E^{2(m-1)}\sim N\left(\overline{x}_E,\frac{\sigma_E^{2\,(m-1)}}{n_E}\right)$,
- $\sigma_E^{2\,(m)}$ from $\sigma_E^2\,|\,\mathbf{x_E},\mu_E^{(m)}\sim IG\left(\frac{n_E}{2},\frac{n_E\overline{s}_E^{2(m)}}{2}\right)$, where $\overline{s}_E^{2(m)}=\frac{\sum_{j=1}^{n_E}\left(x_{Ej}-\mu_E^{(m)}\right)^2}{n_E}$

6. Gibbs sampling from the marginal posteriors $\mu_i\,|\,\mathbf{x}_i$, $i=C,C_0$ for computing a

   For $m=1,\ldots,M$, draw

- $\mu_{C_0}^{(m)}$ from $\mu_{C_0}\,|\,\mathbf{x_{C_0}},\sigma_{C_0}^{2(m-1)}\sim N\left(\overline{x}_{C_0},\frac{\sigma_{C_0}^{2\,(m-1)}}{n_{C_0}}\right)$
- $\sigma_{C_0}^{2(m)}$ from $\sigma_{C_0}^2\,|\,\mathbf{x_{C_0}},\mu_{C_0}^{(m)}\sim IG\left(\frac{n_{C_0}}{2},\frac{n_{C_0}\overline{s}_{C_0}^{2(m)}}{2}\right)$, where $\overline{s}_{C_0}^{2(m)}=\frac{\sum_{j=1}^{n_{C_0}}\left(x_{C_0j}-\mu_{C_0}^{(m)}\right)^2}{n_{C_0}}$
- $\mu_C^{(m)}$ from $\mu_C\,|\,\mathbf{x_C},\sigma_C^{2(m-1)}\sim N\left(\overline{x}_C,\frac{\sigma_C^{2\,(m-1)}}{n_C}\right)$
- $\sigma_C^{2(m)}$ from $\sigma_C^2\,|\,\mathbf{x_C},\mu_C^{(m)}\sim IG\left(\frac{n_C}{2},\frac{n_C\overline{s}_C^{2(m)}}{2}\right)$, where $\overline{s}_C^{2(m)}=\frac{\sum_{j=1}^{n_C}\left(x_{Cj}-\mu_C^{(m)}\right)^2}{n_C}$

Remove burn in and then compute d_H numerically using the two posterior MCMC samples $\mu_{C_0}^{(m)}$ and $\mu_C^{(m)}$, for $m = B+1,...,M$. Then compute a from (10).

7. Gibbs Sampling from the power posterior $\mu_C \mid \mathbf{x}_C, \mathbf{x}_{C_0}, \sigma_C^2, \sigma_{C_0}^2$

   For $m = 1,...,M$, draw

- $\mu_{C_0}^{(m)}$ from $\mu_{C_0} \mid \mathbf{x}_{C_0}, \sigma_{C_0}^{2(m-1)} \sim N\left(\overline{x}_{C_0}, \frac{\sigma_{C_0}^{2\ (m-1)}}{n_{C_0}}\right)$
- $\sigma_{C_0}^{2(m)}$ from $\sigma_{C_0}^2 \mid \mathbf{x}_{C_0}, \mu_{C_0}^{(m)} \sim IG\left(\frac{n_{C_0}}{2}, \frac{n_{C_0}\overline{s}_{C_0}^{2(m)}}{2}\right)$
- $\mu_{C,a}^{(m)}$ from the power posterior $\mu_C \mid \mathbf{x}_C, \mathbf{x}_{C_0}, \sigma_{C,a}^{2(m-1)}, \sigma_{C_0}^{2(m-1)} \sim N\left(\widetilde{\mu}_{C,a}^{(m-1)}, \widetilde{\sigma}_{C,a}^{2(m-1)}\right)$, where $\widetilde{\mu}_{C,a}^{(m-1)}$ and $\widetilde{\sigma}_{C,a}^{2(m-1)}$ are given by (8) with $\overline{X}_C$ and $\overline{X}_{C_0}$ replaced by $\overline{x}_C$ and $\overline{x}_{C_0}$
- $\sigma_{C,a}^{2(m)}$ from $\sigma_C^2 \mid \mathbf{x}_C, \mu_{C,a}^{(m)} \sim IG\left(\frac{n_C}{2}, \frac{n_C\overline{s}_{C,a}^{2(m)}}{2}\right)$, where $\overline{s}_{C,a}^{2(m)} = \frac{\sum_{j=1}^{n_C}\left(x_{Cj} - \mu_{C,a}^{(m)}\right)^2}{n_C}$

8. Compute $\mu_E^{(m)} - \mu_{C,a}^{(m)}$, for each $m = 1,...,M$.
9. Remove burn-in and find the $\frac{\alpha}{2}th$ quantile $\widetilde{L}_{E-C}$ of the empirical distribution of $\mu_E^{(m)} - \mu_{C,a}^{(m)}$, for $m = B+1,...,M$.
10. If $\widetilde{L}_{E-C} > -\delta$ reject H_0.

3 | SIMULATION

3.1 | Frequentist size and power

Regulatory agencies require to analyze frequentist size and power of any new statistical test in clinical trials.[34,35] In the following we will assume that historical data are fixed, whereas we consider current trial data $(\mathbf{X}_E, \mathbf{X}_C)$ as random variables. The power function of the NI test based on the rejection rule (3) is $\eta(\theta) = \mathbb{P}(L_{E-C} > -\delta)$, where $\mathbb{P}(\cdot)$ is the probability computed with respect to the joint distribution of $(\mathbf{X}_E, \mathbf{X}_C)$. To evaluate size and power we fix a design value $\theta^\star = \mu_E^\star - \mu_C^\star = -\delta + \xi$, $\xi \in \mathbb{R}$, so that the power function can be expressed as follows

$$\eta(\xi) = \begin{cases} \alpha(\xi) & \xi \le 0 \\ 1 - \beta(\xi) & \xi > 0 \end{cases},$$

where $\alpha(\xi)$ and $\beta(\xi)$ are the type-I and type-II probability error functions respectively. The size of the test is then given by $\alpha = \alpha(0)$, whereas for $\xi > 0$ we obtain the power. The simulation steps required for evaluating size and power of the test are the following.

1. Steps 1–3 as in the algorithm of Section 2.3.
2. Set $U = 10000$. In current experiment, for $u = 1,...,U$, independently draw n_C iid values $x_{Cj}^{(u)}$ from $N\left(\mu_C^\star, \sigma_C^2\right)$ and n_E iid values $x_{Ej}^{(u)}$ from $N\left(\mu_E^\star, \sigma_E^2\right)$, where $\mu_E^\star = \mu_C^\star - \delta + \xi$, $\xi \ge 0$.
3. Set $\mathbf{x}_E = \mathbf{x}_E^{(u)}$ and $\mathbf{x}_C = \mathbf{x}_C^{(u)}$, where $\mathbf{x}_i^{(u)} = \left(x_{i1}^{(u)}, ..., x_{in_i}^{(u)}\right)$, for $i = E, C$.
4. Repeat U times steps 5–9 as in the algorithm of Section 2.3. Compute $\widetilde{L}_{E-C}^{(u)}$, $u = 1,...,U$.
5. Approximate $\eta(\xi)$ as the fraction of $\widetilde{L}_{E-C}^{(u)} > -\delta$, that is, the empirical size (if $\xi = 0$) or the empirical power (if $\xi > 0$).

Using this algorithm, we perform a simulation study to analyze frequentist properties of our procedure. In addition we compare our results to the ones obtained using the frequentist approach. Simulations are conducted under two scenarios.

Scenario 1: Following Gamalo et al.[10] (Section 4.2), let $\mu_{C_0} = 1$, $\mu_{P_0} = 0$, $\sigma_{C_0} = \sigma_{P_0} = 1$, $n_{C_0} = n_{P_0} = 600$. Historical data $\mathbf{x}_{C_0}$ and $\mathbf{x}_{P_0}$ are simulated at the beginning and then considered as fixed. For step 2 in the above described

TABLE 1 Scenario 1: Size of NI Bayesian and frequentist tests.

(A) Known variances					
	Full borrowing	**Partial borrowing**		**No borrowing**	**Frequentist case**
κ	1	1	0.8	–	–
a	1	0.940	0.752	0	–
λ					
0.0	0.0278	0.0229	0.0256	0.0259	0.0246
0.30	0.0292	0.0247	0.0241	0.0239	0.0248
0.60	0.0258	0.0229	0.0256	0.0259	0.0256
0.90	0.0300	0.0247	0.0241	0.0239	0.0250
(B) Unknown variances					
	Full borrowing	**Partial borrowing**		**No borrowing**	**Frequentist case**
κ	1	1	0.8	–	–
a	1	0.329	0.263	0	–
λ					
0.0	0.0310	0.0265	0.0257	0.0237	0.0253
0.30	0.0271	0.0261	0.0260	0.0219	0.0283
0.60	0.0281	0.0238	0.0237	0.0226	0.0278
0.90	0.0308	0.0238	0.0261	0.0255	0.0259

algorithm, we consider the following design values: $\mu_C^\star = 1$ and $\mu_E^\star = \mu_C^\star - \delta + \xi$. We set $\sigma_C = \sigma_E = 1$ and $n_C = n_E = 30$. In Table 1 we compute the size α for several values of λ and for different choices of a, assuming both known and unknown variances. In Figure 1, we fix $\lambda = 0.3$ and show the power functions for increasing values of ξ, while considering full, partial and no borrowing of historical data as well as the power for the frequentist test. Known and unknown variance cases are displayed in the two panels. Finally, Figure 2 shows expected interval estimates of $\theta = \mu_E - \mu_C$ for different values of λ, in the unknown variance case, both under the Bayesian and the frequentist approaches. Note that the expected intervals depend on λ and, consequently on δ, because the design value is $\theta^\star = -\delta + \xi$. The main comments are the following.

i. All the approaches preserve type-I error for different values of the NI margin (Table 1).
ii. In the partial borrowing case the power prior parameter a is much smaller in Table 1(B) than in Table 1(A). This is due to the additional uncertainty considered under the unknown variance assumption that induces a lower level of compatibility (larger d_H) between current and historical data. Nevertheless, the different values of a do not seem to influence the sizes α that are very close in both known and unknown variance cases.
iii. Figure 1 shows that full borrowing ($a = 1$) and partial borrowing (with $\kappa = 0.8$) approaches are consistently more powerful than the ones in which historical data are not exploited ($a = 0$, frequentist approach).
iv. Figure 2 compares expected lengths of interval estimates for $\mu_E - \mu_C$ for several λ values. First, we note that the no borrowing Bayesian case produces intervals substantially coincident with confidence intervals. Second, whenever past data are borrowed in the current study, the Bayesian methodology produces the shortest interval estimates: indeed, as already discussed, the use of prior knowledge in the current study may produce more accurate estimates. Moreover, the higher the level of borrowing the shorter the expected interval lengths. Finally note that as λ gets larger, since $\theta^\star$ is an increasing function of λ, expected intervals are shifted towards larger values.

Scenario 2: In the previous scenario consistency between $\theta^\star$ and historical control data implies that borrowing is always advantageous. However, it is also important to evaluate the procedure when the use of historical data might lead to an inflation of type-I error probability. Specifically, we now explore the behaviour of the proposed method in terms of type-I error probability when the true $\theta^\star$ belongs to the null hypothesis set (i.e., $\xi = 0$), but historical data support the alternative hypothesis. Keeping the assumptions of Scenario 1, we let vary $\mu_{C_0} = 0.9, 0.8, 0.7$ and $\kappa = 0.8, 0.6, 0.4, 0.2$

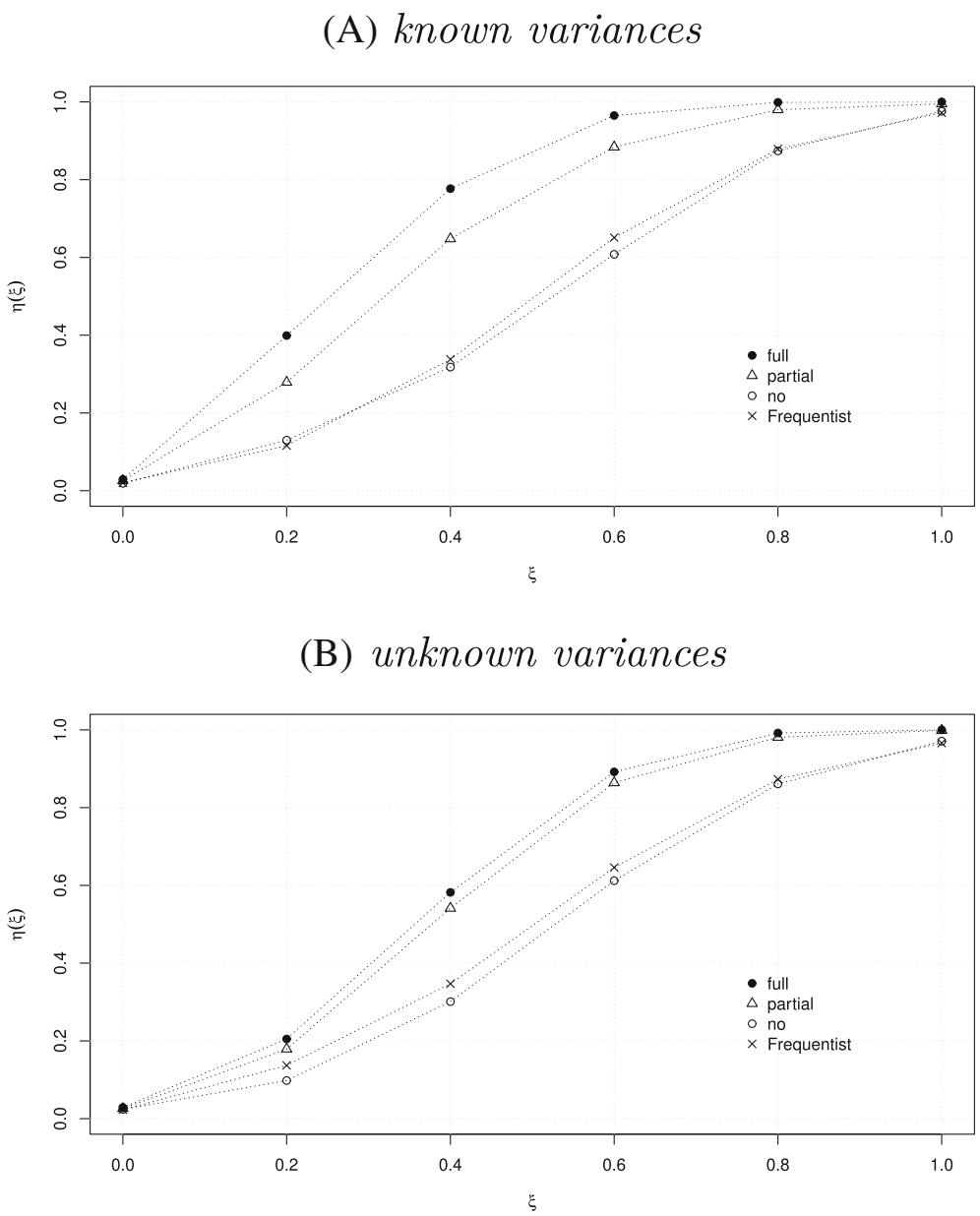

FIGURE 1 Scenario 1: Power function $\eta(\xi)$ for full, partial, no borrowing and frequentist cases.

and we set $\lambda = 0.3$. We investigate how the proposed method preserves the size of the test when $\theta^\star$ belongs to the null hypothesis. In Table 2, values of $1 - d_H(\cdot)$ consistently reduce as μ_{C_0} is smaller and smaller than μ_C. However, for $\mu_{C_0} = 0.8, 0.7$, the size α is substantially preserved only when small values of κ are considered.

3.2 | Comparisons with other dynamic power priors

In this section, the frequentist properties of the Hellinger distance approach (`hellinger`) are compared to the ones of the methods obtained using other similarity measures proposed in the literature. In particular, we select the following competitors.

- `pvalue`: p-value based method (Liu et al.[13]). The power parameter is $a = \exp\left[\frac{\kappa_1}{1-p}\ln(1-p)\right]$ where p is the p-value of the hypothesis test $H_0 : |\mu_{C_0} - \mu_C| > \varepsilon$ *vs* $H_1 : |\mu_{C_0} - \mu_C| < \varepsilon$, where $\varepsilon > 0$ is a pre-specified margin for equivalence (here $\varepsilon = 0.01$) and $\kappa_1 \geq 1$ is an arbitrary coefficient that penalizes the borrowing.

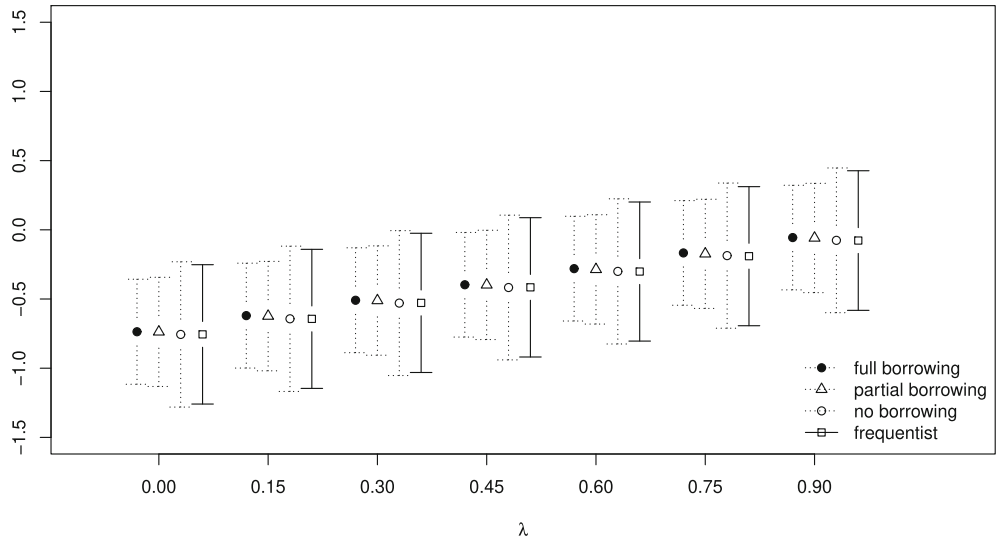

FIGURE 2 Scenario 1: Average 95% credible intervals intervals of $\mu_E - \mu_C$, for full borrowing, partial borrowing with $\kappa = 1$, no borrowing and average 95% confidence interval of $\mu_E - \mu_C$ (frequentist case).

TABLE 2 Scenario 2: Size of NI tests for different values of κ and μ_{C_0}. The true values of $1 - d_H$ are reported in the right column.

μ_{C_0}	$\kappa = 0.8$	$\kappa = 0.6$	$\kappa = 0.4$	$\kappa = 0.2$	$1 - d_H$
0.9	0.0400	0.0375	0.0335	0.0311	0.374
0.8	0.0853	0.0744	0.0631	0.0493	0.286
0.7	0.1440	0.1221	0.1022	0.0720	0.187

- `epw`: equivalence probability weight method (Bennet et al.[26]). The coefficient a is the probability that the random variable $\mu_C - \mu_{C_0} \mid \mathbf{X}_C, \mathbf{X}_{C_0}$ lies within the pre-specified interval of equivalence $(-\gamma, \gamma)$ (here, $\gamma = 0.08$).
- `overlap`: overlap method (Shi et al.[27]). The power parameter is $a = A^{\kappa_2}$, where A is the overlap area under the two distributions $\overline{X}_C \mid \mu_C, \sigma_C^2 \sim N\left(\mu_C, \frac{\sigma_C^2}{n_C}\right)$ and $\overline{X}_{C_0} \mid \mu_{C_0}, \sigma_{C_0}^2 \sim N\left(\mu_{C_0}, \frac{\sigma_{C_0}^2}{n_{C_0}}\right)$ and $\kappa_2 \geq 1$ is an arbitrary calibration parameter.
- `overlap2`: modified overlap method (Shi et al.[27]). The coefficient is $a = (A \cdot p)^{\kappa_2}$, where p is the p-value of the two-sided hypothesis test $H_0 : \mu_C = \mu_{C_0}$ *vs* $H_1 : \mu_C \neq \mu_{C0}$ and A and $\kappa_2 \geq 1$ are defined as before.

Since κ, κ_1 and κ_2 have different interpretations across the selected approaches, we fix $\kappa = \kappa_1 = \kappa_2 = 1$ to ensure a fair comparison. Note that `overlap`, `overlap2` and `hellinger` do not require the specification of further parameters; whereas `pvalue` and `epw` depend respectively on the margin of equivalence ε and on the equivalence bound γ.

Figure 4 displays the true values of the similarity measures when μ_{C_0} varies in the interval $\mu_C \pm 1$. All methods imply similar values of a when $\mu_{C_0} = \mu_C$. The values of a reach 1 in the case of maximum compatibility, that is, when $n_{C_0} = n_C$ (see right panel). The larger $|\mu_{C_0} - \mu_C|$, the smaller a in each panel. The methods depending on p-values (`overlap2`, `pvalue`) return values of a close to 0 even for very small differences between μ_{C_0} and μ_C. All other three methods (`epw`, `overlap`, `hellinger`) allow for larger and similar levels of borrowing even when compatibility between current and historical control data is not stringent.

Let us now discuss the simulation results. Table 3 reports the size of the test and the average values of a across simulations for all the methods under comparison. Figure 3 shows the power of the tests when $\mu_{C_0} = 0.9$. The main comments are the following.

i. Table 3 shows three different values of μ_{C_0} that induce three different levels of disagreement between current and historical data. In general an increasing amount of borrowing yields inflation in type-I error rate. This is more and more evident when the values of μ_{C_0} get smaller (i.e. more in contrast with the current control data).

TABLE 3 Size of NI tests and average a values across simulations (in brackets) for different similarity measures, for $\mu_C = 1$.

μ_{C_0}	0.9		0.8		0.7	
Full	**0.0555**	**(1)**	**0.1399**	**(1)**	**0.2839**	**(1)**
`hellinger`	0.0454	(0.314)	0.0899	(0.272)	0.1539	(0.212)
`overlap`	0.0450	(0.276)	0.0838	(0.232)	0.1368	(0.171)
`epw`	0.0424	(0.228)	0.0750	(0.188)	0.1171	(0.135)
`overlap2`	0.0424	(0.159)	0.0591	(0.119)	0.0774	(0.073)
`pvalue`	0.0326	(0.051)	0.0406	(0.036)	0.0418	(0.019)
`no`	0.0250	(0)	0.0250	(0)	0.250	(0)

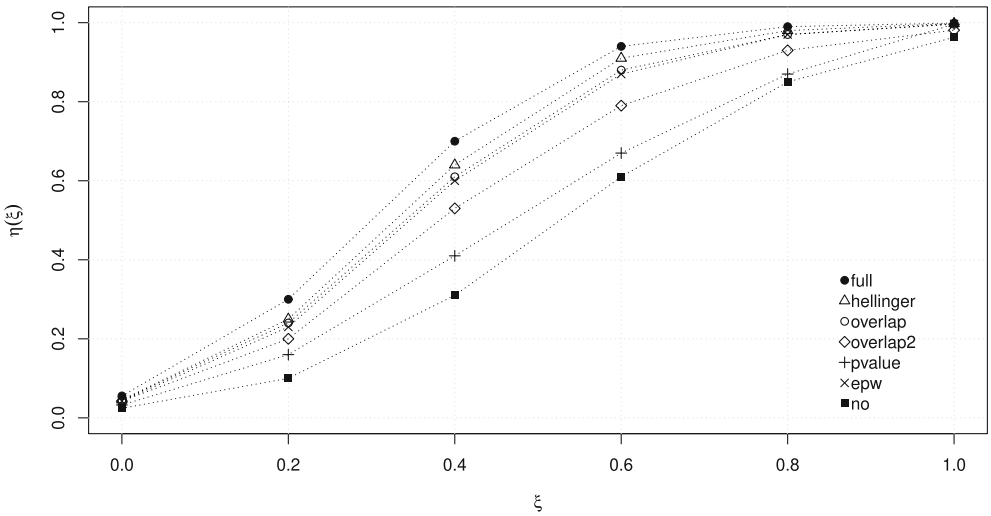

FIGURE 3 Power function $\eta(\xi)$ for full, partial and no borrowing when $\mu_{C_0} = 0.9$ and $\kappa = \kappa_1 = \kappa_2 = 1$.

ii. As a consequence of the behavior of a shown in Figure 4, the methods based on p-values (`pvalue`, `overlap2`) ensure a better preservation of the type-I error rate but, at the same time, these approaches are less powerful than the others (Figure 3).
iii. `hellinger`, `overlap` and `epw` show similar performances both in terms of size and power.
iv. In general, full and partial borrowing methods are more powerful than the no borrowing one.

4 | APPLICATION TO DRUG DATA

In this section, we consider the Example in Gamalo et al.[10] (see Section 4.3, pp. 232-236), based on a NI trial for an iron-containing drug for intravenous administration, which is indicated for the treatment of iron deficiency anemia in adult patients with chronic kidney disease. The parameter of interest in this study is the mean change in hemoglobin (Hgb) from baseline to day 35. The drug is dosed at 510 mg per injection for a total of two daily injections. We consider a NI trial to test whether the effect of a new dosage (i.e., 255 mg four times daily in the experimental arm E) is not inferior to that of the approved dosage (in the control arm C) by a specified margin. First, we assume the setup S1, as in Gamalo et al.,[10] based on a previous placebo-controlled trial: the observed mean change in Hgb in the placebo arm is $\overline{x}_{P_0} = 0.16$ ($s_{P_0} = 1.02$) and $\overline{x}_{C_0} = 0.82$ ($s_{C_0} = 1.24$) in the 510 mg arm, with $n_{P_0} = 76$ and $n_{C_0} = 228$. Assuming non informative priors for μ_{P_0} and μ_{C_0}, we obtain $L_{C_0-P_0} = 0.377$, then $\delta = (1-\lambda) \times 0.377$. In the current trial, the mean change in Hgb in the 510 mg arm is $\overline{x}_C = 0.71$ ($s_C = 1.00$) and $\overline{x}_E = 0.87$ ($s_E = 1.14$) in the 255 mg arm, with $n_C = 64$ and $n_E = 62$. In addition, we consider two different historical setups: in setup S2 we consider $\overline{x}_C = \overline{x}_{C_0} = 0.71$, that is, strong

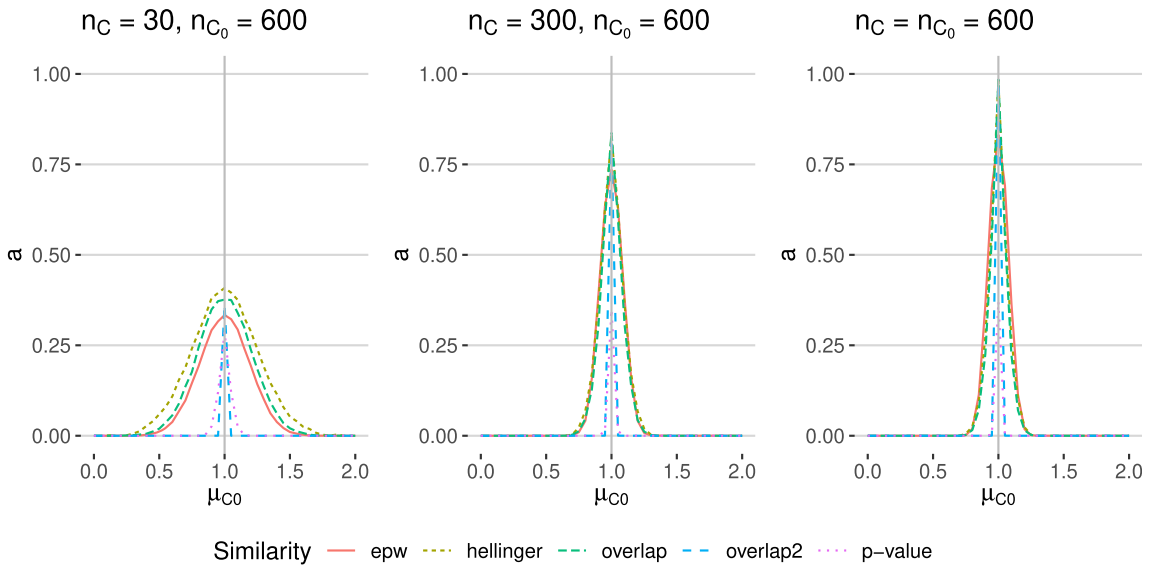

FIGURE 4 True values of different similarity measures as μ_{C_0} varies and for different choices of n_C. We set $\mu_C = 1$, $\kappa = \kappa_1 = \kappa_2 = 1$; the grey vertical line represents complete agreement between historical and current control data.

compatibility between current and historical control data; whereas in setup S3, we set $\overline{x}_{C_0} = 0.67$, that is, we assume that in the current trial the active control performs better than expected from previous data.

Under the three setups S1, S2 and S3, we compare results of our Bayesian procedure, considering full, partial and no borrowing. Table 4 reports, for different values of λ, the posterior probability $\overline{\mathbb{P}}(H_1)$ given in Equation (12), which measures the evidence in favour of H_1, and the corresponding decision for $1 - \frac{\alpha}{2} = 0.975$, (1 = reject H_0, 0 = accept H_0). Note that decisions based on frequentist confidence intervals are not included in the table, as they essentially coincide with decisions obtained under the Bayesian no borrowing case.

Here are the main comments.

i. For each given λ, the values of δ decrease as we move from setup S1 to S2 and S3. In fact, the larger $\overline{x}_{C_0}$ the larger $L_{C_0 - P_0}$ and $|\delta|$.
ii. Across all setups, as λ increases $|\delta|$ gets closer and closer to 0 and $\overline{P}(H_1)$ decreases.
iii. In setup S1 the control treatment performs worse than expected ($\overline{x}_C = 0.71 < 0.82 = \overline{x}_{C_0}$, $a = 0.593$): hence, ignoring historical information ($a = 0$) the difference between experimental and control treatments in the current trial is more remarkable which results in values of $\overline{P}(H_1)$ slightly larger than those obtained for full and partial borrowing. Moreover, the same final decisions are obtained in the full and partial borrowing cases and in the no borrowing case ($a = 0$) – which yields the same decisions of the frequentist approach – except for $\lambda = 0.4$, as already shown in Gamalo et al.[10]
iv. In setup S2, there is a larger level of compatibility between historical and current control data ($a = 0.798$). Since the values of $|\delta|$ are smaller than in the previous scenario, the values of $\overline{P}(H_1)$ are smaller for $a = 0$. However, partial and full borrowing imply a more and more evident difference between experimental and control treatments which favor H_1 (larger and larger values of $\overline{P}(H_1)$). Full and partial borrowing yield different decisions for $\lambda = 0.4$ only, whereas in the no borrowing case H_0 is rejected only for $\lambda = 0$ and $\lambda = 0.1$.
v. Setup S3 shows the opposite situation with respect to setup S1: current control data are closer to experimental data than expected ($\overline{x}_C = 0.71 > 0.67 = \overline{x}_{C_0}$) and therefore we obtain smaller values of $\overline{P}(H_1)$ than in setup S1 and S2, for $a = 0$. When comparing the $a = 0$ case to the borrowing cases, we draw conclusions similar to those of comment (*iv*). Note that in the no borrowing case H_0 is rejected only for $\lambda = 0$.

Summarizing, if the performance of the current control treatment improves with respect to the historical data (moving from S1 to S3), borrowing widens the difference between control and experimental treatments, making more likely rejection of H_0.

TABLE 4 Chronic kidney disease example – Values of $\overline{P}(H_1)$ and corresponding decisions using full, partial and no borrowing approaches, for several values of λ.

Setup S1							
		$a=0$		$a=1-d_H=0.593$		$a=1$	
λ	$\lvert\delta\rvert$	$\overline{P}(H_1)$	**Decision**	$\overline{P}(H_1)$	**Decision**	$\overline{P}(H_1)$	**Decision**
0.0	0.377	0.997	1	0.998	1	0.998	1
0.1	0.339	0.995	1	0.995	1	0.995	1
0.2	0.301	0.991	1	0.991	1	0.991	1
0.3	0.264	0.985	1	0.984	1	0.983	1
0.4	0.226	0.976	1	0.972	0	0.971	0
0.5	0.188	0.963	0	0.955	0	0.953	0
0.6	0.151	0.945	0	0.929	0	0.925	0
0.7	0.113	0.920	0	0.893	0	0.886	0
0.8	0.075	0.887	0	0.846	0	0.835	0
0.9	0.038	0.845	0	0.787	0	0.771	0
1.0	0.0	0.795	0	0.716	0	0.696	0
Setup S2							
		$a=0$		$a=1-d_H=0.798$		$a=1$	
λ	$\lvert\delta\rvert$	$\overline{P}(H_1)$	**Decision**	$\overline{P}(H_1)$	**Decision**	$\overline{P}(H_1)$	**Decision**
0.0	0.266	0.986	1	0.995	1	0.996	1
0.1	0.239	0.980	1	0.992	1	0.993	1
0.2	0.213	0.971	0	0.988	1	0.989	1
0.3	0.186	0.962	0	0.982	1	0.983	1
0.4	0.160	0.950	0	0.974	0	0.975	1
0.5	0.133	0.934	0	0.962	0	0.964	0
0.6	0.106	0.914	0	0.947	0	0.949	0
0.7	0.080	0.891	0	0.927	0	0.930	0
0.8	0.053	0.863	0	0.902	0	0.905	0
0.9	0.027	0.831	0	0.871	0	0.874	0
1.0	0.0	0.794	0	0.834	0	0.837	0
Setup S3							
		$a=0$		$a=1-d_H=0.763$		$a=1$	
λ	$\lvert\delta\rvert$	$\overline{P}(H_1)$	**Decision**	$\overline{P}(H_1)$	**Decision**	$\overline{P}(H_1)$	**Decision**
0.0	0.227	0.977	1	0.994	1	0.995	1
0.1	0.204	0.969	0	0.991	1	0.992	1
0.2	0.181	0.960	0	0.987	1	0.988	1
0.3	0.159	0.949	0	0.981	1	0.983	1
0.4	0.136	0.936	0	0.974	0	0.977	1
0.5	0.113	0.920	0	0.964	0	0.968	0
0.6	0.091	0.901	0	0.952	0	0.957	0
0.7	0.068	0.879	0	0.937	0	0.942	0
0.8	0.045	0.854	0	0.918	0	0.924	0
0.9	0.023	0.826	0	0.896	0	0.902	0
1.0	0.000	0.795	0	0.869	0	0.876	0

5 | CONCLUSIONS

The Bayesian approach has a practical utility in NI trials since it allows to involve past information on the active control that is often available from historical data. This may also represent an advantage in terms of frequentist power. These benefits are obtained as long as historical information is wisely employed and, potentially, discounted in the presence of heterogeneity with respect to current trial data.

In this article we consider a Bayesian approach to NI trials for normal means with unknown variance that exploits historical information on the active control arm for the construction of the prior distribution of μ_C. We consider a dynamic power prior based on the Hellinger distance.

Our analysis follows some key aspects of Gamalo et al.[10]: (i) the choice of the NI margin; (ii) the comparison with respect to the frequentist NI test; (iii) the use of Gibbs sampling for dealing with the unknown variance case. However, our method presents several originalities with respect to Gamalo et al.'s approach. Specifically, whereas they consider full incorporation of historical data for the construction of the prior distribution of μ_C, we adopt a dynamic power prior in order to calibrate the level of borrowing. As a consequence, the simulation procedure of Section 4.1 in Gamalo et al.[10] is modified in our algorithms to compute the power prior coefficient (see step 6 of the algorithm of Section 2.3) and to draw an MCMC sample from the power posterior of μ_C (see step 7) accordingly. Furthermore we assess the impact of different choices of a on the expected length of credible intervals.

One central aspect of the present paper is to study the frequentist properties of the proposed Bayesian methodology, as required by regulatory agencies for new statistical procedures to be introduced in clinical practice. Simulations of Section 3.1 (Scenario 1) show that the proposed method can be more powerful than the frequentist approach while retaining type-I error (size of the test): when historical information is fully or partially borrowed larger values of the power functions are obtained uniformly. We also perform a sensitivity analysis with respect to alternative historical setups (see Scenario 2). We explore the behaviour of the proposed method in terms of type-I error probability when the true $\theta^\star$ belongs to the null hypothesis set, but historical data support the alternative hypothesis. In particular, the proposed approach still yields acceptable values of the type-I error probabilities, especially when κ is small. In addition, in Section 3.2 we compare our method with other existing dynamic power prior approaches: our conclusions are consistent with those obtained using the epw method introduced by Bennet et al.[26] and the overlap method proposed by Shi et al.[27]

Finally, in the real data application of Section 4 based on the original example of Gamalo et al. the proposed and the frequentist methods bring to the same decisions almost all the times, in contrast with the full borrowing methodology. However, if we consider alternative historical setups, we show that conclusions depend on compatibility between historical data and $\theta^\star$. Specifically, if the current control treatment performs better than the historical control, borrowing past information induces a wider difference between control and experimental treatments, so that rejection of H_0 becomes more likely.

Given the simulations and example discussed in the present article, our opinion is that the power prior based on the Hellinger distance has specific advantages over other alternative borrowing methods, such as ease of implementation, interpretability and availability in closed-form for the most common distributions.

Here are some ideas for further developments.

i. In simulation studies, one might take into account potential randomness of historical control data to assess its impact on the performances in terms of size and power.
ii. The methodology can be extended to other models and to non-conjugate priors, resorting to numerical integration or Monte Carlo approximation whenever the Hellinger distance is not available analytically.
iii. As sketched by Gamalo et al.[10] the proposed method can be extended to three-arms NI trials that consist of placebo, reference and experimental treatment. This structure allows one to test simultaneously superiority of the reference over the placebo and non-inferiority of experimental treatment over the reference. In this setup, Bayesian methods ensure the same advantage we have stressed in the present paper, as well as in any active-control trial, that is the possibility of exploiting substantial historical information regarding both placebo and reference treatment. See also Ghosh et al.,[36] Tang et al.,[37] Tang et al.[38] and Ghosh et al.[39]

CONFLICT OF INTEREST STATEMENT

No potential competing interest was reported by the authors.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Francesco Mariani https://orcid.org/0000-0002-0239-1741
Stefania Gubbiotti https://orcid.org/0000-0002-1108-0953

REFERENCES

1. Simon R. Are placebo-controlled clinical trials ethical? Editorial. *Ann Intern Med*. 2000;133:474-475.
2. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013; 310(20):2191-2194.
3. Fleming TR, Krause PR, Nason M, Longini IM, Henao-Restrepo AM. COVID-19 vaccine trials: the use of active controls and non-inferiority studies. *Clin Trials*. 2021;18(3):335-342. doi:10.1177/1740774520988244
4. Jin M, Feng D, Liu G. Bayesian approaches on borrowing historical data for vaccine efficacy trials. *Stat Biopharm Res*. 2020;12(3): 284-292.
5. Nauta J. Vaccine equivalence and non-inferiority trials. *Statistics in Clinical and Observational Vaccine Studies*. Springer; 2020.
6. Wang WWB, Mehrotra DV, Chan ISF, Heyse JF. Statistical considerations for non-inferiority/equivalence trials in vaccine development. *J Biopharm Stat*. 2006;16(4):429-441.
7. Riechelmann RP, Alex A, Cruz L, Bariani GM, Hoff PM. Non-inferiority cancer clinical trials: scope and purposes underlying their design. *Ann Oncol*. 2013;24:1942-1947.
8. Gamalo MA, Tiwari RC, La Vange LM. Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharm Stat*. 2014;13:25-40.
9. Chen MH, Ibrahim JG, Lam P, Yu A, Zhang Y. Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics*. 2011;67:1163-1170.
10. Gamalo MA, Wu R, Tiwari RC. Bayesian approach to non-inferiority trials for normal means. *Stat Methods Med Res*. 2016;25(1):221-240.
11. Gamalo-Siebers M, Gao A, Lakshminarayanan M, et al. Bayesian methods for the design and analysis of noninferiority trials. *J Biopharm Stat*. 2016;26(5):823-841.
12. Li W, Chen MH, Wang X, Dey DK. Bayesian design of non-inferiority clinical trials via the Bayes factor. *Stat Biosci*. 2018;10(2):439-459.
13. Liu GF. A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. *Pharm Stat*. 2018;17:61-73.
14. Psioda MA, Ibrahim JG. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*. 2019;20(3): 400-415.
15. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13(1):41-54. doi:10.1002/pst.1589
16. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7:5-18.
17. Hobbs BP, Carlin BP, Mandrekar SJ, Sergent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011;67:1047-1056.
18. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15:46-60.
19. De Santis F. Power priors and their use in clinical trials. *Am Stat*. 2006;60(2):122-129.
20. De Santis F. Using historical data for Bayesian sample size determination. *J R Stat Soc Ser A*. 2007;170:95-113.
21. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015;34(28):3724-3749.
22. Pan H, Yuan Y, Xia J. A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Appl Stat*. 2017;66(5):979-996.
23. Gravestock I, Held L. Adaptive power priors with empirical Bayes for clinical trials. *Pharm Stat*. 2017;16:349-360.
24. Nikolakopoulos S, van der Tweel I, Roes KCB. Dynamic borrowing through empirical power priors that control type I error. *Biometrics*. 2018;74:874-880.
25. Ollier A, Morita S, Ursino M, Zohar S. An adaptive power prior for sequential clinical trials- application to bridging studies. *Stat Methods Med Res*. 2020;29:2282-2294.
26. Bennet M, White S, Best N, Mander A. A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: a comparison to standard methods. *Pham Stat*. 2021;20(3):462-484.
27. Shi Y, Li W, Liu GF. A novel power prior approach for borrowing historical control data in clinical trials. *Stat Methods Med Res*. 2023; 32(3):493-508.
28. Gravestock I, Held L. Power priors based on multiple historical studies for binary outcomes. *Biomet J*. 2019;61:1201-1218.
29. Box GEP. Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J R Stat Soc (Ser A)*. 1980;143:383-430.
30. De Santis F, Gubbiotti S. A method for incorporating historical information in non-inferiority trials. Book of short papers SIS 2021 – (9788891927361). 2021.
31. De Santis F, Gubbiotti S. A aynamic power prior for Bayesian non-inferiority trials. In: Salvati N, Perna C, Marchetti S, Chambers R, eds. *Studies in Theoretical and Applied Statistics. SIS 2021*. Springer Proceedings in Mathematics & Statistics, Vol 406. Springer, Cham; 2022:15-30. https://doi.org/10.1007/978-3-031-16609-9_2

32. De Santis F, Gubbiotti S. Borrowing historical information for non-inferiority trials on Covid-19 vaccines. *Int J Biostat*. 2022;19:177-189. doi:10.1515/ijb-2021-0120
33. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, https://www.R-project.org/; 2020.
34. CDER/CBER/FDA Guidance for Industry. Non-Inferiority Clinical Trials. 2016 https://www.fda.gov/media/78504/download
35. CDHR/FDA Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Guidance for industry and FDA staff. 2010 https://www.fda.gov/media/71512/download
36. Ghosh P, Nathoo F, Gönen M, Tiwari RC. Assessing noninferiority in a three-arm trial using the Bayesian approach. *Stat Med*. 2011;30:1795-1808.
37. Tang NS, Yu B, Tang ML. Testing non-inferiority of a new treatment in three-arm clinical trials with binary endpoints. *BMC Med Res Methodol*. 2014;14:134.
38. Tang N, Yu B. Bayesian sample size determination in a three-arm non-inferiority trial with binary endpoints. *J Biopharm Stat*. 2022;32(5):768-788.
39. Ghosh S, Paul E, Chowdhury S, Tiwari RC. New approaches for testing non-inferiority for three arm trials with Poisson distributed outcomes. *Biostatistics*. 2022;23:136-156.

SUPPORTING INFORMATION

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

How to cite this article: Mariani F, De Santis F, Gubbiotti S. A dynamic power prior approach to non-inferiority trials for normal means. *Pharmaceutical Statistics*. 2024;23(2):242-256. doi:10.1002/pst.2349

APPENDIX A

Given $\theta \sim N(\mu_1, \sigma_1^2)$ and $\theta \sim N(\mu_2, \sigma_2^2)$. Let $\pi_1(\theta|\mu_1, \sigma_1^2)$ and $\pi_2(\theta|\mu_2, \sigma_2^2)$ be the respective density functions. The Hellinger distance between these two densities is:

$$d_H\left[\pi(\theta|\mu_1, \sigma_1^2), \pi(\theta|\mu_2, \sigma_2^2)\right] = \left(1 - \int_{\mathbf{R}} \sqrt{\pi_1(\theta|\mu_1, \sigma_1^2)\pi_2(\theta|\mu_2, \sigma_2^2)}d\theta\right)^{\frac{1}{2}},$$

where

$$\int_{\mathbf{R}} \sqrt{\pi_1(\theta|\mu_1, \sigma_1^2)\pi_2(\theta|\mu_2, \sigma_2^2)}d\theta = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{-\frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right\}.$$