Research Article

Statistics in Medicine

Received 13 October 2009, Accepted 21 June 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4025

Frequentist performance of Bayesian inference with response-adaptive designs

Bruno Lecoutre,[a]*† Gérard Derzko[b] and Khadija ElQasyr[a]

In controlled clinical trials, where minimizing treatment failures is crucial, response-adaptive designs are attractive competitors to 1:1 randomized designs for comparing the success rates φ_1 and φ_2 of two treatments. In these designs each new treatment assignment depends on previous outcomes through some predefined rule. Here Play-The-Winner (PW), Randomized Play-The-Winner (RPW), Drop-The-Loser, Generalized Drop-the-Loser and Doubly adaptive Biased Coin Designs are considered for new treatment assignments. As frequentist inference relies on complex sampling distributions in those designs, we investigate how Bayesian inference, based on two independent Beta prior distributions, performs from a frequentist point-of-view. Performance is assessed through coverage probabilities of interval estimation procedures, power and minimization of failure count. It is shown that Bayesian inference can be favorably compared to frequentist procedures where the latter are available. The power of response-adaptive designs is generally very close to the power of 1:1 randomized design. However, failure count savings are generally small, except for the PW and Doubly adaptive Biased Coin designs in particular ranges of the true success rates. The RPW assignment rule has the worst performance, while PW, Generalized Drop-the-Loser or Doubly adaptive Biased Coin Designs may outperform other designs depending on different particular ranges of the true success rates. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: adaptive design; Bayesian inference; coverage; power; treatment failures

1. Introduction

In response-adaptive designs, newly accrued subjects are assigned a treatment with a probability that is updated as a function of previous outcomes, according to some predefined rule. The intent is to favor the assignment of the 'most effective treatment' given available information. Even if these designs can only be reasonably implemented in specific clinical situations, they are attractive competitors to 1:1 randomized designs for comparing the success rates φ_1 and φ_2 of two treatments [1, 2].

Several such designs have been proposed in the past, and their asymptotic properties elicited and compared (see References [3–8]). Rosenberger and Hu [2] discussed the issue of maximizing power and minimizing treatment failures for fixed sample size response-adaptive designs with immediately available dichotomous responses. They concluded that the Drop-the-Loser design (DL) and a modification of the Doubly adaptive Biased Coin Design (DBCD) should be preferred over the Randomized Play-the-Winner (RPW) design. An advantage of the Doubly adaptive Biased Coin Design is that it can target any pre-specified allocation proportion, so that a theoretically optimal allocation can be considered. Moreover, their simulations showed that these procedures 'yield a modest reduction in expected treatment failures while preserving power over complete randomization'. Rosenberger and Hu's conclusions were based on the usual Z test statistic for the difference of proportions. It is well known that this test, as well as normal approximation confidence intervals, are not satisfactory for small or moderate-sized samples. An efficient frequentist inference confidence interval procedure must be constructed from a sampling distribution that depends on both the parameter of interest and the specific response-adaptive design being considered. This is a major source of difficulty with most response-adaptive designs, since sampling distributions are overly complicated (Play-The-Winner (PW) design is an exception to this respect, see [9]).

[a] *ERIS, Laboratoire de Mathématique Raphaël Salem, CNRS and Université de Rouen, Avenue de l'Université, BP 12, 76801 Saint-Etienne-du-Rouvray, France*

[b] *Sanofi-Aventis R&D, 371, rue du Professeur Joseph Blayac, 34184 Montpellier, France*

* *Correspondence to: Bruno Lecoutre, Laboratoire de Mathématique Raphaël Salem, CNRS and Université de Rouen, Avenue de l'Université, BP 12, 76801 Saint-Etienne-du-Rouvray, France.*

† *E-mail: bruno.lecoutre@univ-rouen.fr*

On the contrary, Bayesian inference for dichotomous responses is simple and applies equally well to any response-adaptive designs: for all of them indeed, the likelihood function is proportional to the likelihood function associated with the comparison of two independent Binomial proportions. Therefore, we develop mainly here the Bayesian approach to inference with two independent Beta prior distributions, and we investigate how Bayesian interval estimates perform from a frequentist point of view. We assume the same Jeffreys prior as for two independent Binomial proportions, so that the same procedure can be used for different designs. Performances in all the studied designs are then compared through coverage, power and overall treatment failure counts. For assumed fixed values φ_1^*, φ_2^* of the probabilities of success, the frequentist coverage of a Bayesian $100(1-\alpha)$ per cent credible interval procedure for the ratio φ_1/φ_2 (for instance) is defined by the sampling probability that this interval contains the 'true' value φ_1^*/φ_2^*. Its power is defined by the sampling probability (given φ_1^* and φ_2^*) that this interval does not contain one, i.e. the procedure rejects the null hypothesis of no difference.

The response-adaptive designs included in the study are the PW, RPW, Drop-the-Loser, Generalized Drop-the-Loser and Doubly adaptive Biased Coin designs. The specificity of the two latter designs is that they can target any pre-specified allocation proportion. We include the PW design, although this has become unusual in the literature. Indeed, Lecoutre and ElQasyr [8] demonstrated that the PW rule has optimal properties and always converges more rapidly than the RPW rule, sometimes notably so, with less variability. This warrants using it as a theoretical reference for other designs. Moreover, the sampling distribution of this design has been recently elicited by ElQasyr and Lecoutre [9]. This allows a complete comparison of frequentist and Bayesian approaches to analysis for at least one design. The practical use of PW design is briefly discussed in the conclusion.

The paper is organized as follows. In Section 2 we recall the different response-adaptive rules and illustrate their allocation proportions and convergence properties. In Section 3 we summarize available inference procedures for adaptive designs and present the Bayesian solution. In Section 4, we illustrate the good frequentist coverage probabilities of the Bayesian procedures, and show that, upon an appropriate choice of the prior, the Bayesian approach can be favorably compared to available frequentist procedures. In Section 5, we investigate power properties and demonstrate that the loss of power is mild when the response-adaptive designs are compared with the 1:1 randomized design. We conclude that it may be advantageous to use a response-adaptive design rather than a 1:1 randomized Design for preserving the power while minimizing the treatment failure count, and the numerical results provided here may help selecting the most appropriate response-adaptive design depending on the anticipated ranges of true success rates.

2. Description of treatment assignment rules

2.1. PW

The PW rule, introduced by Zelen [3], involves a dichotomous process. If the $(n-1)$-th subject is assigned treatment t then, if t is a success, the n-th subject is assigned treatment t and if t is a failure, the n-th subject is assigned the other treatment. In spite of its apparent determinism, the PW rule is a stochastic process, since it depends on the probabilities of success on each treatment.

2.2. RPW

The RPW rule, introduced by Wei and Durham [4], is generally presented as a Generalized Friedman's Urn (GFU) model. Before the n-th subject comes in, the urn contains (Y_{n-1}^1, Y_{n-1}^2) balls ((Y_0^1, Y_0^2) initially) that represent either treatments. A ball is drawn at random and replaced. The subject is assigned to the corresponding treatment (say t). Balls are added to the urn: if t is a success, u type t balls and v balls of the other type and if t is a failure, v type t balls and u balls of the other type.

The relatively poor performance of RPW rule is due to the fact that, whatever the process history, the total number of balls at step n is a constant, equal to $Y_n^1+Y_n^2=Y_0^1+Y_0^2+n(u+v)$ [8]. An attempt to relax this property was the Drop-the-Loser rule.

2.3. Drop-the-loser

The Drop-the-Loser (DL) rule, proposed by Ivanova [6], is also presented as a GFU model. Extra 'no treatment' balls ('immigration balls') are included, the initial urn composition being (Y_0^1, Y_0^2, Z_0). If an immigration ball is drawn no subject is treated and the ball is returned to the urn together with one ball of each treatment type. If a treatment t ball is drawn, then if t is a success, the ball is replaced (unchanged composition), and if t is a failure, the ball is not replaced (the number of balls is decreased by 1).

2.4. Generalized drop-the-loser

Zhang *et al.* [7] and Sun *et al.* [10] developed a very general model they called the Generalized Drop-the-Loser [GDL] rule. The aim of this rule is to target any pre-specified allocation proportion.

For practical purposes, these authors recommended the following particular rule, restricted here to two treatments. Before the n-th subject comes in, if an immigration ball is drawn, no subject is treated and the ball is returned to the urn together with a_n^1 and a_n^2 balls (instead of one for the DL rule), respectively for each treatment type. The total number of added balls $a_n^1+a_n^2$ is a fixed constant C. If a treatment ball is drawn, then the ball is never replaced, regardless of whether the response is a success or failure.

a_n^1 is chosen for adjusting the desired allocation function for treatment t^1. However, this allocation proportion is a function $\psi_1(\varphi_1, \varphi_2)$ of the unknown probabilities of success. The authors suggest to replace φ_1, φ_2 with their Bayesian estimates for a uniform prior distribution. Let us define $\hat{\psi}_1=\psi_1(\hat{\varphi}_1, \hat{\varphi}_2)$, where $\hat{\varphi}_t$ is 1+ the number of observed successes on treatment t divided by 2+ the number of observed outcomes on treatment t (before the nth subject comes in). Then, if $a_n^1=C\hat{\psi}_1$, the allocation proportion converges to the target proportion.

2.5. Doubly adaptive biased coin design

Eisele [5] introduced and studied the asymptotics of the Doubly adaptive Biased Coin Design, originally indicated when responses are independent random variables from standard exponential family. It was considered in the case of two treatments with immediate responses by Rosenberger and Hu [2]).

As the Generalized Drop-the-Loser, it can target any pre-specified allocation proportion $\psi_1(\varphi_1, \varphi_2)$ for treatment t^1. The probability of assigning patient n to treatment t^1 depends on two variables: the estimate $\hat{\psi}_1$ of the targeted proportion and the proportion of patients P_1 assigned to treatment t^1 before the nth subject comes in. We will consider here the Rosenberger and Hu formula (for a more general formulation, see [11])

$$\frac{\hat{\psi}_1(\hat{\psi}_1/P_1)^\gamma}{\hat{\psi}_1(\hat{\psi}_1/P_1)^\gamma+(1-\hat{\psi}_1)((1-\hat{\psi}_1)/P_1)^\gamma}$$

where γ is a nonnegative integer, which has a limited influence on the performance of the allocation process [2]. Zhang, Chan, Cheung and Hu [7] conclude from simulation results that 'the DBCD's performance is comparable to the GDL rule' (p. 396). However, if we look through their numerical tables (p. 394–395), we find some differences between the two rules, especially for the standard deviations of allocation proportions.

2.6. Targets

With PW, RPW and Drop-the-Loser, the asymptotic allocation proportion (target) for treatment t^1 is:

$$\psi_1(\varphi_1, \varphi_2)=\frac{1-\varphi_2}{1-\varphi_1+1-\varphi_2} \tag{1}$$

That rule will also be used for Generalized Drop-the-Loser (option GDL1) and Doubly adaptive Biased-coin (option DBCD1). We include in comparisons, for Generalized Drop-the-Loser (option GDL2) and Doubly adaptive Biased-coin (option DBCD2), the optimal target allocation proportion proposed by Rosenberger *et al.* [12]

$$\psi_1(\varphi_1, \varphi_2)=\frac{\sqrt{\varphi_1}}{\sqrt{\varphi_1}+\sqrt{\varphi_2}} \tag{2}$$

which, according to these authors, minimizes the expected number of failures under fixed variance of the estimator of the treatment difference.

2.7. Comparisons of allocation proportions

We have retained the parameter values recommended by the authors:

RPW: $Y_0^1=Y_0^2=1u=1v=0$
DL: $Y_0^1=Y_0^2=3Z_0=1$
GDL: $Y_0^1=Y_0^2=3Z_0=1C=2$.
DBCD: $\gamma=2$ (the first two patients being randomly assigned to treatments t^1 and t^2)

Table I gives the expectation and standard deviation of the number of subjects assigned to the less effective treatment, when the sample size is small $N=50$. As a general rule, all reported numerical values in this paper are based on the

Table I. Comparison of the proportions of allocation for five response-adaptive designs: expected number of subjects (standard deviation) allocated to the less effective treatment for $N=50$.

Target		$\frac{1-\varphi_2}{1-\varphi_1+1-\varphi_2}$						$\frac{\sqrt{\varphi_1}}{\sqrt{\varphi_1}+\sqrt{\varphi_2}}$		
		$N=50$						$N=50$		
φ_2	φ_1	PW	RPW	DL	GDL1	DBCD1	$N\to\infty$	GDL2	DBCD2	$N\to\infty$
0.10	0.30	21.9 (1.8)	22.1 (3.1)	22.2 (1.8)	22.6 (2.0)	22.1 (2.4)		21.3 (2.6)	18.9 (4.1)	
		43.8 per cent	*44.1 per cent*	*44.3 per cent*	*45.1 per cent*	*44.2 per cent*	43.8 per cent	*42.5 per cent*	*37.8 per cent*	36.6 per cent
0.20	0.40	21.5 (2.3)	21.7 (3.7)	21.8 (2.2)	22.2 (2.5)	21.7 (2.8)		22.1 (2.5)	20.6 (3.9)	
		43.0 per cent	*43.4 per cent*	*43.7 per cent*	*44.5 per cent*	*43.3 per cent*	42.9 per cent	*44.3* per cent	*41.2* per cent	41.4 per cent
0.30	0.70	15.2 (3.2)	16.4 (4.9)	17.2 (2.8)	17.6 (3.3)	15.9 (3.6)		21.0 (2.1)	19.6 (3.1)	
		30.4 per cent	*32.8 per cent*	*34.4 per cent*	*35.1 per cent*	*31.6 per cent*	30.0 per cent	*42.0 per cent*	39.1 per cent	39.6 per cent
0.60	0.80	16.9 (5.1)	19.1 (7.5)	20.2 (3.7)	19.3 (4.7)	17.5 (5.3)		23.5 (1.6)	23.1 (2.2)	
		33.9 per cent	*38.2* per cent	*40.3* per cent	*38.7* per cent	*35.0* per cent	33.3 per cent	*47.0* per cent	*46.3* per cent	46.4 per cent
0.70	0.90	13.1 (6.1)	17.9 (9.1)	19.9 (3.8)	17.3 (5.2)	14.1 (6.2)		23.7 (1.4)	23.4 (1.9)	
		26.2 per cent	*35.8 per cent*	*39.8 per cent*	*34.6 per cent*	*28.1 per cent*	25.0 per cent	*47.3 per cent*	*46.8 per cent*	46.9 per cent

PW, play-the-winner; RPW, randomized play-the-winner; DL, drop-the-loser; GDL, generalized drop-the-loser; DBCD, Doubly adaptive biased coin design.

exact sampling distribution (and its moments) for the PW and 1:1 randomized designs and are simulated from 10^5 replications for other designs.

In the left part of the table, where the target allocation proportion is (1), the PW and DBCD1 designs always have the expected proportions the closer to their asymptotic values, and a marked advantage over the other designs when the proportions of success are larger. The expected proportion of allocations to the less effective treatment is always smaller for the PW rule than for other rules, while the DL design has the poorest performance.

The number of subjects allocated to one of the treatments is always more variable for the RPW design than for the other designs. The DL design has the smallest standard deviations, which results in power improvement [2, 6]. GDL1 allocation proportion is more variable than for the DL rule, due to the requirement of estimating the probabilities of success at each stage when an immigration ball is drawn. However, if the DL rule is compared to the PW or DBCD1 rules, it appears that the gain in variance reduction is highly correlated with the loss in convergence rate. So, when the DL rule is nearly as efficient as the PW rule for allocation proportion (e.g. $\varphi_1 = 0.30$ and $\varphi_2 = 0.10$), it is not less variable. In fact, when a substantial gain in variability is observed (e.g. $\varphi_1 = 0.90$ and $\varphi_2 = 0.70$), it appears to be an artefact, due to the poor performance of the DL rule in terms of allocation proportion.

In the right part of the table, where the target allocation proportion is (2), the expected allocation proportions of the GDL2 design are far from the asymptotic value when the proportions of success are small, and are very good when they are large. DBCD2 has allocated proportions always very close to their asymptotic values. With respect to the objective of reducing the exposure to the less effective treatment, it must be noted that the modification of the target achieves a sensible reduction in the GDL2, and even more so in the DBCD2, when the success rates are small (e.g. 0.30 and 0.10), but on the contrary can result in a considerable increase when they are large (e.g. 0.90 and 0.70). The gain in variance reduction is sensible for GDL2 compared to DBCD2, but it appears as the consequence of its undesirable property of assigning the two treatments with probabilities closer to 0.5.

3. Inferential procedures

3.1. Some frequentist procedures

Wei [13] developed the permutation test for the RPW design. However, no explicit formula was given and a computationally intensive algorithm was needed to generate all the possible samples and get the exact distribution for the test statistic. Wei *et al.* [14] extended this algorithmic procedure for computing exact conditional confidence intervals for the odds ratio. They noted that these intervals are quite different from those obtained from the noncentral Hypergeometric distribution involved in the complete 50–50 randomization design. Furthermore, they concluded that they were 'rather conservative and not very useful in practice' (p. 158). Consequently, they also developed an unconditional procedure, but the resulting intervals were found to be less powerful than their conditional counterparts. Finally, they considered large-sample confidence intervals, but these intervals are not satisfactory for small or moderate-sized samples.

Similar comments can be made for the PW design. Conditional tests and their associated confidence intervals for the ratio φ_1/φ_2 were derived by ElQasyr and Lecoutre [9]. In this case, formulae, involving usual distributions, were made explicit. Moreover, the conservativeness of 'exact' tests and confidence intervals, due to the discreteness of the sampling distribution, can be overcome by the mid-p approach (e.g. [15–17]). The traditional p-value is the proportion of samples that are 'more extreme' than the observed data (under the null hypothesis), but depends on whether the observed data are included or not in the count of those samples. The usual exact test includes the observed data in the count and is consequently conservative. If the observed data are excluded from the count, the test becomes liberal. A typical easy-to-compute solution to overcome this problem consists of considering a *mid-p*-value, which averages the p-values of the conservative and liberal tests.

For the difference $\varphi_1 - \varphi_2$, it is always possible to consider the usual Z-test statistic (for example, [2]) or the normal approximation confidence interval, as in non-adaptive designs. But this procedure can be expected to be acceptable only for large sample sizes. Another omnibus procedure is to build confidence intervals using appropriate bootstrap techniques. Although these intervals have good coverage properties [18], they, however, do not perform as well as the Bayesian procedures described below.

The main general issue with efficient frequentist methods is that they require different *ad hoc* developments for each design and for each parameter of interest.

3.2. Bayesian procedures

Let n_{11} and n_{21} be the respective numbers of successes to the two treatments, and n_{10} and n_{20} be the numbers of failures. For all considered adaptive designs, the likelihood function is proportional to:

$$\varphi_1^{n_{11}}(1-\varphi_1)^{n_{10}}\varphi_2^{n_{21}}(1-\varphi_2)^{n_{20}} \tag{3}$$

i.e. proportional to the likelihood function associated with the comparison of two independent Binomial (or Negative Binomial) proportions. A simple and usual Bayesian solution assumes two independent Beta priors for φ_1 and φ_2, Beta(v_{11}, v_{10}) and Beta(v_{21}, v_{20}). This is a conjugate prior and the marginal posterior distributions are again two independent Beta distributions:

$$\text{Beta}(v_{11}+n_{11}, v_{10}+n_{10}) \quad \text{and} \quad \text{Beta}(v_{21}+n_{21}, v_{20}+n_{20}) \tag{4}$$

The Bayesian approach allows obtaining the distributions of any derived parameter of interest from the joint posterior distribution: in particular, efficient inferences can be made for the ratio, the difference or the odds-ratio [19]. The posterior distribution for these parameters can be approximated by simulating a large sample from two independent Beta distributions, i.e. repeatedly drawing random φ_1 and φ_2 from their two marginal posterior Beta distributions and computing the associated value for the parameter of interest. The $\alpha/2$ and $1-\alpha/2$ quantiles of the generated values give the (approximate) desired confidence limits. Alternatively, a probability statement about these parameters can be obtained by computing the probability that the two-dimensional variable (φ_1, φ_2) falls in a given region of the space $[0,1]^2$. Dividing this space into small rectangular regions and using the independence of the two marginal Beta distributions, the probability can be approximated from the usual incomplete Beta function, with any desired degree of accuracy, as the sum of the probabilities of these regions. This procedure was described by Novick and Jackson [20, p. 338]. It is implemented in a statistical computer program 'LesProportions' that computes Bayesian interval estimates for the main classical criteria for comparing two proportions.[‡]

For the PW design, there is a correspondence, given in Appendix, between the mid-p value of the conditional test of equality of proportions and the posterior Bayesian probabilities that $\varphi_1 < \varphi_2$ associated with particular choices of the prior.

3.3. Numerical illustration and remarks about the choice of the prior

Assume for instance a fixed number of $N=150$ subjects, and observed success rates: 68 out of 90 attributions for treatment 1 and 38 out of 60 attributions for treatment 2. Since the number of failures are equal (22 in each case), these data can be obtained for any of the considered designs. A simple reasonable solution for an objective prior is to take two marginal independent priors Beta(0.5, 0.5), that is the Jeffreys prior for the 1:1 randomized design. The resulting 90 per cent equal-tailed Bayesian credible intervals are:

[0.996, 1.457] for φ_1/φ_2 (vs conditional mid-p CI [0.989, 1.449])
$[-0.003, 0.247]$ for $\varphi_1-\varphi_2$
[0.986, 3.255] for $\varphi_1(1-\varphi_2)/\{\varphi_2(1-\varphi_1)\}$

In all rigor, the Jeffreys rule gives different priors for the different designs, since it is based on the Fisher information. So, it could be argued that an objective prior should depend on the design (e.g. [21–23]). Most possible refinements of the inference would however need more complicated numerical tools for a small expected improvement in terms of efficiency. This is in agreement with the solution proposed by Berger [24, pp. 5-6] for a problem of medical diagnosis involving three proportions. While expressing his theoretical preference for the *reference prior* approach [25], involving a different prior for each particular parameter of interest, he uses the simple Jeffreys prior in practice.

4. Coverage properties

In this section, we illustrate the frequentist properties of the Bayesian credible interval for the ratio φ_1/φ_2, assuming two Beta(0.5, 0.5) independent priors. We assume that the two treatments have equal probabilities of being assigned to the first subject. For fixed φ_1^* and φ_2^* values, the $100(1-\alpha)$ per cent credible interval for the ratio fails to contain the 'true value' if the lower limit is larger than φ_1^*/φ_2^* or if the upper limit is smaller than φ_1^*/φ_2^*.

The sampling probabilities of error for the lower limit are reported for $N=50$ and $\alpha=0.10$, and for all combinations of φ_1^* and φ_2^* values varying from 0.1 to 0.9 by step of 0.2. Table II gives the sampling probabilities of error for the ratio, for all considered response-adaptive designs and for the 1:1 randomized design. For the PW design, they are compared to the corresponding probabilities for the frequentist conditional mid-p confidence interval.

The Bayesian intervals always have reasonable coverage probabilities, in most cases close to the nominal level. Similar results were found for the difference and the odds ratio [19]. This confirms the conclusions previously obtained for the Jeffreys prior in other situations (see References [24, 26, 27]). For the PW design, the mid-p interval also has fairly good

[‡]*This program is freely available at address http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.html.*

Table II. Coverage error of Bayesian 90 per cent credible intervals for the ratio φ_1/φ_2: Probability of error for the lower limit for the 1:1 randomized design and for seven response-adaptive designs ($N=50$) [mid-p interval added for the PW rule].

		Target	$\frac{1-\varphi_2}{1-\varphi_1+1-\varphi_2}$						$\frac{\sqrt{\varphi_1}}{\sqrt{\varphi_1}+\sqrt{\varphi_2}}$	
φ_2^*	φ_1^*	1:1 RD	PW		RPW	DL	GDL1	DBCD1	GDL2	DBCD2
0.10	0.10	0.056	0.055	[0.037]	0.059	0.058	0.059	0.059	0.060	0.057
	0.30	0.056	0.067	[0.037]	0.062	0.069	0.064	0.066	0.055	0.013
	0.50	0.065	<0.001	[<0.001]	0.036	0.039	0.054	0.027	0.035	<0.001
	0.70	0.069	<0.001	[<0.001]	0.002	<0.001	0.002	<0.001	0.006	<0.001
	0.90	0.072	<0.001	[<0.001]	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
0.30	0.10	0.062	0.057	[0.047]	0.059	0.057	0.056	0.058	0.056	0.056
	0.30	0.050	0.051	[0.047]	0.051	0.051	0.051	0.052	0.055	0.061
	0.50	0.059	0.054	[0.047]	0.053	0.055	0.052	0.052	0.056	0.066
	0.70	0.052	0.065	[0.053]	0.054	0.055	0.048	0.055	0.054	0.063
	0.90	0.046	0.033	[0.014]	0.046	0.049	0.055	0.057	0.055	0.067
0.50	0.10	0.056	0.052	[0.046]	0.054	0.053	0.055	0.055	0.053	0.051
	0.30	0.053	0.054	[0.047]	0.054	0.055	0.054	0.054	0.055	0.055
	0.50	0.059	0.051	[0.048]	0.055	0.053	0.053	0.052	0.058	0.059
	0.70	0.049	0.051	[0.042]	0.053	0.053	0.051	0.051	0.053	0.054
	0.90	0.042	0.060	[0.036]	0.061	0.054	0.054	0.061	0.051	0.054
0.70	0.10	0.054	0.049	[0.048]	0.052	0.059	0.055	0.053	0.051	0.047
	0.30	0.056	0.054	[0.042]	0.054	0.055	0.055	0.054	0.054	0.054
	0.50	0.056	0.056	[0.047]	0.056	0.057	0.055	0.056	0.057	0.057
	0.70	0.050	0.055	[0.050]	0.055	0.058	0.054	0.055	0.053	0.053
	0.90	0.050	0.054	[0.056]	0.056	0.051	0.052	0.055	0.062	0.047
0.90	0.10	0.049	0.050	[0.035]	0.049	0.050	0.049	0.052	0.048	0.051
	0.30	0.056	0.041	[0.046]	0.046	0.049	0.053	0.045	0.052	0.051
	0.50	0.063	0.046	[0.052]	0.050	0.053	0.054	0.048	0.059	0.059
	0.70	0.070	0.051	[0.050]	0.058	0.059	0.056	0.054	0.062	0.060
	0.90	0.056	0.068	[0.049]	0.059	0.061	0.060	0.066	0.054	0.057

RD, randomized design; PW, play-the-winner; RPW, randomized play-the-winner; DL, drop-the-loser; GDL, generalized drop-the-loser; DBCD, doubly adaptive biased coin design. *Note*: For each combination $(\varphi_1^*, \varphi_2^*)$, the probability of error for the upper limit is equal to the probability for the lower limit associated with the reverse combination $(\varphi_2^*, \varphi_1^*)$.

coverage properties. The lower limit of the Bayesian credible and mid-p intervals could be judged very conservative for large values of $\varphi_1-\varphi_2$. However, this is not particular to these procedures: this is a consequence of the strong unbalance resulting from adaptive designs. For instance, for two independent binomial samples with respective sizes 40 and 10, the Koopman's confidence interval for φ_1/φ_2 [28] also has a probability of error near zero for its lower limit when $\varphi_2=0.10$ and φ_1 is larger than 0.50.

5. Power properties

Suppose that we aim to demonstrate that $\varphi_1>\varphi_2$. With the Bayesian approach, it can be concluded that $\varphi_1>\varphi_2$ when the lower limit of the $100(1-\alpha)$ per cent credible interval for φ_1/φ_2 is larger than one, or equivalently when the posterior probability $P(\varphi_1>\varphi_2\,|\,n_{11},n_{10},n_{21},n_{20})$ is larger than a given $1-\alpha/2$. The sampling probability of this event, assuming true values φ_1^* and φ_2^*, evaluates the power of the procedure. For given values of φ_2^*, we have considered the four φ_1^* values that respectively achieve approximatively 50, 80, 90 and 95 per cent power for the 1:1 randomized design with $N=50$ (equal allocation of 25 subjects for each treatment) and $\alpha=0.10$. The corresponding sampling probabilities are reported in Table III for all considered designs. They are computed from the sampling distribution for the PW and the 1:1 randomized design, and they are simulated with 10^5 replications for the other designs. Moreover, for the PW design, the power of the mid-p procedure is also reported.

Focusing on the power, the DL, GDL2 and DBCD2 designs appear to be the most efficient designs. For some combinations of success probabilities, these designs can even be slightly more powerful than the 1:1 randomized design. Previous claims about the superiority of the DL and doubly adaptive biased coin designs (see References [2, 6]) over the RPW design seem to be justified. However, power arguments should be balanced against the performance in terms of allocation proportions and consequently in terms of expected total number of treatment failures.

Table III. Power of the Bayesian 90 per cent credible interval procedure for the 1:1 randomized design and for seven response-adaptive designs ($N=50$) [mid-p confidence interval added for the PW rule]/Expected number of subjects assigned to the less effective treatment (standard deviation).

		Target	$\frac{1-\varphi_2}{1-\varphi_1+1-\varphi_2}$					$\frac{\sqrt{\varphi_1}}{\sqrt{\varphi_1}+\sqrt{\varphi_2}}$	
φ_2^*	φ_1^*	1:1 RD	PW	RPW	DL	GDL1	DBCD1	GDL2	DBCD2
0.10	0.276	0.50	0.479 [0.480]/22.3 (1.7)	0.487/22.4 (3.1)	0.482/22.5 (1.7)	0.490/22.9 (1.9)	0.487/22.5 (2.3)	0.499/21.6 (2.6)	0.506/19.4 (4.2)
	0.389	0.80	0.788 [0.785]/20.2 (2.0)	0.790/20.5 (3.3)	0.790/20.7 (2.0)	0.792/21.3 (2.2)	0.791/20.5 (2.5)	0.798/20.2 (2.4)	0.801/17.4 (3.8)
	0.450	0.90	0.894 [0.887]/19.0 (2.1)	0.890/19.4 (3.5)	0.895/19.6 (2.1)	0.895/20.3 (2.3)	0.894/19.4 (2.7)	0.899/19.6 (2.3)	0.899/16.6 (3.6)
	0.500	0.95	0.947 [0.940]/18.0 (2.2)	0.944/18.4 (3.6)	0.948/18.7 (2.4)	0.948/19.5 (2.4)	0.946/18.4 (2.7)	0.950/19.1 (2.3)	0.952/16.1 (3.5)
0.30	0.518	0.50	0.462 [0.439]/20.5 (2.9)	0.460/20.9 (4.5)	0.469/21.1 (2.7)	0.467/21.5 (3.0)	0.464/20.7 (3.3)	0.487/22.5 (2.3)	0.489/21.4 (3.3)
	0.631	0.80	0.758 [0.742]/17.4 (3.1)	0.753/18.2 (4.8)	0.768/18.7 (2.8)	0.766/19.2 (3.3)	0.759/17.8 (3.5)	0.790/21.5 (2.2)	0.785/20.2 (3.2)
	0.689	0.90	0.870 [0.856]/15.6 (3.2)	0.864/16.7 (4.9)	0.877/17.5 (2.8)	0.876/17.8 (3.3)	0.871/16.1 (3.6)	0.893/21.1 (2.2)	0.892/19.7 (3.1)
	0.734	0.95	0.929 [0.918]/14.0 (3.3)	0.921/15.4 (5.0)	0.935/16.4 (2.8)	0.933/16.7 (3.4)	0.929/14.7 (3.6)	0.946/20.8 (2.1)	0.945/19.3 (3.1)
0.50	0.718	0.50	0.470 [0.440]/18.2 (4.2)	0.466/19.5 (6.3)	0.484/20.1 (3.4)	0.476/20.0 (4.1)	0.469/18.6 (4.5)	0.494/23.2 (1.9)	0.492/22.7 (2.4)
	0.822	0.80	0.771 [0.744]/13.5 (4.4)	0.758/16.1 (6.6)	0.798/17.6 (3.4)	0.784/16.7 (4.2)	0.773/14.2 (4.6)	0.806/22.5 (1.8)	0.805/21.8 (2.4)
	0.868	0.90	0.868 [0.845]/10.9 (4.2)	0.858/14.4 (6.7)	0.902/16.6 (3.3)	0.887/15.0 (4.0)	0.869/11.9 (4.5)	0.904/22.2 (1.7)	0.906/21.4 (2.4)
	0.900	0.95	0.914 [0.889]/8.9 (4.1)	0.907/13.2 (6.7)	0.950/15.8 (3.3)	0.937/13.6 (3.9)	0.919/9.9 (4.2)	0.954/22.0 (1.7)	0.953/21.2 (2.3)
0.70	0.884	0.50	0.465 [0.419]/14.5 (6.1)	0.469/18.6 (8.9)	0.518/20.3 (3.8)	0.490/18.1 (5.3)	0.469/15.3 (6.2)	0.516/23.8 (1.4)	0.511/23.5 (1.9)
	0.948	0.80	0.693 [0.560]/8.4 (5.5)	0.740/15.8 (9.3)	0.807/18.7 (3.7)	0.779/14.6 (4.7)	0.736/9.8 (5.5)	0.807/23.4 (1.4)	0.809/23.1 (1.9)
	0.971	0.90	0.715 [0.560]/5.7 (4.8)	0.830/14.7 (9.4)	0.900/18.2 (3.7)	0.879/13.1 (4.2)	0.839/7.4 (4.8)	0.900/23.3 (1.3)	0.901/22.9 (1.9)
	0.985	0.95	0.664 [0.438]/3.8 (4.0)	0.881/14.0 (9.4)	0.949/17.8 (3.7)	0.933/12.1 (3.8)	0.905/5.8 (4.0)	0.948/23.2 (1.3)	0.949/22.8 (1.9)

RD, randomized design; PW, play-the-winner; RPW, randomized play-the-winner; DL, drop-the-loser; GDL, generalized drop-the-loser; DBCD, doubly adaptive biased coin design.

Table III illustrates the fact that power depends on variability, but as already emphasized in Section 2.7, the gain in variability and power is essentially an artefact due to a poor performance in terms of allocation proportions. For instance, when the true probabilities of success are $\varphi_2^*=0.70$ and $\varphi_1^*=0.971$, the power is 90 per cent with the 1:1 RD , DL, GDL2 and DBCD2 designs, versus only 0.715 and 0.839 respectively for the PW and DBCD1 designs. However, this is compensated by a clear advantage of the latter in terms of allocation proportions, as shown in Table III. For PW and DBCD1, 90 per cent power would require respectively $N=70$ and $N=67$. In terms of the expected total numbers of treatment failures the benefit is clear: 8.2 (1:1 RD), 6.4 (DL), 7.8 (GDL2), 7.7 (DBCD2) versus 4.0 (PW with $N=70$ and DBCD1 with $N=67$). Only PW and DBCD1 provide a noticeable advantage on 1:1 RD.

Note again that the power of the mid-p procedure is virtually always less than the power of the Bayesian procedure.

In most cases, the same power can be obtained for the 1:1 randomized design and for all considered response-adaptive designs with comparable sample sizes. This is illustrated in Table IV, which gives the number of subjects needed to achieve a 90 per cent power with the Bayesian procedure. The corresponding expected numbers of subjects assigned to the less effective treatment are indicated. The corresponding expected total number of treatment failures are given in Table V for the same $(\varphi_1^*, \varphi_2^*)$ pairs and for the five most efficient designs: PW, DL, DBCD1, GDL2 and DBCD2.

In most cases, PW and DBCD1 designs assign quite similar number of subjects to the less effective treatment. They usually perform better than DL design to achieve a 90 per cent power, and this translates into a slightly smaller expected number of treatment failures. The difference is especially noticeable for high proportions of success. In terms of allocation proportions, the GDL2 and DBCD2 designs, due to their different targets, can be either notably superior or inferior to PW and DBCD1 designs, with a trade-off effect of the expected allocation proportion and its variance. However, their superiority, observed for small proportions of success, does not translate into a notable reduction of the number and standard deviation of treatment failures. By contrast with the PW and DBCD1 design, the inferiority of the GDL2 and DBCD2 designs, especially observed for large proportions of success, can result in a notably larger number of treatment failures.

6. Conclusion

For all the considered response-adaptive designs, the Bayesian credible intervals have fairly good frequentist coverage properties, even for small sample sizes. This was illustrated for the ratio of two success rates, but this is equally valid and easily implemented with other measures of discrepancy between two rates. In fact, for all the above-mentioned designs with fixed number of subjects, independent priors Beta(0.5,0.5) for each of the two success rates were shown to work very well for the different parameters. Where the Bayesian inference could be compared to the mid-p and normal approximation, it performed equally well or better.

The power of Bayesian credible intervals, when using response-adaptive designs, was also shown to be very close to the power of 1:1 randomized design. At first sight the GDL2 and DBCD2 designs with optimal target allocation proportion, and to a lesser degree the DL design, showed outstanding results in terms of power. That advantage over PW and DBCD1 designs is the result of variance reduction, but it must be balanced with poorer performance in terms of allocation proportions for GDL2, DBCD2 and DL in small and moderate samples due to slower convergence. In any case the comparison is truly in disfavor of the RPW.

In terms of the total number of failures, it was shown that response-adaptive designs generally provide a modest advantage on 1:1 RD. The only notable exceptions were found for the PW and DBCD1 designs; they correspond to cases where the most effective treatment is clearly superior to the other treatment.

With respect to the goal of minimizing treatment failures while preserving power, our results confirm the previous disappointing conclusions of Rosenberger and Hu [2]. However, they also demonstrate that considerations for determining which design should be used in practice should include the PW rule as a reference. Finally, the choice of the most appropriate response-adaptive design depends on the study conditions. The alternative is mostly between the doubly adaptive biased coin design with the optimal target allocation proportion and PW, depending on the range of the success proportions expected in the two treatment groups.

The PW rule, which we included as reference design in this work for theoretical reasons, has been early discarded from most of comparative studies on the ground that it is conditionally deterministic, therefore likely to produce selection bias. This is true in the framework of a one-center trial, where treatment allocation and evaluation is performed by the same investigator. But most of the modern clinical research is now performed in multiple centers in different countries, and the allocation system is centralized, based on immediate notification of patient eligibility and response, even for non-adaptive designs. There are other situations where the (centralized) allocation and the evaluation are performed completely independently. In these settings, the conditional determinism of the PW design is compatible with a satisfactory protection against selection bias. Here it was demonstrated that the performances of DBCD1 are often very close to those of the PW, while providing a complete protection against selection bias.

Table IV. Number of subjects needed to achieve a 90 per cent power with the Bayesian procedure (first line) and corresponding expected number of subjects assigned to the less effective treatment (standard deviation) for the 1:1 randomized design and for seven response-adaptive designs.

		Target	$\frac{1-\varphi_2}{1-\varphi_1+1-\varphi_2}$					$\frac{\sqrt{\varphi_1}}{\sqrt{\varphi_1}+\sqrt{\varphi_2}}$	
φ_2^*	$\frac{\varphi_1^*}{\varphi_2^*}$	1:1 RD	PW	RPW	DL	GDL1	DBCD1	GDL2	DBCD2
0.10	1.5	1504	1487	1493	1496	1492	1490	1490	1491
		752	722.3 (7.3)	725.2 (14.9)	726.7 (7.4)	725.0 (10.0)	723.8 (11.8)	672.6 (35.1)	669.8 (30.4)
	2.0	426	426	426	426	426	427	426	425
		213	200.5 (4.3)	200.6 (8.4)	200.6 (4.4)	201.0 (5.8)	201.1 (6.6)	180.1 (15.9)	175.2 (16.7)
		106	95.9 (3.4)	96.5 (8.2)	96.1 (3.4)	96.2 (4.2)	96.1 (4.8)	86.4 (9.5)	80.8 (11.9)
	3.0	130	130	130	130	130	129	130	129
		65	56.9 (2.8)	57.1 (5.1)	57.2 (2.8)	57.8 (3.5)	56.7 (3.9)	51.6 (6.1)	46.7 (8.5)
		44	37.4 (2.5)	37.6 (4.4)	37.7 (2.5)	38.4 (3.0)	37.6 (3.4)	34.5 (4.2)	31.0 (6.2)
	4.0	64	67	67	66	66	67	66	64
		32	26.9 (2.3)	27.1 (3.9)	26.9 (2.3)	27.6 (2.6)	27.1 (3.0)	25.8 (3.1)	21.8 (4.6)
	6.0	28	27	29	27	27	26	26	27
		14	8.5 (1.8)	9.7 (2.8)	9.6 (1.6)	10.1 (1.7)	8.6 (2.0)	10.5 (1.2)	8.8 (1.9)
0.20	1.5	642	639	640	637	638	638	637	635
		321	298.2 (7.3)	298.8 (12.6)	297.4 (7.2)	298.4 (9.8)	297.9 (9.7)	287.6 (15.0)	285.2 (13.8)
	2.0	176	177	177	176	177	177	175	174
		88	75.9 (4.3)	76.2 (7.2)	75.8 (4.3)	77.0 (5.4)	76.1 (5.5)	74.2 (6.5)	71.6 (7.4)
		41	32.4 (3.3)	33.2 (5.3)	33.1 (3.2)	33.9 (3.8)	32.4 (4.0)	33.7 (3.6)	31.8 (5.0)
	3.0	48	49	51	49	49	50	48	47
		24	16.5 (2.7)	17.9 (4.3)	17.7 (2.5)	18.3 (2.8)	17.3 (3.2)	19.5 (2.2)	17.1 (3.3)
		16	9.2 (2.3)	10.5 (3.5)	10.8 (1.9)	11.1 (2.2)	9.8 (2.6)	11.9 (1.4)	10.3 (2.2)
	4.0	20	25	25	20	22	24	20	19
		10	5.3 (2.0)	6.7 (2.9)	6.8 (1.5)	7.3 (1.6)	5.9 (2.1)	8.4 (1.0)	6.7 (1.5)
0.30	1.5	352	356	356	355	353	354	354	351
		176	156.7 (7.3)	157.0 (12.0)	156.6 (7.2)	156.5 (9.4)	156.0 (8.9)	160.0 (8.4)	157.6 (8.2)
	2.0	90	93	94	93	93	92	90	89
		45	33.9 (4.2)	35.1 (6.7)	35.0 (4.0)	35.8 (4.8)	34.0 (4.8)	38.4 (3.4)	36.6 (4.3)
	2.5	42	42	44	41	41	42	39	40
		21	11.3 (3.0)	13.4 (4.6)	13.6 (2.5)	13.7 (2.9)	12.0 (3.3)	16.4 (1.7)	15.3 (2.7)
	3.0	20	28	25	19	20	22	18	17
		10	4.0 (2.2)	5.9 (3.3)	6.5 (1.6)	6.3 (1.5)	4.2 (2.0)	7.7 (1.0)	6.3 (1.4)
0.40	1.5	206	214	215	212	215	211	211	209
		103	85.7(7.2)	87.0(11.8)	85.8(7.0)	87.8(9.0)	84.8(8.3)	95.5(5.1)	93.8(5.4)
	2.0	48	52	54	49	50	52	48	46
		24	13.3 (3.8)	16.1 (6.0)	16.2 (3.0)	16.0 (3.7)	14.1 (4.2)	20.7 (1.9)	18.9 (2.6)
0.50	1.5	122	129	131	128	129	129	123	122
		61	43.2 (6.9)	46.4 (11.5)	45.8 (6.2)	46.1 (8.0)	43.7 (7.7)	55.8 (3.1)	54.8 (3.6)
	1.75	50	53	56	48	51	54	46	45
		25	11.1 (4.1)	15.6 (7.3)	15.9 (3.2)	14.7 (3.2)	12.3 (4.6)	20.4 (1.6)	19.2 (2.2)
0.60	1.25	330	335	340	333	334	336	330	329
		165	129.0 (12.8)	134.4 (24.0)	130.8 (12.1)	131.3 (16.2)	129.8 (14.5)	156.1 (4.5)	155.3 (5.4)
	1.5	64	74	78	65	70	74	64	64
		32	15.4 (6.0)	22.5 (10.7)	21.9 (4.2)	19.8 (5.8)	16.5 (6.4)	29.2 (1.8)	28.7 (2.4)
0.70	1.15	574	582	590	578	580	578	573	577
		287	229.4 (20.5)	239.9 (44.7)	231.7 (19.5)	231.6 (26.9)	228.2 (22.9)	276.6 (4.9)	278.4 (6.6)
	1.25	180	188	200	182	189	190	179	180
		90	55.8 (12.0)	69.0 (24.2)	64.0 (9.6)	61.1 (14.0)	57.1 (13.2)	84.7 (2.6)	85.0 (3.6)
	1.40	42	60	59	44	47	58	44	44
		21	5.1 (4.7)	16.2 (10.7)	16.2 (3.4)	17.1 (3.5)	6.9 (4.7)	20.5 (1.3)	20.1 (1.8)

RD, randomized design; PW, play-the-winner; RPW, randomized play-the-winner; DL, drop-the-loser; GDL, generalized drop-the-loser; DBCD, doubly adaptive biased coin design.

Future investigations should consider the case of sequential analysis and of delayed responses. Because of its flexibility, the Bayesian approach is well-suited for these purposes.

The Bayesian predictive approach enables stopping the trial early, or conversely extending it to an adequate size, in a sequential perspective, as illustrated in Lecoutre *et al.* [29]. This fits particularly well with the methodology of adaptive

Table V. Expected total number of failures (standard deviation) for the 1:1 randomized design and for the five most efficient response-adaptive designs, for the probabilities of success considered in Table V.

		Target	$\frac{1-\varphi_2}{1-\varphi_1+1-\varphi_2}$			$\frac{\sqrt{\varphi_1}}{\sqrt{\varphi_1}+\sqrt{\varphi_2}}$	
φ_2^*	$\frac{\varphi_1^*}{\varphi_2^*}$	1:1 RD	PW	DL	DBCD1	GDL2	DBCD2
0.10	1.5	1316.0 (12.8)	1300.1 (12.9)	1307.8 (12.9)	1302.8 (12.8)	1300.1 (12.9)	1300.8 (12.9)
	2.0	362.1 (7.3)	360.8 (7.5)	360.9 (7.5)	361.7 (7.5)	358.8 (7.6)	357.5 (7.6)
	3.0	104.0 (4.4)	102.4 (4.8)	102.5 (4.8)	101.6 (4.8)	101.3 (4.8)	99.7 (4.8)
	4.0	48.0 (3.2)	48.3 (3.9)	48.7 (2.3)	48.3 (3.9)	47.4 (3.1)	44.9 (3.7)
	6.0	18.2 (2.1)	15.0 (2.9)	15.6 (2.7)	14.7 (2.9)	15.6 (2.3)	15.2 (2.4)
0.20	1.5	481.5 (10.9)	477.1 (11.0)	475.6 (11.0)	476.4 (11.1)	474.7 (10.9)	473.0 (10.9)
	2.0	123.2 (5.9)	121.4 (6.4)	120.7 (6.4)	121.4 (6.4)	119.9 (6.1)	118.7 (6.0)
	3.0	28.8 (3.1)	26.2 (3.9)	26.7 (3.7)	26.9 (3.9)	27.0 (3.2)	25.6 (3.2)
	4.0	10.0 (1.8)	8.2 (2.9)	8.1 (2.0)	8.3 (2.8)	9.0 (1.8)	7.8 (1.8)
0.30	1.5	220.0 (9.0)	219.3 (9.3)	218.8 (7.2)	218.1 (8.9)	218.7 (9.0)	216.7 (9.0)
	2.0	49.5 (4.5)	47.4 (5.2)	47.7 (5.1)	47.0 (5.2)	47.5 (4.5)	46.6 (4.5)
	2.5	18.5 (2.8)	15.6 (3.6)	16.4 (3.1)	15.9 (3.6)	17.1 (2.7)	16.9 (2.7)
	3.0	8.0 (1.7)	5.2 (2.5)	5.8 (1.7)	4.7 (2.2)	6.5 (1.5)	5.5 (1.4)
0.40	1.5	103.0 (7.0)	102.7 (7.6)	102.0 (7.5)	101.4 (7.6)	103.5 (7.1)	102.3 (7.0)
	2.0	19.2 (3.1)	15.7 (3.8)	16.3 (3.2)	16.0 (3.9)	17.9 (2.9)	16.7 (2.8)
0.50	1.5	45.8 (5.2)	43.1 (5.8)	43.4 (5.5)	43.2 (5.8)	44.7 (5.0)	44.2 (5.0)
	1.75	15.6 (3.0)	10.8 (3.5)	12.0 (2.7)	11.3 (3.5)	13.4 (2.7)	12.8 (2.6)
0.60	1.25	107.3 (8.4)	103.1 (8.8)	102.8 (8.7)	103.5 (8.8)	105.9 (8.3)	105.5 (8.3)
	1.5	16.0 (3.2)	12.0 (3.7)	13.1 (2.9)	12.3 (3.8)	15.2 (3.0)	15.0 (3.0)
0.70	1.15	142.1 (10.3)	137.6 (10.5)	137.1 (10.4)	136.7 (10.6)	140.7 (10.2)	141.7 (10.2)
	1.25	40.0 (5.5)	33.3 (5.7)	34.0 (5.2)	33.7 (5.8)	37.2 (5.2)	37.4 (5.2)
	1.40	6.7 (2.2)	2.6 (2.0)	5.4 (1.7)	3.1 (1.9)	6.6 (2.0)	6.5 (2.0)

RD, randomized design; PW, play-the-winner; RPW, randomized play-the-winner; DL, drop-the-loser; GDL, generalized drop-the-loser; DBCD, doubly adaptive biased coin design.

designs. An extra advantage of PW design is that easy to compute exact formulae of the predictive distribution are available (see Appendix).

RPW, GDL and DBCD designs cope with delayed responses. PW and DL designs can be adapted for this circumstance. Either some default random allocation, or preferably an allocation based on available responses can be used when the adaptive rule cannot be applied. Frequentist inference procedures must be modified to take into account the effect of these responses on the sampling distribution. On the contrary, the same inference Bayesian procedures can be used. Moreover, it can be expected that the resulting gain in power due to randomized allocations will partly compensate the loss in efficiency of the adaptive process.

Appendix A

ElQasyr and Lecoutre [9] derived the sampling and predictive distributions for the PW rule. Relevant results are summarized hereafter.

Mid-p values of the conditional test for the PW design and Bayesian posterior probabilities

Let n_{11} and n_{21} be the respective numbers of successes to the two treatments, and n_{10} and n_{20} be the numbers of failures. The conditional test is based on the conditional sampling distribution of n_{11}, given fixed numbers of failures n_{10} and n_{20}. This distribution involves only the ratio φ_1/φ_2, allowing to compute the p-value of the exact conditional test of the null hypothesis $\varphi_1=\varphi_2$ against the alternative $\varphi_1>\varphi_2$ in the usual way. Let $\overline{p}_{\text{inc}}$ be this p-value (the subscript inc indicates that n_{11} is included in the computation), and let $\overline{p}_{\text{exc}}$ be its counterpart obtained by excluding n_{11}. For the PW design, the following equalities hold

$$\begin{aligned}\overline{p}_{\text{inc}} &= P_{0,0,1,1} \quad \text{if } n_{10}=n_{20}+1\\ &= P_{0,1,1,0} \quad \text{if } n_{20}=n_{10}+1\\ &= \tfrac{1}{2}(P_{0,1,1,0}+P_{0,0,1,1}) \quad \text{if } n_{10}=n_{20}\end{aligned}$$

$$\begin{aligned}\overline{p}_{\text{exc}} &= P_{1,0,0,1} \quad \text{if } n_{10}=n_{20}+1\\ &= P_{1,1,0,0} \quad \text{if } n_{20}=n_{10}+1\\ &= \tfrac{1}{2}(P_{1,1,0,0}+P_{1,0,0,1}) \quad \text{if } n_{10}=n_{20}\end{aligned}$$

where $P_{v_{11},v_{10},v_{21},v_{20}}$ is the posterior probability that $\varphi_1<\varphi_2$, given $(n_{11},n_{10},n_{21},n_{20})$, associated with the prior defined by $(v_{11},v_{10},v_{21},v_{20})$. Note that a consequence of the PW rule is that n_{10} and n_{20} are equal or different by only one unit.

This extends the previous correspondence between conditional tests and Bayesian procedures obtained for the Binomial and Negative Binomial sampling [17] and for the multinomial sampling [26, 30]. Moreover, in the present case of integer (v_{ij}), $P_{v_{11},v_{10},v_{21},v_{20}}$ is the probability that a Beta-Binomial distribution with parameters $n_{11}+n_{10}+v_{11}+v_{10}-1$, $n_{21}+v_{21}$ and $n_{20}+v_{20}$ is greater or equal to $n_{11}+v_{11}$. This allows a simple computation of the mid-p value, defined by $\overline{p}_{\text{mid}}=(\overline{p}_{\text{inc}}+\overline{p}_{\text{exc}})/2$.

It can be noted that the Jeffreys prior for the 1:1 randomized design—two marginal independent priors Beta(1/2, 1/2)—corresponds to the average of the prior weights of the four different Beta priors involved by the mid-p value. Consequently, it also appears as a simple reasonable prior for the PW design.

Predictive distribution for the PW design

The Bayesian predictive probability of observing $(n_{11},n_{10},n_{21},n_{20})$ is

$$\Pr(n_{11},n_{10},n_{21},n_{20})=\int_0^1\int_0^1 \Pr(n_{11},n_{10},n_{21},n_{20}|\varphi_1,\varphi_2)p(\varphi_1,\varphi_2)\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2$$

i.e. a mixture of the sampling probabilities, the weights being given by the prior density $p(\varphi_1,\varphi_2)$. For all designs, it can be obtained by simulation. For PW, a simple expression can be derived. If either treatment is randomly assigned to the first subject with probability 0.5, assuming two independent Beta priors for φ_1 and φ_2, Beta(v_{11},v_{10}) and Beta(v_{21},v_{20}), we get

$$\begin{aligned}\Pr(n_{11},n_{10},n_{21},n_{20}) &= \tfrac{1}{2}\,p_{\text{B-Bin}}(n_{21};n_{21}+n_{20},v_{21},v_{20})\,p_{\text{B-NBin}}(n_{11};n_{10},v_{11},v_{10})\mathbb{1}_{\{0,1\}}(n_{10}-n_{20})\\ &\quad+\tfrac{1}{2}\,p_{\text{B-Bin}}(n_{11};n_{11}+n_{10},v_{11},v_{10})\,p_{\text{B-NBin}}(n_{21};n_{20},v_{21},v_{20})\mathbb{1}_{\{0,1\}}(n_{20}-n_{10})\end{aligned}$$

expressed from Beta-Binomial and Beta-Negative-Binomial distributions

$$p_{\text{B-Bin}}(j;r,a,b)=\binom{r}{j}\frac{B(j+a,r-j+b)}{B(a,b)}=\frac{\binom{j+a-1}{j}\binom{r-j+b-1}{r-j}}{\binom{r+a+b-1}{r}}$$

$$p_{\text{B-NBin}}(j;r,a,b)=\binom{j+r-1}{j}\frac{B(j+a,r+b)}{B(a,b)}=\frac{\binom{j+a-1}{j}\binom{r+b-1}{r}}{\binom{j+r+a+b-1}{j+r}}$$

where $B(a,b)$ denotes the beta function.

The predictive probability of observing $(n'_{11},n'_{10},n'_{21},n'_{20})$ in a future independent sample of size N' can be obtained in the same way, replacing the prior Beta distributions with the posterior Beta distributions.

References

1. Pullman D, Wang X. Adaptive designs, informed consent, and the ethics of research. *Controlled Clinical Trials* 2001; **22**:203–210. DOI: 10.1016/S0197-2456(01)00122-2.
2. Rosenberger WF, Hu F. Maximizing power and minimizing treatment failures. *Clinical Trials* 2004; **1**:141–147. DOI: 10.1191/1740774504cn016oa.
3. Zelen M. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 1969; **64**:131–146.
4. Wei LJ, Durham S. The randomized play-the-winner rule in medical trial. *Journal of the American Statistical Association* 1978; **73**:840–843.
5. Eisele JR. The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference* 1994; **38**:249–261.
6. Ivanova A. A play-the-winner-type urn design with reduced variability. *Metrika* 2003; **58**:1–13.
7. Zhang LX, Chan WS, Cheung SH, Hu F. Generalized drop-the-loser urn clinical trials with delayed response. *Statistica Sinica* 2007; **17**:387–409.
8. Lecoutre B, ElQasyr K. Adaptive designs for multi-arm clinical trials: the play-the-winner rule revisited. *Communications in Statistics—Simulation and Computation* 2008; **37**:590–601. DOI: 10.1080/03610910701812402.

9. ElQasyr K, Lecoutre B. Comparing two success rates with play-the-winner designs. Retrieved June 4, 2010. Available from: http://hal.archives-ouvertes.fr/hal-00422985, 2009.
10. Sun R, Cheung SH, Zhang L-X. A generalized drop-the-loser rule for multi-treatment clinical trials. *Journal of Statistical Planning and Inference* 2007; **137**:2011–2023. DOI: 10.1016/j.jspi.2006.06.039.
11. Hu F, Zhang L-X. Asymptotic properties of doubly-adaptive biased coin designs for multitreatment clinical trials. *The Annals of Statistics* 2004; **32**:268–301. DOI: 10.1214/aos/1079120137.
12. Rosenberger WF, Stallard N, Ivanova A, Harper CN, Ricks ML. Optimal adaptive designs for binary response trials. *Biometrics* 2001; **57**:909–913. DOI: 10.1111/j.0006-341X.2001.00909.x.
13. Wei LJ. Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* 1988; **75**:603–606. DOI: 10.1093/biomet/75.3.603.
14. Wei LJ, Smythe RT, Lin DY, Park TS. Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association* 1990; **85**:156–162.
15. Routledge RD. Practicing safe statistics with the mid-p^*. *The Canadian Journal of Statistics* 1994; **22**:103–110.
16. Berry G, Armitage P. Mid-P confidence intervals: a brief review. *The Statistician* 1995; **44**:417–423.
17. Lecoutre B. Bayesian methods for experimental data analysis. In *Handbook of Statistics*: *Epidemiology and Medical Statistics*, vol. 27, Rao CR, Miller J, Rao DC (eds). Elsevier: Amsterdam, 2007; 775–812.
18. Rosenberger WF, Hu F. Bootstrap methods for adaptive designs. *Statistics in Medicine* 1999; **18**:1757–1767.
19. ElQasyr K. *Modélisation et Analyse Statistique des Plans d'Expérience Séquentiels*. *Doctoral thesis*, Université de Rouen, 2008. Retrieved June 4, 2010. Available from: http://tel.archives-ouvertes.fr/docs/00/37/71/14/PDF/these.pdf.
20. Novick MR, Jackson PH. *Statistical Methods for Educational and Psychological Research*. McGraw-Hill: New York, 1974.
21. de Cristofaro R. On the foundations of likelihood principle. *Journal of Statistical Planning and Inference* 2004; **126**:401–411. DOI: 10.1016/j.jspi.2003.08.011.
22. Bunouf P, Lecoutre B. Bayesian priors in sequential binomial design. *Comptes Rendus de l'Académie des Sciences*, *Série I* 2006; **343**:339–344. DOI: 10.1016/j.crma.2006.06.029.
23. Sun D, Berger JO. Objective Bayesian analysis under sequential experimentation. *IMS Collections*, *Pushing the Limits of Contemporary Statistics*: *Contributions in Honour of Jayanta K Ghosh* 2008; **3**:19–32.
24. Berger J. The case for objective Bayesian analysis. *Bayesian Analysis* 2004; **1**:1–17.
25. Berger JO, Bernardo JM. On the development of the reference prior method (with discussion). In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Oxford University Press: Oxford, 1992; 35–60.
26. Lecoutre B, Charron C. Bayesian procedures for prediction analysis of implication hypotheses in 2×2 contingency tables. *Journal of Educational and Behavioral Statistics* 2000; **25**:185–201. DOI: 10.3102/10769986025002185.
27. Agresti A, Min Y. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics* 2005; **61**:515–523. DOI: 10.1111/j.1541-0420.2005.031228.x.
28. Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 1984; **40**:513–517.
29. Lecoutre B, Derzko G, Grouin J-M. Bayesian predictive approach for inference about proportions. *Statistics in Medicine* 1995; **14**:1057–1063. DOI: 10.1002/sim.4780140924.
30. Altham PME. Exact Bayesian analysis of a 2×2 contingency table and Fisher's 'exact' significance test. *Journal of the Royal Statistical Society B* 1969; **31**:261–269.