Statistical Papers (2022) 63:157–180
https://doi.org/10.1007/s00362-021-01234-3

REGULAR ARTICLE

A simple solution to the inadequacy of asymptotic likelihood-based inference for response-adaptive clinical trials

Likelihood-based inference for RAR trials

Alessandro Baldi Antognini[1] · **Marco Novelli**[1] · **Maroussa Zagoraiou**[1]

Received: 4 September 2020 / Revised: 25 January 2021 / Accepted: 2 April 2021 /
Published online: 17 April 2021

Abstract
The present paper discusses drawbacks and limitations of likelihood-based inference in sequential clinical trials for treatment comparisons managed via Response-Adaptive Randomization. Taking into account the most common statistical models for the primary outcome—namely binary, Poisson, exponential and normal data—we derive the conditions under which (i) the classical confidence intervals degenerate and (ii) the Wald test becomes inconsistent and strongly affected by the nuisance parameters, also displaying a non monotonic power. To overcome these drawbacks, we provide a very simple solution that could preserve the fundamental properties of likelihood-based inference. Several illustrative examples and simulation studies are presented in order to confirm the relevance of our results and provide some practical recommendations.

Keywords Confidence intervals · Ethics · Hypothesis testing · Power · Target allocations · Type-I errors

1 Introduction

Over the past decades a growing stream of statistical papers on the topic of Response-Adaptive Randomization (RAR) has flourished, especially in the context of phase-III clinical trials for treatment comparisons, also due to the encouragement of U.S. gov-

✉ Alessandro Baldi Antognini
a.baldi@unibo.it

Marco Novelli
marco.novelli4@unibo.it

Maroussa Zagoraiou
maroussa.zagoraiou@unibo.it

[1] Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy

ernment agencies and Health Authorities (CHMP 2007; FDA 2018). RAR procedures are sequential allocation rules in which the allocation probabilities change on the basis of earlier responses and past assignments; the aim is to balance the experimental goals of drawing correct inferential conclusions and caring about the welfare of each patient, the so-called *individual-versus-collective ethics* dilemma (for a recent review, see Hu and Rosenberger 2006; Atkinson and Biswas 2014; Baldi and Giovagnoli 2015; Rosenberger and Lachin 2015). A cornerstone example is the randomized Play-the-Winner (PW) suggested for binary trials (see, e.g., Wei and Durham 1978; Ivanova 2003). The peculiarity of the PW rule is that the allocation proportion of each of the two treatments converges to the relative risk of the other, so that (asymptotically) the majority of patients will receive the best treatment. Another example, for normal and survival outcomes, is the treatment effect mapping (Rosenberger 1993), where the assignments are based on a function that links the difference between the treatment effects to the ethical skew of the allocation probability (Rosenberger and Seshaiyer 1997; Bandyopadhyay and Biswas 2001; Atkinson and Biswas 2005b).

Since the statistical object of drawing correct inferential conclusions about the identification of the best treatment and its relative superiority often conflicts with the ethical aim of maximizing the subjects care, some authors formalize these goals into suitable combined/constrained optimization problems (see, e.g., Rosenberger et al. 2001; Baldi Antognini and Giovagnoli 2010). The ensuing optimal allocations, usually referred to as targets, depend in general on the unknown treatment effects; although *a priori* unknown (the so-called *local optimality* problem), they can be approached by RAR procedures that estimate sequentially the model parameters in order to progressively approach the chosen target. Classical examples are the Efficient Randomized Adaptive DEsign (ERADE) proposed by Hu et al. (2009) and the doubly-adaptive biased coin design (Hu and Zhang 2004). Under a different perspective, the same trade-off between ethics and inference represents a special case of the so-called *exploration-versus-exploitation* dilemma in the Bayesian literature of bandit problems, where at each step an agent wants to simultaneously acquire new knowledge and optimize his/her decisions based on existing information (see for review Villar et al. 2015a, b).

Although the adaptation process induces a complex dependence structure, several authors provide the conditions under which the classical asymptotic likelihood-based inference is still valid for RAR procedures (see, e.g., Durham et al. 1997; Melfi and Page 2000; Baldi Antognini and Giovagnoli 2005). Essentially, the crucial one regards the limiting allocation proportion induced by the chosen RAR rule, that should be a non-random quantity different from 0 and 1. Excluding some extremely ethical procedures, such as the randomly reinforced urn designs (May and Flournoy 2009), such condition is generally satisfied by the existing RAR rules and therefore the usual asymptotic properties of the MLEs are preserved; indeed the large majority of the literature has been focussed on the asymptotic likelihood-based inference, where the Wald test is the cornerstone (Rosenberger and Sriram 1996; Rosenberger et al. 1997; Melfi et al. 2001; Hu and Zhang 2004; Atkinson and Biswas 2005a, b; Geraldes et al. 2006; Tymofyeyev et al. 2007; Azriel et al. 2012). Under RAR procedures, Yi and Li (2018) theoretically prove that the Wald statistics is first order efficient, while Yi and Wang (2011) show via simulations that, although asymptotically equivalent to likelihood ratio and score tests, it performs better in small samples. However, several

simulation studies exhibit that, in some circumstances, such an approach presents anomalies in terms of coverage probabilities of confidence intervals, as well as inflated type-I errors (see, e.g., Rosenberger and Hu 1999; Yi and Wang 2011; Atkinson and Biswas 2014; Baldi Antognini et al. 2018), especially for targets with a strong ethical component.

The aim of this paper is to demonstrate the inadequacy of asymptotic likelihood-based inference for RAR procedures, in terms of both confidence intervals and hypothesis testing. We stress the crucial role played by the chosen target, the variance function of the statistical model and the presence of nuisance parameters, that could (i) compromise the quality of the Central Limit Theorem (CLT) approximation of the standard MLEs and (ii) lead to a vanishing Fisher information. In particular, these degeneracies could happen when the variance function is unbounded or when the target allocations approach either 0 or 1 (that depends on both the chosen ethical component and on the relative superiority of a given treatment wrt the other), showing also how the functional form of the target could induce a non monotonic power function. We prove that the Wald test could become inconsistent, it may display a strong dependence on the nuisance parameters, and the standard confidence intervals could degenerate.

Since the common approach of the practitioners consists in superimposing a minimum percentage of allocations to each treatment, we demonstrate that by re-scaling the target some of these drawbacks could be circumvented. We show how a suitable choice of the threshold can be matched with a strong ethical skew of the target without compromising the inferential precision. Several illustrative examples are provided for normal, binary, Poisson and exponential data and simulation studies are performed in order to confirm the relevance of our results.

The paper is structured as follows. Starting from the notation and some preliminaries in Sect. 2, Sect. 3 deals with likelihood-based inference, highlighting its inadequacy for RAR procedures in Sect. 4, with several examples showing the practical implication of the above-mentioned drawbacks. Section 5 discusses our proposal of re-scaling the target and its properties and Sect. 6 deals with some concluding remarks.

2 Preliminaries

2.1 Notation and model

Suppose that statistical units come to the trial sequentially and are assigned to one of two competing treatments, say A and B. At each step $i \geq 1$, let δ_i be the indicator managing the allocation of the ith subject, namely $\delta_i = 1$ if he/she is assigned to A and 0 otherwise. Given the treatment assignments, the observed outcomes Ys relative to either treatment are assumed to be independent and identically distributed belonging to the natural exponential family with quadratic variance function $Y \sim NQ(\theta; v(\theta))$, where $\theta \in \Theta \subseteq \mathbb{R}$ denotes the mean and the variance $v = v(\theta) > 0$ is at most a quadratic function of the mean (Morris 1982). In this setting, $\boldsymbol{\theta} = (\theta_A; \theta_B)^t$ denotes the treatment effects and from now on we let $\overline{\Theta} = \sup \Theta$ and $\underline{\Theta} = \inf \Theta$. Special cases of particular relevance for applications are the Bernoulli distribution (with $\theta_j \in (0; 1)$ and

$v(\theta_j) = \theta_j(1-\theta_j)$) for binary outcomes, the Poisson model ($\theta_j \in \mathbb{R}^+$ and $v(\theta_j) = \theta_j$) for count data, the exponential distribution ($\theta_j \in \mathbb{R}^+$ and $v(\theta_j) = \theta_j^2$) for survival outcomes, while the normal homoscedastic model is also encompassed for continuous responses (where $\theta_j \in \mathbb{R}$ and $v(\theta_j) = v \in \mathbb{R}^+$ is the common nuisance parameter). In this setting, the treatment outcomes are stochastically ordered on the basis of their effects and from now on (without loss of generality) we assume that high responses are preferable. As it is well known, the NQ class contains two more basic models, such as the negative binomial and the generalized hyperbolic secant distribution, which however may be less appealing for practical applications, especially in the clinical context.

After n steps, let $N_{An} = \sum_{i=1}^{n} \delta_i$ and $N_{Bn} = n - N_{An}$ be the assignments to both treatments, so that $\pi_n = n^{-1} N_{An}$ is the allocation proportion to A (respectively, $1-\pi_n$ to B). Then, the MLEs of the treatment effects coincide with the sample means, namely $\hat{\theta}_{An} = N_{An}^{-1} \sum_{i=1}^{n} \delta_i Y_i$ and $\hat{\theta}_{Bn} = N_{Bn}^{-1} \sum_{i=1}^{n} (1-\delta_i) Y_i$, while the normalized Fisher information is $\mathbf{M}_n = \text{diag}\,(\pi_n/v_A; [1-\pi_n]/v_B)$ (see Baldi and Giovagnoli 2015).

2.2 Target allocations and RAR rules

Motivated by ethical demands, Response-Adaptive procedures have been proposed with the aim of skewing the assignments towards the treatment that appears to be superior or, more in general, of converging to suitable limiting allocation proportions—say $\rho = \rho(\boldsymbol{\theta}) \in (0; 1)$ to A (and $1-\rho$ to B, respectively)—namely ideal allocations of the treatments representing a valid trade-off among ethics and inference.

In the context of binary trials, a classical example is the PW rule (Zelen 1969), under which a success on a given treatment leads to assigning the same treatment to the next unit, while a failure implies switching to the competitor. Under this procedure, the allocation proportion of treatment A converges to

$$\rho_{PW}(\boldsymbol{\theta}) = \frac{1-\theta_B}{2-\theta_A-\theta_B}, \tag{1}$$

which is also the limiting allocation of the randomized PW (Wei and Durham 1978) and of the Drop-the-Loser rule (Ivanova 2003). Differently, for normal homoscedastic trials Bandyopadhyay and Biswas (2001) and Atkinson and Biswas (2005b) suggested RAR procedures targeting

$$\rho_N(\boldsymbol{\theta}) = \Phi\left(\frac{\theta_A-\theta_B}{T}\right), \tag{2}$$

where Φ is the cumulative distribution function (cdf) of the standard normal and $T > 0$ a tuning parameter. Although ρ_{PW} and ρ_N are considered ethical targets, as the majority of subjects are assigned to the best treatment, they do not have a formal mathematical justification. On the other hand, by expressing ethical aims and inferential goals into suitable design criteria, several authors provided optimal allocations via combined/constrained optimization problems. An example for binary trials is the

target proposed by Rosenberger et al. (2001) and further generalized by Tymofyeyev et al. (2007), namely

$$\rho_Z(\boldsymbol{\theta}) = \frac{\sqrt{\theta_A}}{\sqrt{\theta_A} + \sqrt{\theta_B}},$$

which is aimed at minimizing the expected number of failures for a given variance of the estimated treatment difference, while

$$\rho_R(\boldsymbol{\theta}) = \frac{\theta_A}{\theta_A + \theta_B},$$

corresponds to the so-called A- and E-optimal design for exponential and Poisson data, respectively (Baldi and Giovagnoli 2015). Clearly, these targets also encompass normal homoscedastic data provided that the treatment effects are positive (Zhang and Rosenberger 2006).

In order to favour the best treatment, the targets should depend on a suitable discrepancy measure between the unknown treatment effects (like, e.g., the treatment difference in ρ_N, the ratio between the effects for ρ_R or the relative risk in ρ_{PW}), so that the target function ρ links the relative superiority of a given treatment to the ethical skewness of the allocations. Moreover, as for (2), the targets could also depend on a non-negative constant T—chosen by the experimenter—managing their ethical skew (i.e., for low values of T the target tends to strongly skew the assignments to the best treatment, while as T grows the ethical component vanishes and ρ tends to balance the allocations). Therefore, common assumptions are:

$A1$: ρ is a continuous function invariant under label permutation of the treatments, namely $\rho(\theta_A; \theta_B) = 1 - \rho(\theta_B; \theta_A)$,

$A2$: ρ is increasing in θ_A and decreasing in θ_B,

ensuring that (i) both treatments are treated likewise and (ii) the best treatment should be favoured increasingly as its relative superiority grows.

Remark 1 Note that, on the basis of the underlying statistical model, the well-known Neyman allocation $\rho(\boldsymbol{\theta}) = \sqrt{v_A}/(\sqrt{v_A} + \sqrt{v_B})$ - i.e., the A-optimal design—may not have any ethical appeal, since the majority of patients could be assigned to the worst treatment. Indeed, for binary and normal outcomes it does not satisfy assumption $A2$, while for Poisson and exponential data the Neyman target is ethical and corresponds to ρ_Z and ρ_R, respectively.

Given a desired ρ, RAR rules based on sequential estimation could be employed to converge to it. After a starting sample of n_0 subjects assigned to both treatments to derive non-trivial estimates of the unknown parameters, at each step $n > 2n_0$ the treatment effects are estimated by means of $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{An}; \hat{\theta}_{Bn})^t$ and the target is estimated accordingly by $\rho(\hat{\boldsymbol{\theta}}_n)$, so the next assignment is forced to converge to ρ. For instance, ERADE (Hu et al. 2009) randomizes the allocations by

$$\Pr(\delta_{n+1} = 1 \mid \delta_1, Y_1, \ldots, \delta_n, Y_n) = \begin{cases} \gamma\rho(\hat{\boldsymbol{\theta}}_n), & \text{if } \pi_n > \rho(\hat{\boldsymbol{\theta}}_n) \\ \rho(\hat{\boldsymbol{\theta}}_n), & \text{if } \pi_n = \rho(\hat{\boldsymbol{\theta}}_n), \\ 1 - \gamma\left[1 - \rho(\hat{\boldsymbol{\theta}}_n)\right], & \text{if } \pi_n < \rho(\hat{\boldsymbol{\theta}}_n) \end{cases}$$

where $\gamma \in [0; 1)$ is the randomization parameter of the allocation process.

3 Asymptotic likelihood-based inference for RAR procedures

Assuming that the inferential goal consists in estimating/testing the superiority of a given treatment with respect to the gold standard (say A wrt B), the parameter of interest is the treatment difference $\Delta = \theta_A - \theta_B$, while θ_B is usually regarded as a nuisance parameter (namely, θ_B is a common baseline while Δ represents the additive effect of the relative superiority/inferiority of A over B). Although the MLEs remain the same as the non-sequential setting's ones, this is not true for their distribution because of the complex dependence structure generated by the adaptation process. However, if the RAR design is chosen so that

$C1$: $\lim_{n\to\infty} \pi_n = \rho(\boldsymbol{\theta}) \in (0; 1) \quad a.s.$

with $\rho(\boldsymbol{\theta})$ satisfying assumptions $A1$-$A2$, then the standard asymptotic inference is allowed. Indeed,

$$\lim_{n\to\infty} \mathbf{M}_n = \mathbf{M} = \text{diag}\left(\frac{\rho(\boldsymbol{\theta})}{v_A}; \frac{1-\rho(\boldsymbol{\theta})}{v_B}\right) \quad a.s.$$

and the MLEs are still consistent and asymptotically normal with $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \hookrightarrow \mathcal{N}(\mathbf{0}_2, \mathbf{M}^{-1})$, where $\mathbf{0}_2$ is the 2-dim vector of zeros. Thus, let $\hat{\Delta}_n = \hat{\theta}_{An} - \hat{\theta}_{Bn}$, then $\sqrt{n}(\hat{\Delta}_n - \Delta) \hookrightarrow \mathcal{N}\left(0, \sigma^2\right)$, where

$$\sigma^2_\rho = \frac{v_A}{\rho(\boldsymbol{\theta})} + \frac{v_B}{1-\rho(\boldsymbol{\theta})} \tag{3}$$

and, due to the continuity of the target, $\lim_{n\to\infty} \rho(\hat{\boldsymbol{\theta}}_n) = \rho(\boldsymbol{\theta})\ a.s.$ Letting $\hat{v}_{jn}$s be consistent estimators of the treatment variances, then

$$\hat{\sigma}^2_n = \frac{\hat{v}_{An}}{\rho(\hat{\boldsymbol{\theta}}_n)} + \frac{\hat{v}_{Bn}}{1-\rho(\hat{\boldsymbol{\theta}}_n)}$$

is a consistent estimators of σ^2_ρ and the $(1-\alpha)\%$ asymptotic confidence interval is

$$CI(\Delta)_{1-\alpha} = \left(\hat{\Delta}_n \pm \frac{z_{1-\alpha/2}\hat{\sigma}_n}{\sqrt{n}}\right), \tag{4}$$

where z_α is the α-percentile of Φ.

For what concerns hypothesis testing, the inferential aim typically lies in testing $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$ (or $H_1 : \Delta \neq 0$). The asymptotic test is usually performed via the Wald statistic $W_n = \sqrt{n}\hat{\Delta}_n\hat{\sigma}_n^{-1}$ which, under H_0, converges to the standard normal distribution. Thus, given the alternative $H_1 : \Delta > 0$, the power of the

α-level test is $\Pr\left(\sqrt{n}(\hat{\Delta}_n - \Delta) > z_{1-\alpha}\hat{\sigma}_n - \sqrt{n}\Delta\right)$, which can be approximated by

$$\Phi\left(\sqrt{n}\Delta\sigma_\rho^{-1} - z_{1-\alpha}\right), \quad \Delta \geq 0, \tag{5}$$

due to the consistency of $\hat{\sigma}_n^2$. As stated by several authors (Lehmann 1999; Hu and Rosenberger 2006; Tymofyeyev et al. 2007), this approximation is accurate and particularly effective in the moderate-large sample setting of phase-III trials therefore neither for early phase studies with small sample sizes, nor asymptotically (where different approaches aimed at providing proper local approximation of the power around specific value of Δ as $n \to \infty$ could be suitable like e.g. the *local alternative* framework).

Even if less interesting in the actual practice, the two-sided alternative $H_1 : \Delta \neq 0$ can be encompassed analogously. Under H_0, W_n^2 converges in distribution to a central chi-square χ_1^2 with 1 degree of freedom; while under H_1, W_n^2 could be approximated by a non-central χ_1^2 with non-centrality parameter $n\Delta^2\sigma_\rho^{-2}$, namely the square of the crucial quantity in (5). As is well-known, the power is an increasing function of the non-centrality parameter and it is maximized by the Neyman allocation, also minimizing (3).

4 Inadequacy of likelihood-based inference

Note that condition $C1$ avoids the extreme scenarios $\rho = 0$ or 1; however, most of the targets suggested in the literature satisfy the following property:

$$\lim_{\theta_A \to \overline{\Theta}} \rho(\theta_A; \theta_B) = 1 \quad \text{or} \quad \lim_{\theta_A \to \underline{\Theta}} \rho(\theta_A; \theta_B) = 0, \quad \text{for every } \theta_B. \tag{6}$$

It is worth stressing that, even if the symmetric assumption $A1$ holds, $\rho \to 1$ as $\theta_A \to \overline{\Theta}$ does not imply that $\rho \to 0$ as $\theta_A \to \underline{\Theta}$ and vice-versa (see, e.g., ρ_{PW} in (1)).

If ρ satisfies (6) or if the variance function of the statistical model is unbounded, then the asymptotic variance σ_ρ^2 tends to diverge and the quality of the CLT approximation could be damaged, thus compromising any likelihood-based inferential procedure. This translates in both i) unreliable asymptotic confidence intervals and ii) anomalous behaviour of the power of the Wald test.

4.1 Confidence Intervals

The following Theorem shows the drawbacks of the asymptotic likelihood-based confidence intervals, that could degenerate not only for statistical models with unbounded variance, but also when the chosen target is characterized by a strong ethical component, i.e., if ρ satisfies (6).

Theorem 1 *The asymptotic variance σ_ρ^2 and the width of the asymptotic $CI(\Delta)_{1-\alpha}$ diverge if the variance function is unbounded, i.e. when $\overline{\Theta} = \infty$ and $\lim_{\theta \to \overline{\Theta}} v(\theta) = \infty$, or if ρ is chosen so that*

$$\lim_{\theta_A \to \overline{\Theta}} \rho(\theta_A; \theta_B) = 1 \quad or \quad \lim_{\theta_A \to \underline{\Theta}} \frac{v(\theta_A)}{\rho(\theta_A; \theta_B)} = \infty, \quad for\ every\ \theta_B \in \Theta.$$

In particular, for exponential and Poisson data, the width of $CI(\Delta)_{1-\alpha}$ *diverges as* Δ *grows regardless of the chosen target, while for normal homoscedastic outcomes, the asymptotic CI degenerates for every target satisfying* (6)*. As regards binary trials,* $CI(\Delta)_{1-\alpha}$ *degenerates under* ρ_{PW}*, while it does not diverge adopting* ρ_R*.*

Proof The proof follows directly from (3) by noticing that condition $\lim_{\theta_A \to \underline{\Theta}} \rho(\theta_A; \theta_B) = 0$ for every $\theta_B \in \Theta$ is only necessary but not sufficient, since the variance function could vanish as $\theta_A \to \underline{\Theta}$. For normal homoscedastic, exponential and Poisson data the proof is straightforward. For binary trials, under the PW target, the asymptotic $CI(\Delta)_{1-\alpha}$ degenerates, since $\lim_{\theta_A \to \overline{\Theta}} \rho_{PW}(\theta_A; \theta_B) = 1$ for every $\theta_B \in (0; 1)$ (although $\lim_{\theta_A \to \underline{\Theta}} \rho_{PW}(\theta_A; \theta_B) = (1 - \theta_B)/(2 - \theta_B) > 0$). Adopting ρ_R instead, $CI(\Delta)_{1-\alpha}$ does not diverge since, for every $\theta_B \in (0; 1)$, $\lim_{\theta_A \to \overline{\Theta}} \rho_R(\theta_A; \theta_B) = (1 + \theta_B)^{-1} < 1$ and $\lim_{\theta_A \to \underline{\Theta}} \rho_R(\theta_A; \theta_B) = 0$, but $\lim_{\theta_A \to \underline{\Theta}} v(\theta_A)/\rho_R(\theta_A; \theta_B) = \theta_B < 1$. □

The divergence of the asymptotic CIs strongly depends on the speed of convergence of the target to 0 or 1. For instance, taking into account ρ_N in (2), this can be severely accentuated by the effect of the tuning constant, since T induces a scaling effect by contracting/expanding the treatment difference Δ (for $T > 1$ or $T < 1$, respectively). Thus, small choices of T may deteriorate the quality of the CLT approximation as well as accelerate the divergence of the asymptotic variance σ^2_ρ, even for values of θ_A close to θ_B (i.e., for values of Δ close to 0) and not only as θ_A tends either to $\underline{\Theta}$ or $\overline{\Theta}$).

Example 1 In order to stress how small values of T could severely undermine the precision of likelihood-based inferential procedure, we perform a simulation study with 100000 normal homoscedastic trials ($v = 1$) by employing ERADE ($\gamma = 0.5$) with $n = 250$. Taking into account ρ_N, Fig. 1 shows the simulated distributions of the MLE $\hat{\Delta}_n$, as Δ and T vary, while Table 1 summarizes the behaviour of the simulated 95% asymptotic confidence intervals for Δ, where Lower (L) and Upper (U) bounds are obtained by averaging the endpoints of the simulated trials (within brackets the corresponding theoretical values derived by (4)).

When $\Delta = 0$, low values of T severely damage the CLT approximation leading to a non-negligible increase of the density in the tails; whereas, for $\Delta > 0$ the distribution of $\hat{\Delta}_n$ presents a positive skewness, regardless of the value of T.

For $T \geq 1$, analytical and simulated confidence bounds are quite close; however, as Δ grows, the impact of the skewness affects the quality of the CLT approximation. Regardless of Δ, small values of T severely damage the accuracy of the $CI(\Delta)_{0.95}$, that tends to diverge extremely fast. The empirical coverage confirms the above-mentioned behaviour and tends to 1 as the width of the intervals grows. Moreover, as showed by many authors (see, e.g., Coad and Woodroofe 1998), although asymptotically unbiased, the MLEs under RAR procedures are biased for finite samples. Even for $n = 250$, $\hat{\Delta}_n$ tends to overestimate Δ for positive values of the treatment difference and this effect is exacerbated for low values of T.

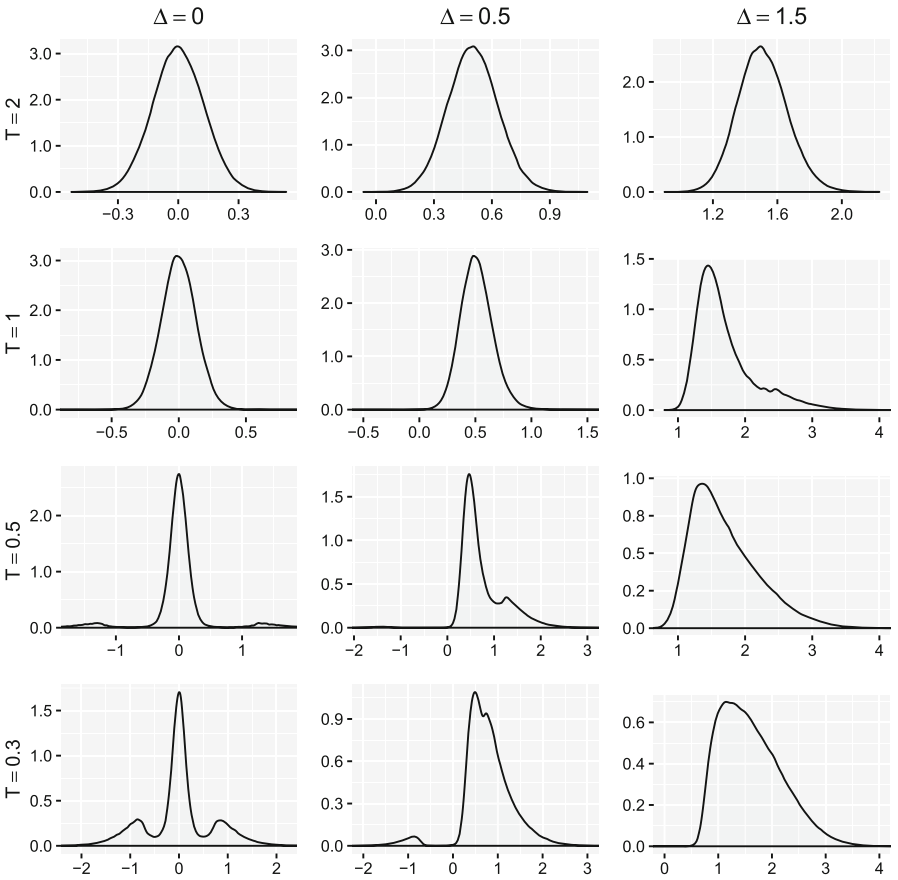

Fig. 1 Simulated distribution of $\hat{\Delta}_n$ adopting ρ_N as T and Δ vary

4.2 Hypothesis Testing

Taking now into account hypothesis testing, for every fixed value of the nuisance parameter $\theta_B \in \Theta$ (and $v \in \mathbb{R}^+$ for normal homoscedastic data), the power function (5) is governed by the non-negative function

$$t_\rho(\Delta) = \frac{\Delta}{\sigma_\rho} = \frac{\theta_A - \theta_B}{\sqrt{\frac{v(\theta_A)}{\rho(\theta_A;\theta_B)} + \frac{v(\theta_B)}{1-\rho(\theta_A;\theta_B)}}}, \quad \theta_A - \theta_B \geq 0. \tag{7}$$

Notice that the Wald test could present inflated type-I errors. Indeed, when $\theta_A = \theta_B$, from assumption $A1$, $\rho(\boldsymbol{\theta}) = 1 - \rho(\boldsymbol{\theta}) = 1/2$ and therefore $t_\rho(0) = 0$ for every $\theta_B \in \Theta$ regardless of the chosen target. Moreover, since in this case $\sigma_\rho = 2\sqrt{v(\theta_B)}$, inflated type-I errors could be present only if $v(\theta_B) \simeq 0$. This is the reason why a slightly inflation is detected in several simulation studies of both binary trials with low

Table 1 Likelihood-based simulated asymptotic $CI(\Delta)_{0.95}$ by adopting ρ_N as T and Δ vary

	Δ												
		0				0.5				1.5			
		L	$\hat{\Delta}_n$	U	EC	L	$\hat{\Delta}_n$	U	EC	L	$\hat{\Delta}_n$	U	EC
T	2	−0.25	0.00	0.25	0.95	0.24	0.50	0.76	0.96	1.16	1.51	1.86	0.98
		(−0.25)		(0.25)		(0.25)		(0.75)		(1.20)		(1.80)	
	1	−0.25	0.00	0.25	0.95	0.23	0.52	0.80	0.95	0.88	1.70	2.52	0.99
		(−0.25)		(0.25)		(0.23)		(0.77)		(1.00)		(2.00)	
	0.5	−1.44	0.00	1.44	0.97	−9.86	0.73	11.33	0.98	−2472	1.71	2476	1
		(−0.25)		(0.25)		(0.16)		(0.84)		(−1.88)		(4.88)	
	0.3	−10931	0.00	10931	1	−46537	0.77	46538	1	−326300	1.72	326303	1
		(−0.25)		(0.25)		(−0.08)		(1.08)		(−230.03)		(233.03)	

L, *U* average lower and upper simulated bounds (theoretical endpoints in brackets), *EC* empirical coverage

success probabilities and normal trials with small values of v (Zhang and Rosenberger 2012; Atkinson and Biswas 2014; Rosenberger and Lachin 2015).

Under the alternative hypothesis, the power could exhibit anomalous behaviour, especially when ρ has a strong ethical skew. In particular, we shall show that, for a given statistical model, some target allocations may induce a non monotonic power—that could also degenerate as the difference between the treatment effects grows—making the Wald test not consistent. Indeed, for every size, if $t_\rho(\Delta)$ in (7) vanishes as Δ grows, from (5) the power tends to $\Phi(-z_{1-\alpha}) = \alpha$ (i.e., the significance level), as the following Theorem shows.

Theorem 2 *When $\overline{\Theta} < \infty$, if $\lim_{\theta_A \to \overline{\Theta}} \rho(\theta_A; \theta_B) = 1$ for every $\theta_B \in \Theta$, then the Wald test is not consistent. The same conclusion still holds when $\overline{\Theta} = \infty$, provided that $\lim_{\theta_A \to \infty} (\theta_A - \theta_B)^2 [1 - \rho(\theta_A; \theta_B)] = 0$, for every $\theta_B \in \Theta$. In particular, for binary trials the Wald test is consistent under ρ_R, while it is not adopting ρ_{PW}. Taking into account Poisson, exponential and normal homoscedastic models, ρ_R guarantees the consistency of the Wald test, while ρ_N induces the inconsistency of the test.*

Proof Given a chosen target ρ, the Wald test is not consistent when $t_\rho(\Delta)$ in (7) vanishes as Δ grows. For $\overline{\Theta} < \infty$, from Theorem 1 this is satisfied iff $\lim_{\theta_A \to \overline{\Theta}} \rho(\theta_A; \theta_B) = 1$ for every $\theta_B \in \Theta$. For $\overline{\Theta} = \infty$, the same conclusion still holds provided that as $\theta_A \to \infty$, σ_ρ^2 diverges faster than θ_A^2. Since for the NQ class the variance function $v(\cdot)$ is at most quadratic, this holds iff $\lim_{\theta_A \to \infty} (\theta_A - \theta_B)^2 \{v(\theta_B)/[1 - \rho(\theta_A; \theta_B)]\}^{-1} = \lim_{\theta_A \to \infty} (\theta_A - \theta_B)^2 [1 - \rho(\theta_A; \theta_B)] = 0$, for every $\theta_B \in \Theta$. For binary trials, assuming the PW target in (1) the power tends to α as Δ grows, since $\lim_{\theta_A \to \overline{\Theta}} \rho_{PW}(\theta_A; \theta_B) = 1$, for every $\theta_B \in (0; 1)$. Whereas, adopting ρ_R, $\lim_{\theta_A \to \overline{\Theta}} \rho_R(\theta_A; \theta_B) = (1 + \theta_B)^{-1} < 1$ for every $\theta_B \in (0; 1)$ and therefore the test is consistent. Taking into account Poisson, exponential and normal homoscedastic models, adopting ρ_R the test is consistent since $\lim_{\theta_A \to \infty} (\theta_A - \theta_B)^2 [1 - \rho_R(\theta_A; \theta_B)] = \theta_B (\theta_A - \theta_B)^2 (\theta_A + \theta_B)^{-1} = \infty$ for every $\theta_B \in \mathbb{R}$ (even if $\lim_{\theta_A \to \infty} \rho_R(\theta_A; \theta_B) = 1$). By using ρ_N the test is not consistent since $\lim_{\theta_A \to \infty} (\theta_A - \theta_B)^2 [1 - \rho_N(\theta_A; \theta_B)] = 0$ for every $\theta_B \in \mathbb{R}$. □

Remark 2 Although condition $\lim_{\theta_A \to \overline{\Theta}} \rho(\theta_A; \theta_B) = 1$ is always necessary for the inconsistency of the Wald test, for binary trials it is also sufficient, making the PW rule unsuitable for likelihood-based inference. Excluding the binary case, in order to reliably apply the Wald statistic, ρ should satisfy $\lim_{\theta_A \to \infty} (\theta_A - \theta_B)^2 [1 - \rho(\theta_A; \theta_B)] > 0$ for every $\theta_B \in \Theta$.

Remark 3 Although our approach complements the one of Yi and Li (2018), Theorems 1 and 2 clearly conflict with their results. In particular, the authors show that the Wald statistic achieves the upper bound of the asymptotic power and derive the rates of coverage error probability of the corresponding confidence intervals. Their results depend on the boundedness of the remainder term in the Taylor expansion of Lemma 1 in Yi and Li (2018), where the authors state that if $\rho \in (0; 1)$ then there exists $r \in (0; 1/2]$ such that $r \leq \rho \leq 1 - r$. However, this condition does not hold for targets satisfying (6) (for instance, $\nexists r \in (0; 1/2]$ bounding ρ_N).

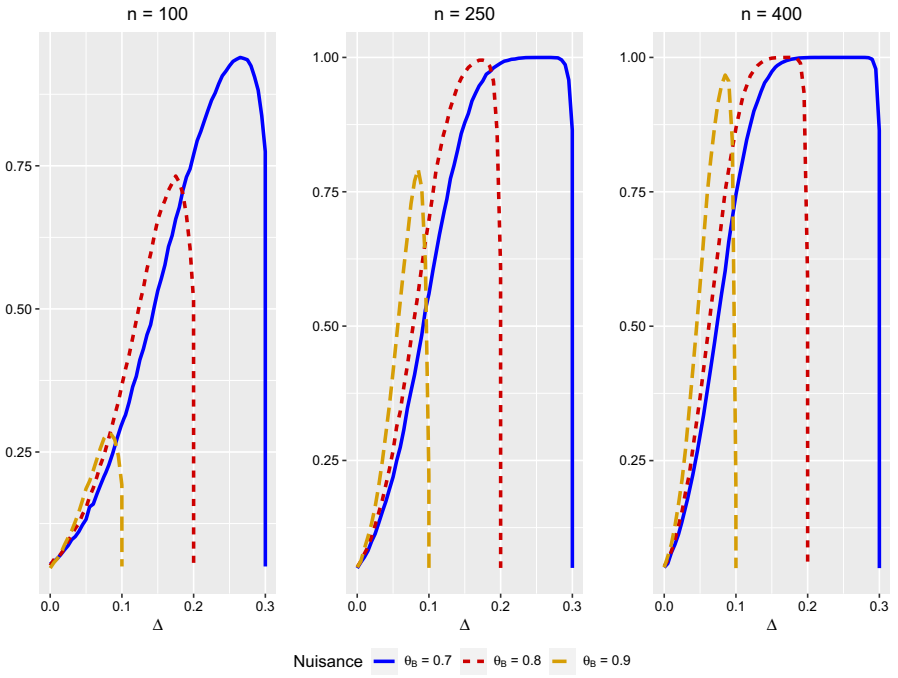

Fig. 2 Simulated power of the Wald test adopting ρ_{PW} as θ_B and n vary

Example 2 To underline how the adoption of the PW target could severely undermine the reliability of the Wald test, we perform a simulation study with 100,000 binary trials by employing ERADE ($\gamma = 0.5$). Figure 2 shows the simulated power as Δ varies for $\theta_B = 0.7, 0.8$ and 0.9 for different sample sizes.

As theoretically proved, the power tends to the significance level α regardless of the sample size. Moreover, the power function is decreasing not only at $\theta_A \approx 1$ but also for smaller and potentially crucial differences between the treatment effects, especially for small samples. For instance, when $n = 100$, for $\theta_B = 0.9$ the maximum power is about 25% attained at $\Delta = 0.07$ (i.e., $\theta_A = 0.97$), while for $\theta_B = 0.8$ the power is always lower than 75% and rapidly decreases for $\Delta \geq 0.16$. Even with $n = 250$, the power does not reach 1 when $\theta_B > 0.8$; although such a degenerating behaviour is attenuated as the sample size increases, it still persists also for $n = 400$.

An additional drawback of the PW target is related to its functional form. Indeed, although condition $A2$ is satisfied (namely, ρ_{PW} is decreasing in θ_B and therefore $1 - \rho_{PW}$ is increasing in θ_B), for any fixed difference $\Delta = \theta_A - \theta_B$, the allocation to B is decreasing in θ_B as the following table shows.
Indeed, the PW target could be rewritten as

$$\rho_{PW}(\theta_A; \theta_B) = \frac{1 - \theta_B}{2(1 - \theta_B) - (\theta_A - \theta_B)}$$

Table 2 The behaviour of the treatment allocation proportions adopting ρ_{PW} for $\Delta = 0.2$

θ_A	θ_B	ρ_{PW}	$1-\rho_{PW}$
0.9	0.7	0.750	0.250
0.8	0.6	0.667	0.333
0.5	0.3	0.583	0.417
0.3	0.1	0.563	0.437

and therefore, for a fixed difference $\theta_A - \theta_B$, its derivative wrt θ_B is

$$\frac{\theta_A - \theta_B}{[2(1-\theta_B) - (\theta_A - \theta_B)]^2} > 0, \quad \theta_A > \theta_B$$

leading to a negative derivative wrt θ_B of $1-\rho_{PW}$ (i.e., the target allocation of treatment B).

Besides consistency, an additional natural requirement of the test is that the power should be monotonically increasing in Δ (i.e., in θ_A for every $\theta_B \in \Theta$), in order to identify with high precision the best treatment as its relative superiority grows. From (7), provided that ρ is differentiable, the power of the Wald test is increasing iff, for every $\theta_B \in \Theta$,

$$\frac{2\sigma_\rho^2}{\theta_A - \theta_B} \geq \frac{v'(\theta_A)}{\rho(\theta_A;\theta_B)} + \rho'_{\theta_A}(\theta_A;\theta_B)\left\{\frac{v(\theta_B)}{[1-\rho(\theta_A;\theta_B)]^2} - \frac{v(\theta_A)}{\rho^2(\theta_A;\theta_B)}\right\}, \quad \theta_A > \theta_B \tag{8}$$

where $f'_x = \partial f/\partial x$ denotes the partial derivative of f wrt x (to avoid cumbersome notation, we shall omit the subscript for the derivative of scalar functions). In addition to the statistical model, condition (8) regards the chosen target and needs to be satisfied for every $\theta_A > \theta_B$, involving the entire functional form of ρ (not only its limits and the speed of convergence to them as in Theorems 1 and 2). Clearly, if the target induces the inconsistency of the test, then (8) fails to hold, instead if ρ guarantees the consistency of the test, it does not necessarily ensure the monotonicity of the power, as shown in Fig. 5. For instance, as also discussed by Baldi Antognini et al. (2018), for normal homoscedastic data $v' = 0$ and the power is increasing in Δ iff ρ is chosen so that, for every $\theta_B \in \mathbb{R}$

$$\rho(\theta_A;\theta_B)[1-\rho(\theta_A;\theta_B)] \geq (\theta_A-\theta_B)[\rho(\theta_A;\theta_B)-1/2]\rho'_{\theta_A}(\theta_A;\theta_B), \quad \theta_A > \theta_B. \tag{9}$$

Clearly, this condition fails to hold for ρ_N, while it is satisfied by ρ_R. Analogously, for binary trials adopting ρ_{PW} the power of the Wald test is not monotonically increasing. Indeed, condition (8) can be restated as

$$\frac{2}{\theta_A - \theta_B} - \frac{\theta_A - \theta_B}{(2-\theta_A-\theta_B)(1-\theta_A)} \geq \frac{(1-2\theta_A)(1-\theta_A) + \frac{(\theta_A-\theta_B)(\theta_A+\theta_B-1)(1-\theta_B)}{2-\theta_A-\theta_B}}{\theta_A(1-\theta_A)^2 + \theta_B(1-\theta_B)^2},$$

where, for every $\theta_B \in (0; 1)$, as θ_A tends to $\overline{\Theta} = 1$ the LHS tends to $-\infty$ while the RHS tends to $1/(1-\theta_B) > 0$.

Proposition 1 *For normal, binary, exponential and Poisson data, ρ_R always guarantees that the power of the Wald test is monotonically increasing in Δ.*

Proof For the normal homoscedastic model, inequality (9) is trivially satisfied since $2\theta_A(\theta_A + \theta_B) \geq (\theta_A - \theta_B)^2$ for every $\theta_A \geq \theta_B > 0$. For Poisson and exponential data, condition (8) still holds since, for every $\theta_B \in \mathbb{R}^+$,

$$\frac{\theta_B}{\theta_A + \theta_B} \leq 1 \leq 1 + \frac{4\theta_A\theta_B}{\theta_A^2 - \theta_B^2}, \quad \theta_A \geq \theta_B > 0.$$

In the context of binary trials, inequality (8) becomes

$$\theta_A\{\theta_A - \theta_B + 2\theta_B(2 - \theta_A - \theta_B)\} \geq 0$$

which is clearly satisfied for $1 > \theta_A \geq \theta_B > 0$. □

As previously discussed, ρ_R is able to preserve the fundamental properties of the Wald test, namely the consistency and the monotonicity of its power. However, this target strongly depends on the nuisance parameter θ_B; indeed, for a fixed difference Δ, as θ_B grows $\rho_R(\theta_A; \theta_B) \to 1/2$ and, therefore, its ethical improvement tends to vanish as well as the induced power. For instance, from (7), under exponential outcomes $t_{\rho_R}(\Delta) = \Delta/(\theta_A + \theta_B)$, while for Poisson data $t_{\rho_R}(\Delta) = \Delta/\sqrt{2(\theta_A + \theta_B)}$ and both of them vanish as θ_B grows, for every fixed θ_A. Figure 3 confirms graphically the crucial role played by θ_B in terms of power: given a difference $\Delta = 0.5$, under the exponential model the power decreases from 0.94 to 0.10 as θ_B grows from 1 to 10 (while for Poisson data it goes from 0.97 to 0.34).

5 A possible solution for likelihood-based inference: the re-scaled target

From Theorems 1 and 2, it is quite evident that some anomalous behaviours could be prevented by assuming a target that is not characterized by a strong ethical component, namely under which (6) fails to hold. Indeed, if the target is chosen so that $0 < l_1 \leq \rho(\boldsymbol{\theta}) \leq l_2 < 1$ for every $\boldsymbol{\theta}$, then the Wald test is consistent, while $CI(\Delta)_{1-\alpha}$ does not diverge provided that $v(\cdot)$ is bounded.

Moreover, to mitigate the effects of the nuisance parameters, a possible way consists in adopting targets that depend only on the treatment difference Δ and not on θ_B, namely $\rho = \rho^\star(\Delta)$; however, this is only a partial solution, since the nuisance parameter affects any likelihood-based inferential procedure through the variance function. In this setting, assumptions $A1$-$A2$ become

A: $\rho^\star$ is continuous and increasing with $\rho^\star(\Delta) = 1 - \rho^\star(-\Delta)$.

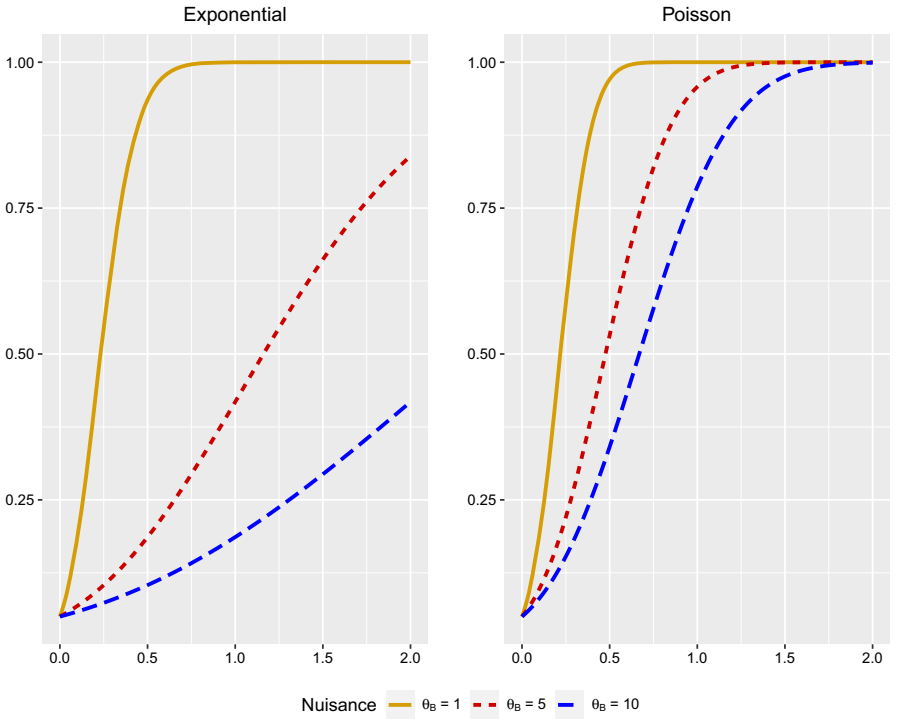

Fig. 3 Simulated power of the Wald test for exponential and Poisson data adopting ρ_R with $n = 250$

For instance, under normal, Poisson and exponential data $\rho^\star$ could be interpreted as the cdf of a continuous r.v. with support in $\mathbb{R}$ and symmetric around 0, as $\rho_N = \rho_N^\star$ in (2) (Baldi Antognini et al. 2018). While, for binary trials, the target

$$\rho_G^\star(\Delta) = \frac{1}{2} + \frac{\omega\Delta}{2(2-\omega)}, \quad \Delta \in (-1; 1),$$

is the asymptotic allocation of the doubly-adaptive weighted difference design, suggested by Geraldes et al. (2006). It is obtained by a suitable weighted combination of two linear randomization functions, one for ethics and the other dictated by balance, where $\omega \in [0; 1]$ reflects the relative importance of ethics. Note that $\rho_G^\star$ guarantees the consistency of the Wald test and the reliability of the CIs, since as $\theta_A \to \overline{\Theta} = 1$, $\rho_G^\star(\Delta) \to (2-\omega)^{-1} < 1$, while as $\theta_A \to \underline{\Theta} = 0$, $\rho_G^\star(\Delta) \to (1-\omega)/(2-\omega) > 0$, for every $\omega < 1$.

By combining these suggested solutions, even when the desired $\rho^\star$ is characterized by a strong ethical improvement, a possible way to overcome some degeneracies consists in re-scaling the target, namely by letting

$$\rho_r^\star(\Delta) = 1 - r + \rho^\star(\Delta)(2r-1), \quad \text{with} \quad r \in (1/2; 1). \tag{10}$$

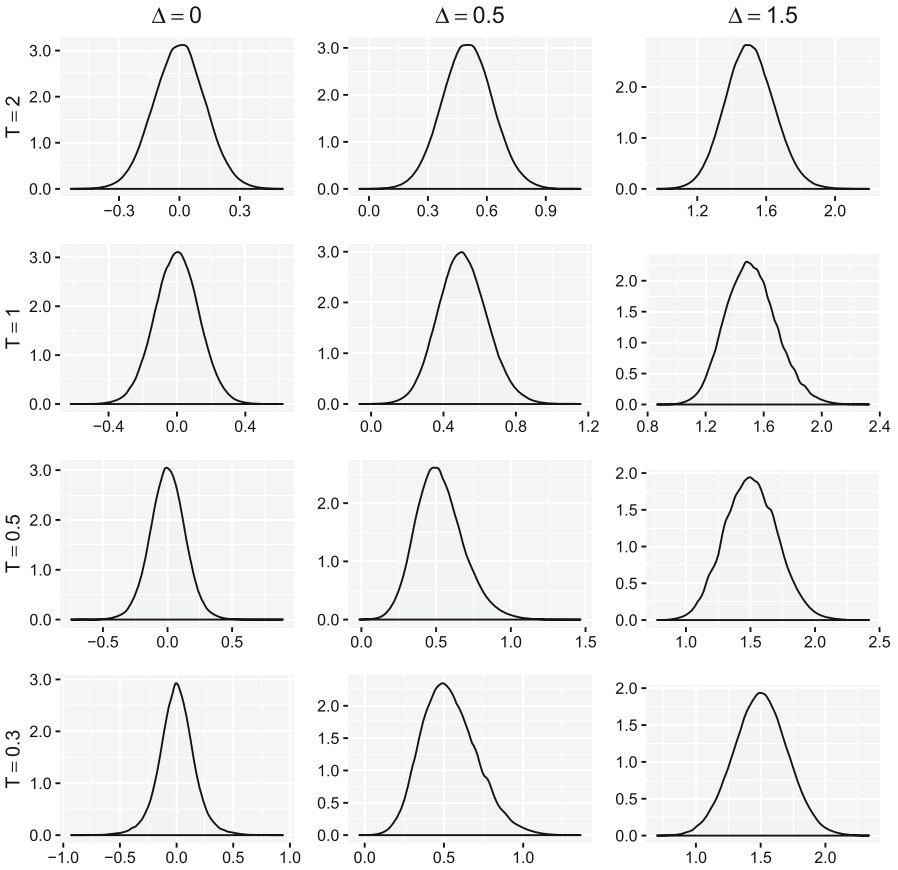

Fig. 4 Simulated distribution of $\hat{\Delta}_n$ adopting $\rho^{\star}_{N_r}$ ($r = 0.9$) as T and Δ vary

Transformation (10) simply contracts the image of $\rho^{\star}$, which is re-scaled in $[1-r; r]$, while preserving its functional form. Clearly, for $r = 1$ no re-scaling transformation is applied, namely $\rho^{\star}_1(\Delta) = \rho^{\star}(\Delta)$, while the case $r = 1/2$ corresponds to the balanced allocation.

Although the anomalous scenarios induced by the unboundedness of the variance function—i.e., by the statistical model—cannot be overcome, by adopting $\rho^{\star}_r$ some degeneracies caused by the target could be avoided, since the Wald test is consistent and $CI(\Delta)_{1-\alpha}$ does not diverge.

Remark 4 Since under condition $C1$ the treatment allocation proportion π_n of a RAR design is a consistent estimator of the target, another possible way to overcome some drawbacks of likelihood-based asymptotic procedures consists in estimating σ^2_{ρ} by $\breve{\sigma}^2_n = \hat{v}_{An}/\pi_n + \hat{v}_{Bn}/[1-\pi_n]$. Indeed, given a starting sample of $2n_0$ assignments, for any fixed n, $\pi_n \in [\eta_n; 1-\eta_n]$, where $\eta_n = n_0/n \in (0; 1/2)$ is the percentage of (non-adaptive) allocations initially made on either treatment. In practice, $\pi_n \simeq \rho(\hat{\boldsymbol{\theta}}_n)(1-\eta_n) + [1-\rho(\hat{\boldsymbol{\theta}}_n)]\eta_n$, that substantially corresponds to assume a re-scaled

target with $r = r(n) = 1 - \eta_n$. Unfortunately, this approach could be useful only for clinical trials where η_n is non-negligible (i.e., for quite small samples), otherwise n_0 should be chosen as an increasing function of n (Baldi Antognini et al. 2018).

Although the re-scaling correction could also be applied to targets depending on nuisance parameters, in general it does not protect against the non monotonicity of the power function discussed in Section 4. However, since $0 < \rho'_{r_{\theta_A}} = (2r-1)\rho'_{\theta_A} < \rho'_{\theta_A}$, then monotonicity condition (8) tends to be satisfied as r decreases (namely when the target tends to be balanced); thus, as it will be shown in Examples 3 and 4, this drawback could be strongly mitigated/overcome by re-scaling the target with a proper choice of r.

Example 3 To show how a re-scaled target not depending on the nuisance could improve the precision of likelihood-based inference, we perform a simulation study in the same setting of Example 1 by adopting $\rho^{\star}_{N_r}$ with $r = 0.9$. Figure 4 shows the simulated distributions of $\hat{\Delta}_n$ as T and Δ vary, while Table 3 summarises the behaviour of the simulated 95% asymptotic confidence interval for Δ, where Lower (L) and Upper (U) bounds are obtained by averaging the endpoints of the simulated trials (within brackets the theoretical values derived by (4)).

Adopting $\rho^{\star}_{N_r}$, the reliability of the $CI(\Delta)_{0.95}$ drastically increases: analytical and simulated bounds almost coincide for every value of T and Δ. Although for small values of T (i.e., for a high ethical component) the width of the confidence intervals slightly grows, this does not compromise the inferential precision. By limiting the skewness and the variability of the MLE's distribution, the re-scaled target significantly improves the accuracy of the asymptotic confidence intervals, also confirmed by the empirical coverage which is always quite close to the nominal value. Note that the re-scaling correction seems also to reduce the bias of the MLEs, in particular for higher values of the treatment difference.

As regards hypothesis testing, Fig. 5 shows the power of the Wald test adopting $\rho^{\star}_{N_r}$ as T and r vary (the case $r = 1$ corresponds to $\rho^{\star}_{N}$).

Regardless of the values of T, the re-scaled target (i.e., $r < 1$) always preserves the consistency of the test. However, this target does not satisfies condition (9) and, for small values of T, the decreasingness of the power is accentuated as r tends to 1. Even for $T = 0.5$ or $T = 0.3$, by selecting $r \leq 0.95$, monotonicity condition (9) is fulfilled; in this way the ethical component of the target could be strongly improved without compromising inference.

Example 4 Ideally, the re-scaling correction should be applied to targets with a strong ethical skew—i.e., satisfying (6)—that (i) fulfill (8) to guarantee a monotonic power function of the Wald test and (ii) depend on the treatment effects only through the difference Δ (to mitigate the effects of the nuisance parameters). As previously shown, when adopting ρ_{PW} none of these conditions is satisfied; however, the re-scaled version ρ_{PW_r} could still overcome or mitigate some of the above-mentioned drawbacks. To see this, we perform a simulation study in the same setting of Example 2, by comparing the performances of ρ_{PW} and ρ_{PW_r} with $r = 0.9$. Figure 6 shows the simulated power of the Wald test as Δ varies for $\theta_B = 0.7$, 0.8 and 0.9 for $n = 100$, 250 and 400, while Table 4 summarizes the behaviour of the simulated 95% asymptotic confidence

Table 3 Likelihood-based simulated asymptotic $CI(\Delta)_{0.95}$ adopting $\rho^{\star}_{Nr}$ ($r = 0.9$) as T and Δ vary

	Δ												
		0				0.5				1.5			
		L	$\hat{\Delta}_n$	U	EC	L	$\hat{\Delta}_n$	U	EC	L	$\hat{\Delta}_n$	U	EC
T	2	−0.25	0.00	0.25	0.95	0.24	0.50	0.76	0.95	1.17	1.51	1.84	0.98
		(−0.25)		(0.25)		(0.25)		(0.75)		(1.22)		(1.78)	
	1	−0.25	0.00	0.25	0.95	0.24	0.51	0.77	0.96	1.12	1.51	1.91	0.98
		(−0.25)		(0.25)		(0.24)		(0.76)		(1.16)		(1.84)	
	0.5	−0.25	0.00	0.25	0.95	0.22	0.53	0.83	0.95	1.05	1.50	1.95	0.97
		(−0.25)		(0.25)		(0.20)		(0.80)		(1.09)		(1.91)	
	0.3	−0.26	0.00	0.26	0.94	0.17	0.54	0.90	0.96	1.05	1.50	1.95	0.97
		(−0.25)		(0.25)		(0.14)		(0.86)		(1.09)		(1.91)	

L, *U* average lower and upper simulated bounds (theoretical endpoints in brackets), *EC* empirical coverage

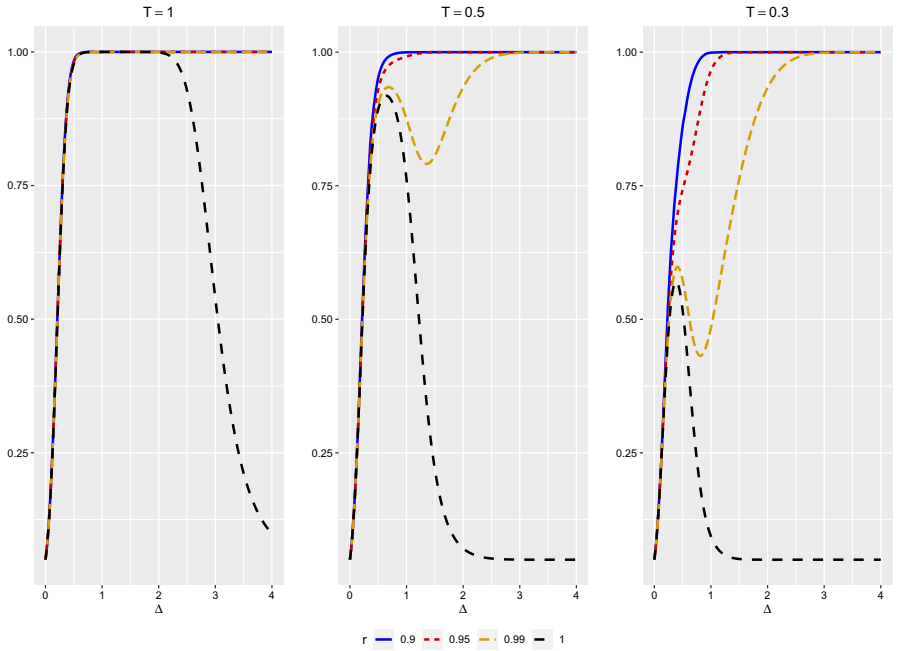

Fig. 5 Power of the Wald test for normal outcomes adopting $\rho^{\star}_{N_r}$ and $\rho^{\star}_{N}$ as T and r vary

interval for Δ, where Lower (L) and Upper (U) bounds are obtained by averaging the endpoints of the simulated trials (within brackets the theoretical values derived by (4)). If compared to ρ_{PW} (see Fig. 2), the re-scaled target ρ_{PW_r} guarantees the consistency of the Wald test, also strongly improving the behaviour of the power function. The improvement in the inferential precision is remarkable: for instance, with $n = 100$ and $\theta_B = 0.9$, for $\Delta = 0.08$ the power is about 40% with a gain of 13% wrt the non re-scaled version, while for $n = 250$ the power increases of 18%. For what concerns CIs, although ρ_{PW} performs quite well, the asymmetric distribution of the MLEs causes a right shift of the CI with a slight increase in the width (that is exacerbated for $\theta_A > 0.95$). On the other hand, the adoption of ρ_{PW_r} leads to narrower and centered CIs with a correct empirical coverage.

6 Discussion

This paper explores in depth the limitations of the likelihood-based approach for RAR experiments, in terms of asymptotic confidence intervals and hypothesis testing. Although clinical trials represent one of the most actual fields of application of this methodology (because of the main concern about the ethical impact on the subjects' care), RAR procedures could be a useful tool for local optimality problems also in different contexts like, e.g., industrial experiments. First of all, we show that some RAR rules as well as some targets can compromise the asymptotic likelihood-based

Table 4 Likelihood-based simulated asymptotic $CI(\Delta)_{0.95}$ by adopting ρ_{PW} and ρ_{PW_r} $(r = 0.9)$ with $n = 250$ and $\theta_B = 0.7$, as Δ varies

	Δ												
		0				0.15				0.25			
		L	$\hat{\Delta}$	U	EC	L	$\hat{\Delta}$	U	EC	L	$\hat{\Delta}$	U	EC
T	ρ_{PW}	−0.12	0.00	0.12	0.95	0.05	0.16	0.28	0.95	0.13	0.27	0.44	0.96
		(−0.11)		(0.11)		(0.04)		(0.26)		(0.10)		(0.40)	
	ρ_{PW_r}	−0.11	0.00	0.11	0.95	0.04	0.15	0.26	0.95	0.13	0.26	0.39	0.95
		(−0.11)		(0.11)		(0.04)		(0.26)		(0.12)		(0.38)	

L, U average lower and upper simulated bounds (theoretical endpoints in brackets), *EC* empirical coverage

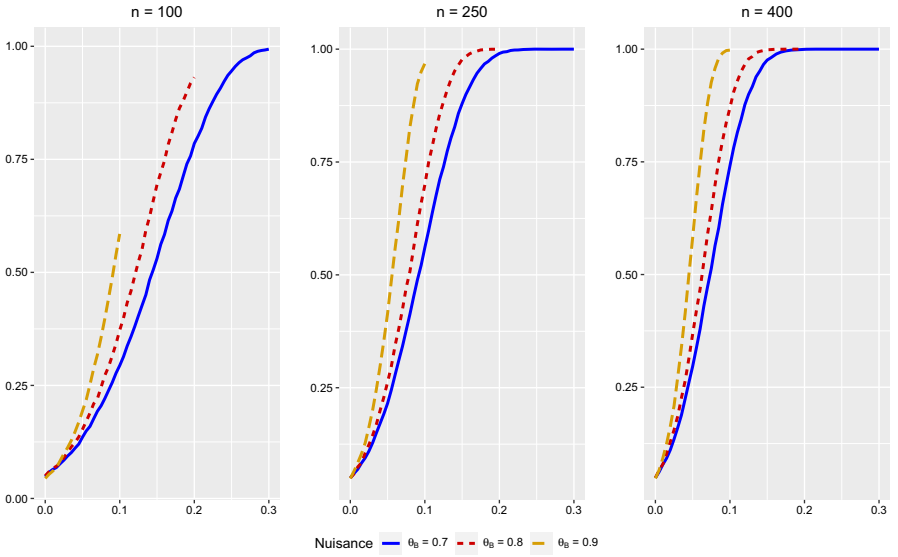

Fig. 6 Simulated power of the Wald test adopting ρ_{PW_r} (with $r = 0.9$) as θ_B and n vary

inference, inducing a degenerating behaviour of the power of the Wald test and unreliable CIs. This is particularly true when the empirical evidence strongly suggests the superiority of one treatment wrt the other or when the ethical component of the target is remarkable, since this could induce the target to approach either 0 or 1. Furthermore, these anomalies may also be caused by statistical models with unbounded variance, and inference could also be strongly compromised due to the effect of nuisance parameters.

Our results show that, in general, ρ_R is able to preserve the fundamental properties of hypothesis testing, because it guarantees the consistency of the Wald test as well as the monotonicity of its power; however, its dependence on the nuisance parameter could damage the inferential precision. On the other hand, the PW rule confirms its practical inadequacy since i) the asymptotic CIs diverge and ii) the power of the Wald test is decreasing and tends to the significance level as the difference between the treatment effects grows, thus severely undermining the inferential precision.

Inspired by the common practice of superimposing a minimum percentage of allocations for each treatment, several authors have recently taken into account RAR procedures with a minimum prefixed threshold in the assignments to avoid possible degeneracies (see Tymofyeyev et al. 2007; Sverdlov et al. 2011; Sverdlov and Rosenberger 2013; Villar et al. 2015b). In this paper, we prove how a re-scaling correction of the target could preserve some of the fundamental properties of likelihood-based inference. In particular, we show that, by adopting a re-scaled target, the consistency of the Wald test and the reliability of the CIs are ensured (provided that the variance function is bounded), even with a high ethical component. Moreover, choosing a suitable threshold r significantly improves the accuracy of the asymptotic likelihood-based CIs (also confirmed by the empirical coverage which is quite closed to the nominal value)

and overcomes the non monotonicity of the power function. Generally, a choice of $r = 0.9$ preserves the inferential accuracy, regardless of the statistical model and of the adopted target. As regards ρ_N, $r = 0.9$ matched with $T \geq 0.5$ guarantees good performances in terms of both ethics and inference. Clearly, these results could also be applied to the class of Bayesian RAR designs, where frequentist likelihood-based inference is performed at the end of the trial. Indeed, Bayesian RAR procedures could also present possible degeneracy in the treatment allocation proportions and therefore a re-scaling correction could represent a valid tool for inference. For instance, as recently discussed by Villar et al. (2018) for the case of several treatments, superimposing a minimum percentage of allocation to the control group produces robust inference by preserving type-I errors even in the case of time trends.

However, in some circumstances, other critical issues related to the unboundedness of the variance function and the effect of the nuisance parameters cannot be circumvented by simply re-scaling the target. This is the case, for example, of ρ_R and ρ_Z under exponential and Poisson responses, respectively (namely, the corresponding Neyman allocations); their re-scaled versions, while maintaining the same inferential performances of the non re-scaled counterparts, do not protect against neither the strong dependence on the nuisance parameter nor the unboundedness of the variance function. In such situations, alternative inferential approaches could be preferable and one of the most promising is randomization-based inference (Wei 1988; Rosenberger 1993). Under this framework, the equality of treatment groups corresponds to an allocation in which the assignments are unrelated to the responses; inference is thus carried out by computing the distribution of the treatment allocations conditionally on the observed outcomes, that are treated as deterministic. Since the distribution of the test depends on the chosen RAR rule, exact results are quite few and, generally, p-values and the endpoints of confidence intervals are computed by Monte Carlo methods (for recent contributions see Wang et al. 2020 for randomization tests and Wang and Rosenberger 2020 for randomization-based interval estimation).

Our results are focussed on the case of two treatments, but a suitable extension to the multi-armed case could be very relevant. Indeed, for $K > 2$ treatments, multiple comparisons between the treatment groups should be taken into account for inference (some of them with possibly different importance, due to e.g., previous knowledge about a gold standard, the presence of a control arm). As showed by Tymofyeyev et al. (2007), Sverdlov et al. (2011) and Baldi Antognini et al. (2019), the optimal design maximizing the power of the Wald test of homogeneity is a degenerate allocation involving only the best and the worst treatments without observations on the intermediate ones (here, the treatment order is the usual stochastic order between random variables). This clearly leads to unreliable inference about the treatment contrasts and, at the same time, problems also arise from the ethical viewpoint, since more than half of the patients could be assigned to the less effective treatment. A re-scaling transformation can still be applied for multidimensional target $\boldsymbol{\rho}^t = (\rho_1, \ldots, \rho_K)$ with $\rho_i \geq 0$ and $\sum_{i=1}^K \rho_i = 1$ by letting, analogously to (10),

$$\rho_{ir} = (1 - r)(1 - \rho_i)/(K - 1) + r\rho_i, \quad \text{for } i = 1, \ldots, K, \quad \text{with } r \in (1/K; 1),$$

which ensures that $\rho_{ir} \in [(1-r)/(K-1); r]$ and $\sum_{i=1}^{K} \rho_{ir} = 1$. However, in this setting the impact of the re-scaling correction in terms of estimation efficiency and power needs to be studied. This topic, as well as proper comparisons between likelihood-based and randomization-based inference, is left for future research.

Acknowledgements We are very grateful to the referees for their helpful comments, which led to a substantially improved version of the paper. We also wish to thank Professor Laura Anderlucci for the helpful discussion on this article and her constructive suggestions.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Atkinson AC, Biswas A (2005a) Adaptive biased-coin designs for skewing the allocation proportion in clinical trials with normal responses. Stat Med 24:2477–2492

Atkinson AC, Biswas A (2005b) Bayesian adaptive biased-coin designs for clinical trials with normal responses. Biometrics 61:118–125

Atkinson AC, Biswas A (2014) Randomised response-adaptive designs in clinical trials. Chapman & Hall/CRC Press, Boca Raton

Azriel D, Mandel M, Rinott Y (2012) Optimal allocation to maximize power of two-sample tests for binary response. Biometrika 99:101–113

Baldi Antognini A, Giovagnoli A (2005) On the large sample optimality of sequential designs for comparing two or more treatments. Seq Anal 24:205–217

Baldi Antognini A, Giovagnoli A (2010) Compound optimal allocation for individual and collective ethics in binary clinical trials. Biometrika 97:935–946

Baldi Antognini A, Giovagnoli A (2015) Adaptive Designs for Sequential Treatment Allocation. Chapman & Hall/CRC Biostatistics, Boca Raton

Baldi Antognini A, Vagheggini A, Zagoraiou M (2018) Is the classical wald test always suitable under response-adaptive randomization? Stat Methods Med Res 27:2294–2311

Baldi Antognini A, Novelli M, Zagoraiou M (2019) Optimal designs for testing hypothesis in multiarm clinical trials. Stat Methods Med Res 28:3242–3259

Bandyopadhyay U, Biswas A (2001) Adaptive designs for normal responses with prognostic factors. Biometrika 88:409–419

CHMP (2007) Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Available on line

Coad D, Woodroofe M (1998) Approximate bias calculations for sequentially designed experiments. Seq Anal 17(1):1–31

Durham SD, Flournoy N, Rosenberger WF (1997) Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. J Stat Plann Inference 60:69–76

FDA (2018) Guidance for industry. Adaptive design clinical trials for drugs and biologics (draft document). https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf, available on line

Geraldes M, Melfi V, Page C, Zhang H (2006) The doubly adaptive weighted difference design. J Stat Plann Inference 136:1923–1939

Hu F, Rosenberger WF (2006) The theory of response-adaptive randomization in clinical trials. John Wiley & Sons, New York
Hu F, Zhang LX (2004) Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. Ann Stat 32:268–301
Hu F, Zhang LX, He X (2009) Efficient randomized adaptive designs. Ann Stat 37:2543–2560
Ivanova A (2003) A play-the-winner type urn model with reduced variability. Metrika 58:1–13
Lehmann EL (1999) Elements of large-sample theory. Springer Verlag, New York
May C, Flournoy N (2009) Asymptotics in response-adaptive designs generated by a two-color, randomly reinforced urn. Ann Stat 37(2):1058–1078
Melfi V, Page C (2000) Estimation after adaptive allocation. J Stat Plann Inference 29:353–363
Melfi V, Page C, Geraldes M (2001) An adaptive randomized design with application to estimation. Can J Stat 29:107–116
Morris CN (1982) Natural exponential families with quadratic variance functions. Ann Stat 10:65–80
Rosenberger WF (1993) Asymptotic inference with response-adaptive treatment allocation designs. Ann Stat 21:2098–2107
Rosenberger WF, Hu F (1999) Bootstrap methods for adaptive designs. Stat Med 18(14):1757–1767
Rosenberger WF, Lachin JM (2015) Randomization in clinical trials: theory and practice. John Wiley & Sons, New York
Rosenberger WF, Seshaiyer E (1997) Adaptive survival trials. J Biopharm Stat 7:617–624
Rosenberger WF, Sriram TN (1996) Estimation for an adaptive allocation design. J Stat Plann Inference 59:309–19
Rosenberger WF, Flournoy N, Durham SD (1997) Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. J Stat Plann Inference 60:69–76
Rosenberger WF, Stallard N, Ivanova A, Harper CN, Ricks ML (2001) Optimal adaptive designs for binary response trials. Biometrics 57:909–913
Sverdlov O, Rosenberger WF (2013) On recent advances in optimal allocation designs in clinical trials. J Stat Theory Pract 7(4):753–773
Sverdlov O, Tymofyeyev Y, Wong WK (2011) Optimal response-adaptive randomized designs for multi-armed survival trials. Stat Med 30:2890–2910
Tymofyeyev Y, Rosenberger WF, Hu F (2007) Implementing optimal allocation in sequential binary response experiments. J Am Stat Assoc 102:224–234
Villar SS, Bowden J, Wason J (2015a) Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. Stat Sci 30(2):199
Villar SS, Wason J, Bowden J (2015b) Response-adaptive randomization for multi-arm clinical trials using the forward looking gittins index rule. Biometrics 71(4):969–978
Villar SS, Bowden J, Wason J (2018) Response-adaptive designs for binary responses: how to offer patient benefit while being robust to time trends? Pharm Stat 17:182–197
Wang Y, Rosenberger WF (2020) Randomization-based interval estimation in randomized clinical trials. Stat Med 39:2843–2854
Wang Y, Rosenberger WF, Uschner D (2020) Randomization tests for multiarmed randomized clinical trials. Stat Med 39:494–509
Wei LJ (1988) Exact two-sample permutation tests based on the randomized play-the-winner rule. Biometrika 75:603–606
Wei LJ, Durham S (1978) The randomized play-the-winner rule in medical trials. J Am Stat Assoc 73:840–843
Yi Y, Li X (2018) Response adaptive designs with asymptotic optimality. Canad J Stat 46:458–469
Yi Y, Wang X (2011) Comparison of Wald, score, and likelihood ratio tests for response adaptive designs. J Stat Theory Appl 10:553–569
Zelen M (1969) Play-the-winner rule and the controlled clinical trials. J Am Stat Assoc 64:131–146
Zhang LX, Rosenberger WF (2006) Response-adaptive randomization for clinical trials with continuous outcomes. Biometrics 62:562–569
Zhang LX, Rosenberger WF (2012) Adaptive randomization in clinical trials. In: Hinkelmann K (ed) Design and analysis of experiments, vol 3. Special Designs and Applications. Wiley, New York, pp 251–282