Contemporary Clinical Trials Communications 33 (2023) 101141

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Simulation study for evaluating an adaptive-randomisation Bayesian hybrid trial design with enrichment

Valentin Vinnat[a,*], Jean-Daniel Chiche[b], Alexandre Demoule[c,d], Sylvie Chevret[a]

[a] ECSTRRA team, INSERM U1153, Université Paris Cité, Paris, France
[b] Service de médecine intensive adulte, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland
[c] Sorbonne Université, INSERM, UMRS1158 Neurophysiologie Respiratoire Expérimentale et Clinique, France
[d] AP-HP, Groupe Hospitalier Universitaire APHP-Sorbonne Université, site Pitié-Salpêtrière, Service de Médecine Intensive et Réanimation (Département R3S), Paris, France

ARTICLE INFO

Keywords:
Response-adaptive randomisation
Marker-based strategy designs
Bayesian designs
Precision medicine

ABSTRACT

Background: As we enter the era of precision medicine, the role of adaptive designs, such as response-adaptive randomisation or enrichment designs in drug discovery and development, has become increasingly important to identify the treatment given to a patient based on one or more biomarkers. Tailoring the ventilation supply technique according to the responsiveness of patients to positive end-expiratory pressure is a suitable setting for such a design.

Methods: In the setting of marker-strategy design, we propose a Bayesian response-adaptive randomisation with enrichment design based on group sequential analyses. This design combines the elements of enrichment design and response-adaptive randomisation. Concerning the enrichment strategy, Bayesian treatment-by-subset interaction measures were used to adaptively enrich the patients most likely to benefit from an experimental treatment while controlling the false-positive rate.

The operating characteristics of the design were assessed by simulation and compared to those of alternate designs.

Results: The results obtained allowed the detection of the superiority of one treatment over another and the presence of a treatment-by-subgroup interaction while keeping the false-positive rate at approximately 5\% and reducing the average number of included patients. In addition, simulation studies identified that the number of interim analyses and the burn-in period may have an impact on the performance of the scheme.

Conclusion: The proposed design highlights important objectives of precision medicine, such as determining whether the experimental treatment is superior to another and identifying wheter such an efficacy could depend on patient profile.

1. Background

In recent years, adaptive designs have become increasingly popular because of their flexibility as an alternative to fixed randomised clinical trials. Adaptive design allows ongoing clinical trials to be modified based on information gathered during the trial without compromising the integrity and validity of the trial. These designs are attractive to investigators, sponsors,and even patients, as they can shorten the duration of trial, limiting the number of patients receiving ineffective/inferior treatments and increasing the chances of subsequent success in drug development. This was exemplified during the COVID-19 pandemic, which has fostered rapid completion of these studies, allowing us to obtain answers that rapidly influenced clinical management [1–3].

In addition to enabling early stopping due to efficacy issues or futility, adaptive designs allow for changes in operational characteristics during trials, such as dropping an arm or changing the probability of randomisation between arms. Those designs, known as response-adaptive randomisation (RAR) designs, are increasingly considered a In addition to enabling early stopping due to efficacy issues or futility, adaptive designs allow for changes in operational characteristics during trials, such as dropping an arm or changing the probability of randomisation between arms. Those designs, known as response-adaptive randomisation (RAR) designs, are increasingly considered as an alternative

* Corresponding author.
E-mail address: vinnatvalentin@gmail.com (V. Vinnat).

https://doi.org/10.1016/j.conctc.2023.101141
Received 4 August 2022; Received in revised form 22 March 2023; Accepted 12 April 2023
Available online 14 April 2023

to fixed randomisation in controlled trials. The goal of these RAR designs is to maximise the number of patients receiving optimal treatment by sequentially updating treatment allocation probabilities based on accumulated outcome data. Therefore, RAR can be considered more ethical than equal randomisation (ER) because it allows more patients to receive better treatment [4–6]while providing information on treatment efficacy. Nevertheless, the use of RAR in confirmatory trials has been questioned, notably for two-arm trials, where it may have poorer performances than fixed randomisation [7,8]

Moreover, whichever the number of arms, it is often criticized for operational issues into the running of the trial, with high probability of resulting in extreme imbalances. However, its use in multiarm trials has been shown to reach similar or better operating characteristics in terms of power and type 1 error rate in the detection of the best treatment among all [9–12].In the area of precision medicine, tailori ng the treatment given to a patient according to one or more personal characteristics appears desirable. Indeed, the plausible heterogeneity in patient response cannot always be addressed using traditional designs. It is necessary to create new statistical methodologies that best address this issue [13].

Thus, to restrict the inclusion to a subset of patients who are likely to benefit the most from the treatment during trial accrual, enrichment designs have been proposed [14,15]. The response-adaptive randomisation is potentially used by these models [16,17]. All these adaptive-randomisation designs that use biomarker or multi-marker predictive signature information differ notably according to the unavailability at the trial onset of any classifier or if the patient population may be divided into biomarker-defined subgroups in which the efficacy of the treatment is supposed to differ. In the former case, the design and analysis must incorporate a cross-validation signature for identifying sensitive patients before providing any enrichment and adaptive randomisation, such as that recently proposed by Xia et al. [17]. In the latter case, where the subsets are known at the trial onset, several proposals have been published for those so-called “stratified trial designs” or “marker-stratified designs” [18,19]. In this setting, Bayesian proposals are aimed at optimising the adaptation rules throughout the decision, framework [16] or using dynamic borrowing to assess the evidence for efficacy in a specific subgroup and an overall positive effect [20].

We placed ourselves in the setting of Marker-based Strategy Designs (MSD) [19,21,22], that is, where the trial is designed to further assess the benefit of stratifying the population into nonoverlapping subgroups, and each patient is randomly assigned either to have therapy determined by their marker status or to receive therapy independent of marker status [23]. We aimed at (i) evaluating the benefit of treatments in terms of a binary outcome measure and (ii) to test whether such an effect could depend on a patient profile.

The contributions of this paper are as follows. First, we proposed a Bayesian hybrid adaptive randomisation design for clinical trials evaluating a marker-based strategy, with sequential rules for assessing the efficacy of both the stratification and the treatment effect. This is the first novelty of our proposal, given previous MSD often placed themselves in a frequentist framework, by maximising the power of some test to detect the predictive marker effect [21,22]. Moreover, to perform response-adaptive-randomisation (RAR) we used the optimal allocation criterion proposed by Rosenberger [24], modified by plugging Bayesian estimates [25]. Such a RAR also differs from that proposed by Zhou et al. [48] and Gu et al. [26], who directly derived the allocation ratio from the estimated posterior mean of response rates. It also differs from the proposed ratio of posterior probability of superiority [27] that appears inappropriate in the setting of MSD, where the “treatment” superiority relies, rather than on whether the biomarker is of interest for treatment choice, on the two treatments to be compared. Otherwise our design allows for early decision rules, not only to stop recruitment if the treatment under study is likely to be beneficial (rather than unlikely in Ref. [28]), but also in case of predictive biomarker, measured on the subset-by-treatment interaction. Such a detection used Bayesian measures of heterogeneity in treatment effect across the biomarker strata, allowing indirect enrichment. In that sense, it differs from the previous designs, while it takes advantage of our previous work [29]. Last, while most of the previous works have considered the use of targeted treatments where many biomarkers may be of interest, we were only interested in one marker related to a specific setting of intensive care. Indeed, we considered mechanically ventilated patients, for whom how ventilation strategy may use patient profile is debated. The motivating randomised clinical trial designed in intensive care patients aimed at evaluating the benefit of the ventilation supply technique according to the responsiveness of patients to positive end-expiratory pressure (PEEP). None of the available designs appeared to directly apply to this trial.

To get further insights into our proposed approach, a simulation study was conducted to assess its operating characteristics.

2. Methods

2.1. Trial setting

Despite the best supportive care, acute respiratory distress syndrome (ARDS) still kills 35–46% of the 3 million patients affected annually. In the absence of specific treatments, providing safe and efficient mechanical ventilation (MV) is key to survival. There are still open questions regarding the best approach to MV. The PESETAS (PEep SElection Test in ArdS) trial aims to test adults admitted to the intensive care unit (ICU) with moderate or severe ARDS, a personalised approach to set positive end-expiratory pressure (PEEP). We hypothesised that i) patients with greater amounts of recruitable lung may benefit from higher PEEP levels, provided that at-tention is paid to drive pressure, ii) setting PEEP based on the results of a PEEP-responsiveness test improves survival compared to the systematic use, independent of the patient response, of either a low- (“minimal distension”, further denoted *a*) or a high-PEEP (“maximal recruitment”, denoted *b*) strategy. The main outcome was 28-day all-cause mortality after inclusion.

Therefore, the PESETAS trial aims to evaluate whether the ventilation strategy, that is, the choice of these two techniques *a or b*, depends on the patient PEEP responsiveness. Thus, based on the response to the PEEP-responsiveness test (PRT) within the first 30 min, two subsets of interest within the trial population will be defined, either $R = 1$ in case of response and $R = 0$ otherwise.

Therefore, to assess the benefit of tailoring the ventilation supply method to the patient PEEP-responsiveness, a Marker-Based Strategy Design was used. Patients were randomly allocated between three randomisation arms, namely, (i) Arm A, where all allocated patients, regardless of their PEEP responsiveness, were ventilated using a minimal distension strategy, *a*, which corresponds to the standard of care; (ii) Arm B, where all allocated patients, regardless of their PEEP responsiveness, were ventilated using a maximal recruitment strategy, b; and (iii) Arm C, where the results of a PRT were used to tailor the ventilatory strategy, i.e., patients with a negative PRT were ventilated with the same strategy a as patients randomised to comparator Arm A, and patients with a positive PRT were ventilated with the same strategy *b* as patients randomised to comparator Arm B. The study design is summarised in Fig. 1.

The main objectives of the trial were to compare the effect of each intervention (a vs. b) on the patient outcome overall and to test whether such an effect could depend on patient PEEP responsiveness. We further decided to use a Bayesian hybrid response-adaptive randomisation design, as described below.

2.2. Basic concepts for Bayesian adaptive design

Let p_j, $j = a, b$ denote the 28-day mortality rates in patients who received the ventilation supply *a* or *b*, respectively. We consider the three-arm (A, B, C) randomised clinical trial described above.

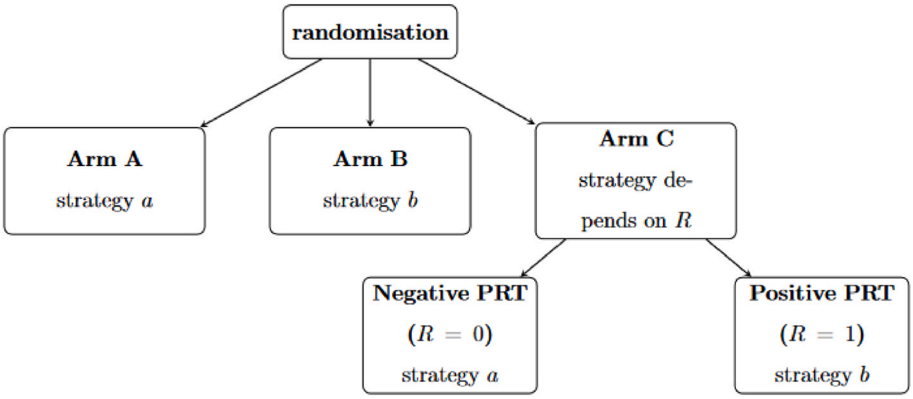

Fig. 1. Study Design of the PESETAS trial
PRT denotes the PEEP-Responsiveness Test, a denotes the minimal distension strategy, and b the maximal recruitment strategy.

Let $\mathrm{Y_{ik}}$ denote the binary outcome of the i^{th} patient in arm k = A, B, C, i = 1, 2, …, n_k, and $\mathrm{n_k}$ is the number of patients enrolled in arm k: $\mathrm{Y_{ik}} = 1$ in case of patient death at day 28, and $\mathrm{Y_{ik}} = 0$ otherwise. The total number of 28-day deaths in arm k is denoted $s_k = \sum_{i=1}^{n_k} Y_{ik}$.We consider a Bayesian framework, whe re the 28-day mortality rate in arms A, B and C, $\pi_k = (\pi_A, \pi_B \, and \, \pi_C)$, respectively, had a uniform non informative prior distribution Beta(1, 1). The use of such a flat non informative prior was motivated by several considerations. First, it allows the posterior to be dominated by the data rather than by any prior overoptimistic views regarding the experimental arms. Thus, it ensures that a critical amount of clinical information is required as a basis for deciding whether the experi-mental arm will be allocated to a large number of patients. In addition, such domination by the data allows the trial results to be used by others who have their own priors.

After observing sk deaths at day 28 from nk patients, the posterior probability of 28-day death, πk|sk, in arm k followed a beta distribution Beta($1 + s_k$, $1 + n_k - s_k$) due to the natural conjugate property of the beta family for binomial sampling.

2.3. Decision rules

Let $\tau_R = \Pr(R = r)$, denote the prevalence of subset $r \in \{0, 1\}$ (that is, of the PRT result). Let θ_R denote the treatment effect in patient subset R, measured in terms of absolute outcome difference, $\theta_R = |\pi_{a|R} - \pi_{b|R}|$, where $\pi_{t|R}$ is the probability of 28-day death in patients who received the ventilation supply t (= a, b) from subset $r \in \{0, 1\}$. As described above, the posterior distribution of that probability will be sequentially computed for each intervention (a vs. b). The resulting decision criteria and one-sided stopping rule defined at the mth stage can be summarised as follows:

- Stop the trial if there is sufficient information on the benefit of intervention b At each interim analysis, the trial can be stopped early for efficacy if the posterior probability of observing at least the minimum expected treatment difference in mortality between interventions in favour of b is higher than a predetermined threshold, computed with equation (1) described below. If this probability exceeds a certain threshold, the trial is stopped early due to demonstrated efficacy in favour of intervention b.

$$P\left(p_a - p_b > \Delta | s_{ki}, n_{ki}, k = A, B, C\right) > \nu \tag{1}$$

where Δ is the minimal expected treatment difference and ν denotes the threshold for stopping decision.

- Detection of subset (R) by treatment (a vs b) interaction for enrichment

If the previous stopping rule is not fulfilled, the interaction between the patient subset (PRT) and the intervention effect is estimated with the measures proposed by Ref. [29] computed from equations (2) and (3). If the posterior probability of these measures exceeding a predefined cutoff η is high enough and clinically relevant, then the trial population is enriched in the subset of interest.

$$P\left(\theta_{R=1}/\theta_{R=0} > \eta | \theta_{R=1} \geq \theta_{R=0}, s_{ki}, n_{ki}\right) > \varepsilon \tag{2}$$

$$P\left(\theta_{R=0}/\theta_{R=1} > \eta | \theta_{R=0} \geq \theta_{R=1}, s_{ki}, n_{ki}\right) > \varepsilon \tag{3}$$

where η is the threshold for interaction and ε denote the threshold for enrichment decision. ν and ε are set at stringent values, as recommended by Harrel [30].

2.4. Allocation probabilities

If the trial was not stopped at this interim analysis, the allocation probabilities are updated for the next cohort of patient. We propose the following Bayesian hybrid response-adaptive randomisation with eNrichment desiGn (BRING). The algorithm is summarised in the following box.

The BRING algorithm is summarised below.

A sample of N patients is to be enrolled, and randomly allocated to 3 marker-based strategy arms, A, B, and C.

A total of m interim analyses are scheduled to be performed once after N/m patients have been enrolled, with available outcome measure.

2.4.1. Step 1: Initialisation

First, patients are assigned to either arm k = A, B, C, with probabilities of $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ that is, using a 1:1:1 balance fixed-randomisation design. After a burn-in enrolment period, patient outcomes will be recorded. Then, interim analysis can be performed as described below

2.4.2. Step 2: Estimation

Then, data collected on the first cohort of enrolled patients are used to calculate the posterior mean of the 28-day death rate in each arm k, given accumulated data on n $= \sum n_k$, that is:

$$\widehat{\pi k} = E(\pi_k | s_k, n_k), k = A, B, C \tag{4}$$

2.4.3. Step 3: Optimal allocation computation

To allocate the randomisation arm k to n patients, we aimed to minimise the expected number of treatment failures (here, 28-day deaths), for a fixed asymptotic variance – which reflects the power of the test of comparing the benefit of each marker strategy against all others [24]. Thus, we used the allocation proportions $\rho^*{}_k(\pi_A, \pi_B, \pi_C)$ and k = A, B, C, as proposed by Hu and Zhu [31].

Step 1: Assign the first n = N/m patients to arms A, B, C with probabilities of 1/3 each. Record patient outcomes.

At the time of the first interim analysis, based on those n patients, and unless no event has

been observed, proceed to Step 2.

Step 2: Calculate the posterior mean of the 28-day death rate in each arm as well as posterior

probabilities of quantities of interest.

If there is evidence of any treatment effect, stop the trial; otherwise, go to Step 3.

Step 3: Calculate the optimal allocation proportion for each arm using equation (5).

Update the allocation probability function using equation (6).

Assign n further patients to the arm with the highest allocation probability.

Repeat Steps 2 and 3 until the sample size N has been reached orany stopping rule is fulfilled, at which point the trial is stopped.

The Bayesian posterior mean estimates of failure rates were plugged in the allocation ratio, as previously proposed [25,24] minimising the total number of deaths over the trial sample:

$$\rho^{*}{}_{k}(\pi_A,\pi_B,\pi_C)=\frac{\sqrt{1-\widehat{\pi_k}}}{\sqrt{1-\widehat{\pi_A}}+\sqrt{1-\widehat{\pi_B}}+\sqrt{1-\widehat{\pi_C}}} \tag{5}$$

where $\widehat{\pi A},\widehat{\pi B},\widehat{\pi C}$ are mean posterior estimates of πA, πB, πC.

We then considered that interim analyses are scheduled to be performed after every n patients have been allocated to either arm, A, B and C. Let define the optimal allocation probability function ψk proposed by Jeon and Hu [32] as the probability of allocation arm *k*:

$$\psi_k=\frac{\rho_k^{*}\left(\frac{\rho_k^{*}}{\frac{n_k}{n}}\right)^{\gamma}}{\sum_{k=A}^{C}\rho_k^{*}\left(\frac{\rho_k^{*}}{\frac{n_k}{n}}\right)^{\gamma}} \tag{6}$$

with γ = 2, tuning parameters for controlling the degree of randomness [33] and where n_k is the sample size in arm k after n enrolled patients. Although equations (4)–(6) could be computed after each patient's outcome is available, this design only plan to update the allocation probabilities at each pre-planed interim analysis.

All this processes are repeated until the sample size has been reached or any stopping rule is fulfilled, at which point the trial is stopped.

2.5. Simulation study design

Through simulations, we evaluated the operating characteristics of our trial design across a range of scenarios corresponding to various reasonable sizes of the treatment effect in each intervention and the treatment-by-subset interaction. The simulation setting aimed to mimic the PESETAS trial setting under various realistic underlying scenarios, described in Table 1. Scenario 1 refers to the null hypotheses of no treatment effect and no treatment-by-subset interaction; Scenario 2 refers to the benefit of strategy b over a in the whole population (no treatment-by-subset interaction); Scenario 3 refers to situations with a benefit of b over a, which is increased in patients with positive PRT (*r* = 1) compared to those with negative PRT (*r* = 0). While scenario 4 indicates the situation where there is a benefit of the intervention of *b* compared to *a*, which is increased in patients with a negative PRT (*r* = 0) compared to those with a positive PRT (*r* = 1); finally, in Scenarios 5 and 6, a qualitative treatment-by-subset interaction exists, with *b* beneficial in patients with positive PRT, while *a* is favoured in those with negative PRT.

The maximum sample size of 1200 patients was computed to detect a mortality difference of 5%–10% for any of the interventions, assuming a death rate at day 28 in the control arm ranging from 0.3 up to 0.45, with a power ranging from 72% up to 99%. The simulation procedure includes a burn-in period with a fixed equal allocation ratio. Then, interim analyses were conducted based on a prespecified number of recruited patients, checking the efficacy boundaries based on each intervention's posterior probabilities of death. If the trial was not stopped, the randomisation allocation ratio was updated based on the algorithm described above. Decision thresholds and stopping boundaries were defined and selected by simulations to ensure that the false-positive rate under the null was close to 5%. The minimal expected treatment difference Δ was set at 0.05, the threshold for interaction η was set at 1.05,

Table 1
Description of the simulated scenarios.

Scenarios	**28-day death probability**							
	Treatment (t)	$\pi_{t\vert R=1}$	$\pi_{t\vert R=0}$	π_A	π_B	π_C	p_a	p_b
Scenario 1	(a)	0.400	0.400	0.400	0.400	0.400	0.400	0.400
	(b)	0.400	0.400					
Scenario 2	(a)	0.400	0.400	0.400	0.350	0.375	0.400	0.350
	(b)	0.350	0.350					
Scenario 3	(a)	0.400	0.400	0.400	0.325	0.350	0.400	0.316
	(b)	0.300	0.350					
Scenario 4	(a)	0.400	0.400	0.400	0.325	0.375	0.400	0.333
	(b)	0.350	0.300					
Scenario 5	(a)	0.400	0.300	0.400	0.325	0.275	0.333	0.300
	(b)	0.250	0.400					
Scenario 6	(a)	0.600	0.300	0.400	0.325	0.275	0.400	0.300
	(b)	0.250	0.400					

a refers to the minimal distension strategy, b to the maximal recruitment strategy, *r* denotes the PEEP-responsiveness Test (PRT), with r = 1 in case of responsiveness and 0 otherwise. *πk* denotes the 28-day mortality rates in Arms *k* = A, B, C, and p_a ans p_b, denote the 28-day mortality rates in patients allocated to minimal or maximal recruitment strategy, respectively.

and ν and ε denote thresholds for stopping decisions, set at stringent values, that was, $\nu = 0.80$, and $\varepsilon = 0.75$, respectively.

The performance of the algorithm was then evaluated based on trial efficiency and patient ethics. Trial efficiency was measured by (i) the false positive rate or "type I error rate", estimated under the null scenario, by the proportion of simulation runs where treatment arm b was considered superior to arm a in terms of outcome, whichever the randomisation A, B, and C, and the underlying interaction, (ii) the true-positive rate, or "power" estimated under Scenarios 2–6, by the proportion of simulation runs with conclusion of treatment efficacy (b over a), note also that, given we only considered one-sided differences favouring the experimental arm b, there were no false positive decisions. (iii) the mean bias, normalised relatively to the true event rates and (iv) the relative mean square error (RMSE) of the mortality rates of the two interventions a and b.

Patient ethics considerations were measured by the sample size and the proportion of patients assigned to the best arm in terms of 28-day mortality rate, throughout allocation probabilities. In addition to the mean values of these measures, their standard deviations are provided to better understand the performance variability of the design. All analyses were performed using R version 4.0.1 [34] and the package "R2jags" [35]. It allowed to compute the measures of subset interactions, from equations (2) and (3).

2.6. Sensitivity analyses

We evaluated the robustness of the propos ed design to the burn-in sample size, allocation update frequency, and subset prevalence.

2.7. Comparison to alternative designs

In addition to the Bayesian response-a daptive randomisation algorithm with enrichment (BRING), a Bayesian fixed equal allocation without stopping rules (Fixed ER) design, a Bayesian response adaptive randomisation (AR) and a Bayesian equal allocation with enrichment (ERENCH) were also included in the simulation to serve as comparisons to the proposed scheme.

3. Results

The operating characteristics of the p roposed design BRING in comparison to the alternate designs are summarised in Table 2.

In Scenario 1, expectedly, the false-positive rates were similar among the three different algorithms, while in the fixed equal allocation design, the false positive rate decreased to 0.005. In the other scenarios, the fixed equal randomisation design and the Bayesian adaptive randomisation without enrichment design (AR) had lower true-positive rates than designs allowing enrichment; that is, the Bayesian adaptive randomisation and equal randomisation with enrichment designs (BRING and ERENCH), with the exception of Scenario 2. In the other scenarios, the fixed equal randomisation design and the Bayesian adaptive randomisation without enrichment design (AR) had lower true-positive rates than designs allowing enrichment; that is, the Bayesian adaptive randomisation and equal randomisation with enrichment designs (BRING and ERENCH), with the exception of Scenario 2. Concerning the fixed design, the true-positive rate was always lower than those of the remaining designs. Moreover, except in Scenarios 3, 4 and 6, the difference in the sample size among all three allocated arms was very small. This is mainly because in these scenarios, there was a qualitative interaction that was quickly detected by the treatment-by-subset interaction measure and that allowed an enrichment in the subgroup of interest and thus increased the probability of early stopping for efficiency in later analyses.

Otherwise, the bias and the mean square errors were somewhat low, indicating good performances of our design in estimating the probability of death in each treatment group and in each subset. However, bias tends to remain relatively higher from BRING than other designs in all the scenarios. However, regarding patient ethics criteria, the patient probability to be assigned to the most effective arm was not particularly different between response-adaptive and equal randomisation designs.

Table 2
Operating characteristics of the proposed design for the three-arm trial.

Scenarios	Algorithm	Number of analyses	False/True Positive rate	Sample size mean (SD)	Allocation probability Mean			Bias p_a	Bias p_b	RMSE p_a	RMSE p_b
					A	B	C				
Scenario 1	Fixed ER	1	0.005	1200 (0)	0.333	0.333	0.333	0.000	0.000	0.000	0.001
	ERENCH	4	0.055	1155 (192)	0.333	0.333	0.333	0.000	0.001	0.000	0.002
	AR	4	0.057	1154 (193)	0.335	0.332	0.333	0.002	0.002	0.002	0.001
	BRING	4	0.055	1157 (193)	0.335	0.332	0.333	0.000	0.002	0.000	0.002
Scenario 2	Fixed ER	1	0.201	1200 (0)	0.333	0.333	0.333	0.000	0.001	0.001	0.001
	ERENCH	4	0.366	957 (368)	0.333	0.333	0.333	0.018	−0.016	0.002	0.002
	AR	4	0.379	953 (368)	0.320	0.347	0.333	0.018	−0.015	0.002	0.002
	BRING	4	0.372	952 (370)	0.320	0.347	0.333	0.020	−0.016	0.002	0.002
Scenario 3	Fixed ER	1	0.644	1200 (0)	0.333	0.333	0.333	0.001	0.004	0.001	0.001
	ERENCH	4	0.772	695 (381)	0.333	0.333	0.333	0.025	−0.022	0.002	0.003
	AR	4	0.760	695 (383)	0.314	0.346	0.340	0.024	−0.018	0.003	0.003
	BRING	4	0.773	691 (382)	0.313	0.346	0.340	0.025	−0.023	0.003	0.003
Scenario 4	Fixed ER	1	0.405	1200 (0)	0.333	0.333	0.333	0.001	0.003	0.001	0.001
	ERENCH	4	0.619	821 (390)	0.333	0.333	0.333	0.023	−0.028	0.002	0.003
	AR	4	0.577	833 (393)	0.316	0.357	0.327	0.021	−0.021	0.003	0.003
	BRING	4	0.615	821 (390)	0.315	0.357	0.328	0.023	−0.028	0.003	0.003
Scenario 5	Fixed ER	1	0.068	1200 (0)	0.333	0.333	0.333	0.000	0.001	0.001	0.001
	ERENCH	4	0.526	937 (332)	0.333	0.333	0.333	0.040	−0.021	0.004	0.004
	AR	4	0.208	1056 (310)	0.312	0.324	0.364	0.012	−0.014	0.002	0.002
	BRING	4	0.517	945 (177)	0.316	0.320	0.364	0.038	−0.024	0.004	0.004
Scenario 6	Fixed ER	1	0.835	1200 (0)	0.333	0.333	0.333	0.001	0.003	0.001	0.001
	ERENCH	4	0.998	469 (197)	0.333	0.333	0.333	0.103	−0.056	0.005	0.004
	AR	4	0.860	592 (360)	0.274	0.335	0.391	0.014	−0.025	0.003	0.003
	BRING	4	0.996	471 (203)	0.270	0.337	0.393	0.089	−0.059	0.005	0.005

BRING, Bayesian response-adaptive randomisation algorithm with enrichment; Fixed ER, Bayesian fixed equal allocation; AR, Bayesian response-adaptive randomisation; ERENCH, Bayesian equal allocation with enrichment; Max sample size = 1200, burn-in period = 300, allocation update frequency = every 300 subjects, efficacy stopping boundary = 0.80, interaction stopping boundary = 0.75, simulation iteration = 10,000. Bias and RMSE correspond to the mean normalised bias and the mean square error of the 28-day mortality rates in patients allocated to intervention a or b, respectively.

A strong treatment effect was indeed required to observe a clear-cut difference in favour of the best arm, as exemplified in Scenarios 4 and 6.

Sensitivity analyses are summarised in Table 3 and Figs. 2 and 3.

Table 3 shows that the number of interim analyses, with a fixed burn-in period of 300 patients, with equal and balanced randomisation, did not have a marked impact on the false positive rate of our proposal design. In fact, the false-positive rate increased as the number of interim analyses increased, ranging from 0.047 for one interim analysis to 0.066 for 9 interim analyses.

Concerning the sensitivity to the prevalence of the subset (that is, of PEEP responders), Fig. 2 illustrates that the subset prevalence had no marked impact on the total study population in Scenarios 1 and 2. The results were consistent with the scenarios set up, as there was no interaction in these two scenarios. By contrast, in Scenarios 3, 5, and 6, the sample size decreased as the prevalence of the subset increased.

Moreover, the impact was increased in the case of a qualitative interaction, as shown in Scenarios 5 and 6. In contrast, in Scenario 4, the sample size increased as the prevalence of the subset increased. These results are consistent with the scenario set up, as there was, in this case, a quantitative interaction in favour of the subgroup with patients not responding to the PEEP test.

The reliability of our design was affected by the burn-in period length and the frequency of interim analyses. As described in Fig. 3, in Scenario 1, a shorter burn-in period yielded an increased number of interim analyses and an inflation of the false-positive rate. When the burn-in size period was set at 100 patients, the false-positive rate was approximately 15% while it decreased to 2% when it was set to 700 patients. Similarly, the same trend was observed in Scenarios 2 to 5, where the true-positive rate decreased as the burn-in period increased and the number of intermediate analyses decreased.

For Scenario 6, the results were similar regardless of the burn-in period because the difference in treatment between the two interventions in favour of intervention *b* was marked and quickly detectable by our proposed design.

4. Discussion

In precision medicine, drugs are developed to target subgroups of patients with specific biomarkers, such as clinical-pathological, molecular, or genetic variations. In recent years, there has been much work on developing designs in this setting, often focused on identifying and validating biomarker subgroups [36–40]. Once the subgroups are identified, the objective is to detect which response allows us to reach the best response to the experimental treatment.

We have proposed a Bayesian adaptive-randomisation design with enrichment to address this challenge. Using the PESETAS trial as a motivating example, we defined the process involved in constructing the design, described the adaptive-randomisation allocation, stopping and enrichment rules and studied the behaviour of the design through its operating characteristics across a range of scenarios and in comparison with alternate designs.

Using the Bayesian adaptive design, we elaborated several decision rules for detecting treatment efficacy and treatment-by-subset interaction to ensure a false-positive rate close to 5%, updated the randomisation allocation in favour of the most effective arm, and enriched the trial by enrolling more patients in the Arm that were most likely to benefit from the most effective treatment. Our design performed better than the competing ones with a higher power and decreased average sample size than the fixed and Bayesian response-adaptive randomisation designs without enrichment. In contrast, the use of RAR generally offered similar power and average sample size to the adaptive designs that employed equal randomisation, which is consistent with the results discussed by Du et al. [41] However, our design appears more ethical for patients and maximises the number of patients receiving the optimal treatment.

Table 3 and Fig. 3 illustrate that the interim analysis frequency and burn-in period had the most critical impacts on the false positive and true positive rates. Simulation studies are essential for choosing the appropriate number of interim analyses and burn-in period to evaluate this impact. Based on these results, we recommend a burn-in period of 300 and a frequency of interim analyses performed every 200–300 patients. Our results also showed that an enrichment strategy allowed the oversampling of sensitive patients and the undersampling of non-sensitive patients that benefited trial participants and thus reduced the average sample size of the study. This could be of interest in other settings where the patient condition under study is rare. This could be of interest in the setting fo histology-agnostic targeted therapeutic agents, based on specific genomic or molecular alterations, possibly resulting in very small sample sizes [42], or in the comparison of three drugs commonly used to treat various forms of isolated skin vasculitis [43].

The enrichment strategy described here achieves findings previously reported by other biomarker adaptive designs [17,44,45]. For example, the proposed enriched biomarker stratified design and the auxiliary-variable- enriched biomarker stratified design also showed that applying an enrichment strategy can result in a more cost efficient design in terms of power and reduce the average sample size [46,47]

There are however, important disadvantages and risks of adaptive designs. The limited practicability of the RAR is a main issue, with complicated logistic compared to those of standard trials. Indeed, such designs require real-time information on treatment responses to update the randomisation allocation probabilities, which can limit its practicality in certain settings. It points out the practical issue of updating the allocation probabilities to take place during the course of randomisation. This requires that database of marker, treatment, and outcome data, must be connected to the software that makes the treatment assignment. It also needs to know how to handle delayed responses when defining the allocation rate over time. This further justifies the use of interim analyses rather than continuous monitoring, as well as its use

Table 3
Operating characteristics of the proposed design for the three-arm trial.

Scenarios	Allocation Update frequency	Number of analyses	False/True Positive rate	Sample size mean (SD)	Allocation probability Mean			Bias pa	Bias pb	RMSE pa	RMSE pb
					A	B	C				
Scenario 1	900	2	0.047	1160 (186)	0.335	0.332	0.333	0.001	0.001	0.002	0.002
	450	3	0.049	1162 (177)	0.335	0.332	0.333	0.000	0.002	0.002	0.002
	300	4	0.055	1157 (193)	0.335	0.332	0.333	0.000	0.002	0.002	0.002
	225	5	0.058	1154 (191)	0.334	0.332	0.334	0.001	0.002	0.002	0.002
	180	6	0.062	1151 (198)	0.335	0.332	0.333	0.000	0.002	0.002	0.002
	150	7	0.063	1150 (200)	0.334	0.332	0.334	0.000	0.002	0.002	0.002
	100	10	0.066	1148 (202)	0.335	0.332	0.333	0.000	0.000	0.001	0.002

Max sample size = 1200, algorithm = RING, burn-in period = 300, efficacy stopping boundary = 0.80, interaction stopping boundary = 0.75, simulation iteration = 10,000. Bias and RMSE correspond to the mean normalised bias and the root mean square error of the 28-day mortality rates in patients allocated to intervention *a* or *b*, respectively.

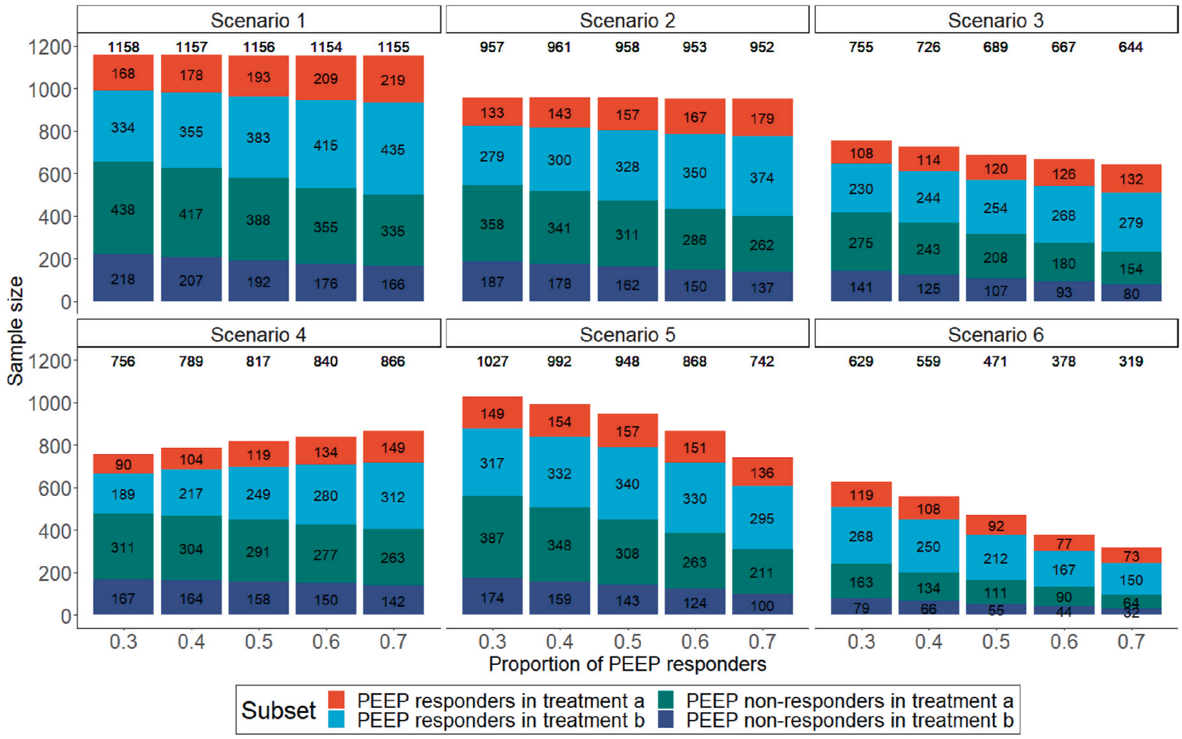

Fig. 2. Influence of the prevalence of the subset of PEEP responders defined on the PEEP-Responsiveness test (PRT) on the total enrolled sample size; Max sample size = 1200, algorithm = BRING, burn-in period = 300, allocation update frequency = every 300 subjects, prevalence of the subset of PRT = {0.3, 0.4, 0.5, 0.6, 0.7}, efficacy stopping boundary = 0.80, interaction stopping boundary = 0.75, simulation iteration = 10,000.

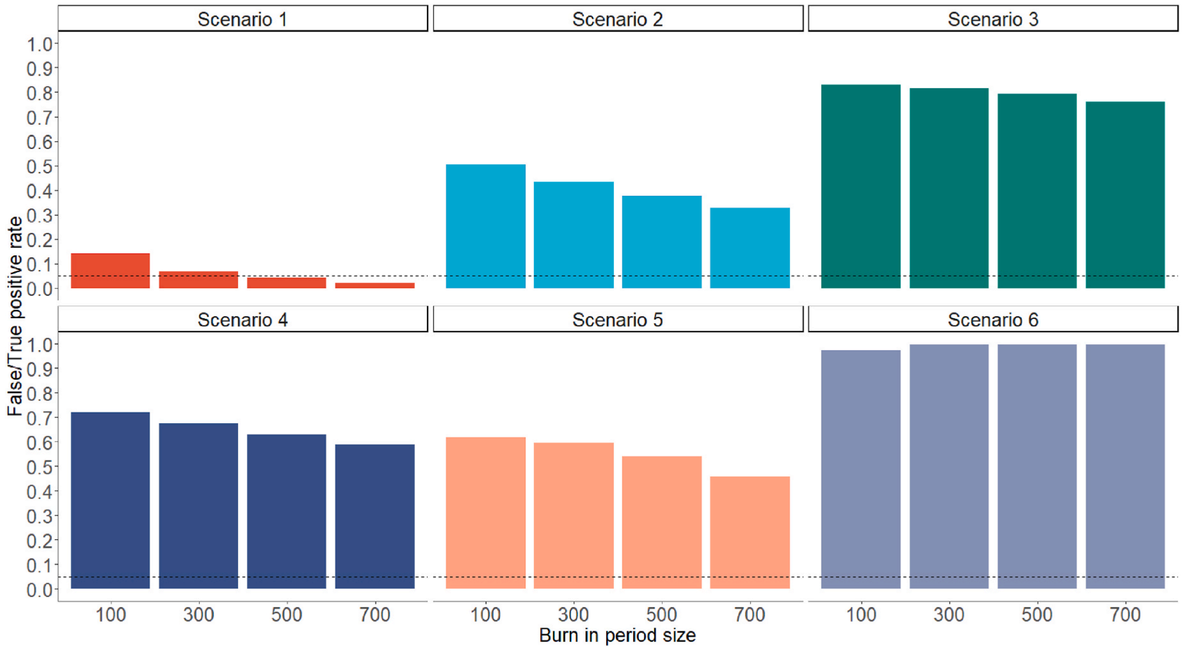

Fig. 3. Impact of the burn-in period on the proportion of trial conclusion in terms of overall treatment efficacy (a versus b)
Max sample size = 1200, algorithm = BRING, burn-in period = {100, 300, 500, 700}, allocation update frequency = every 100 subjects, efficacy stopping boundary = 0.80, interaction stopping boundary = 0.75, simulation iteration = 10,000.

in settings where long recruitment periods during which sufficient number of immediate or moderately delayed responses are accumulated, such as in cancer. One other concern is information leakage, with modifications that occur during the trial may convey information outside the sphere of confidentiality of the Data and Safety Monitoring Board (DSMB) and affect the types of patients who are accrued to the trial. Although of reported interest in multiarmed trials [48], it achieves a loss of power in paired comparisons.

Although our design provides some advantages, some limitations cannot be ignored. In fact, our design was designed for the PESETAS trial. According to the context of the study, it is essential to elaborate appropriate decision rules and stopping boundaries to control the false positive rate. The time needed to assess the patient biomarkers should be short enough, as patient treatment allocation depends on the determination of the marker profile. In the motivating example, a PEEP-responsiveness test is performed within 30 min of admission to the intensive care unit, but this can take much longer depending on the biomarker. In addition, the time for outcome assessment must be relatively short so that the decision based on updated data can provide appropriate guidance for subsequent treatment assignments. If a trial has a fast accrual rate, many patients may have been enrolled in the trial before the outcome data became available to provide helpful information for the adaptive randomisation. Therefore, quick and easily computed end points and slow to moderate accrual rates (relative to the outcome assessment time) are most suitable for designs with response-adaptive randomisation. Last, we only considered one marker, though the design could be extended to the use of more than one marker. Note however, that this design is mostly dedicated to a situation where one wish to assess the interest of segregating treatment across a known marker.

In conclusion, adaptive designs are exciting and promising for answering questions of clinical interest as quickly as possible, but they need to ensure that their conclusions are controlled for decision errors. We have proposed a Bayesian hybrid adaptive randomisation with an enrichment design to be applied to the search for the benefit of interventions that may differ between the patient subgroups. Such an approach appears promising in the large context of stratified or precision medicine.

Funding

None.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2023.101141.

References

[1] D.C. Angus, S. Berry, R.J. Lewis, F. Al-Beidh, Y. Arabi, W. van Bentum-Puijk, Z. Bhimani, M. Bonten, K. Broglio, F. Brunkhorst, et al., The remap-cap (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. Rationale and design, Ann. Am. Thoracic Soc. 17 (2020) 879–891.

[2] T.R.C. Group, Dexamethasone in hospitalized patients with covid-19, N. Engl. J. Med. 384 (2021) 693–704, https://doi.org/10.1056/NEJMoa2021436.

[3] B.L. Houston, P.R. Lawler, E.C. Goligher, M.E. Farkouh, C. Bradbury, M. Carrier, V. Dzavik, D.A. Fergusson, R.A. Fowler, J.P. Galanaud, et al., Anti-thrombotic therapy to ameliorate complications of covid-19 (attacc): study design and methodology for an international, adaptive bayesian randomized controlled trial, Clin. Trials 17 (2020) 491–500.

[4] J.T. Connor, B.R. Luce, K.R. Broglio, K.J. Ishak, C.D. Mullins, D.J. Vanness, R. Fleurence, E. Saunders, B.R. Davis, Do bayesian adaptive trials offer advantages for comparative effectiveness research? protocol for the re-adapt study, Clin. Trials 10 (2013) 807–827.

[5] L. Trippa, E.Q. Lee, P.Y. Wen, T.T. Batchelor, T. Cloughesy, G. Parmigiani, B. M. Alexander, Bayesian adaptive randomized trial design for patients with recurrent glioblastoma, J. Clin. Oncol. 30 (2012) 3258.

[6] J.M. Wason, L. Trippa, A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials, Stat. Med. 33 (2014) 2206–2221.

[7] E.L. Korn, B. Freidlin, Outcome-adaptive randomization: is it useful? J. Clin. Oncol. 29 (2011) 771.

[8] P. Thall, P. Fox, J. Wathen, Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials, Ann. Oncol. 26 (2015) 1621–1628.

[9] Y. Jiang, W. Zhao, V. Durkalski-Mauldin, Impact of adapta- tion algorithm, timing, and stopping boundaries on the performance of bayesian response adaptive randomization in confirmative trials with a binary endpoint, Contemp. Clin. Trials 62 (2017) 114–120.

[10] E.G. Ryan, S.E. Lamb, E. Williamson, S. Gates, Bayesian adaptive designs for multi-arm trials: an orthopaedic case study, Trials 21 (2020) 1–16.

[11] K. Viele, K. Broglio, A. McGlothlin, B.R. Saville, Comparison of methods for control allocation in multiple arm studies using response adaptive randomization, Clin. Trials 17 (2020) 52–60.

[12] J.K. Wathen, P.F. Thall, A simulation study of outcome adaptive randomization in multi-arm clinical trials, Clin. Trials 14 (2017) 432–440.

[13] Y. Park, Personalized risk-based screening design for comparative two-arm group sequential clinical trials, J. Personalized Med. 12 (2022) 448.

[14] Y. Park, S. Liu, P.F. Thall, Y. Yuan, Bayesian group sequential enrichment designs based on adaptive regression of response and survival time on baseline biomarkers, Biometrics (2021) 1–12, https://doi.org/10.1111/biom.13421.

[15] N. Simon, R. Simon, Adaptive enrichment designs for clinical trials, Biostatistics 14 (2013) 613–625.

[16] N. Ballarini, T. Burnett, T. Jaki, C. Jennison, F. König, M. Posch, Optimizing subgroup selection in two-stage adaptive enrichment and umbrella designs, Stat. Med. 40 (2021) 2939–2956.

[17] F. Xia, S.L. George, J. Ning, L. Li, X. Huang, A signature enrichment design with bayesian adaptive randomization, J. Appl. Stat. 48 (2021) 1091–1110, https://doi.org/10.1080/02664763.2020.1757048.

[18] H. Janes, M.D. Brown, M.S. Pepe, Designing a study to evaluate the benefit of a biomarker for selecting patient treatment, Stat. Med. 34 (2015) 3503–3515, https://doi.org/10.1002/sim.6564.

[19] D. Sargent, B. Conley, C. Allegra, L. Collette, Clinical trial designs for predictive marker validation in cancer treatment trials, J. Clin. Oncol. 23 (2005) 2020–2027.

[20] N. Best, R. Price, I. Pouliquen, O. Keene, Assessing efficacy in important subgroups in confirmatory trials: an example using bayesian dynamic borrowing, Pharmaceut. Stat. 20 (2021) 551–562.

[21] Y. Zang, J. Jack Lee, Y. Yuan, Two-stage marker-stratified clinical trial design in the presence of biomarker misclassification, J. Roy. Stat. Soc. C Appl. Stat. 65 (2016) 585–601.

[22] Y. Zang, S. Liu, Y. Yuan, Optimal marker-strategy clinical trial design to detect predictive markers for targeted therapy, Biostatistics 17 (2016) 549–560.

[23] J. Wason, R. Dunn, C. Stein, N. Stallard, Adaptive designs for clinical trials assessing biomarker-guided treatment strategies, Br. J. Cancer 110 (2014) 1950–1957.

[24] W.F. Rosenberger, N. Stallard, A. Ivanova, C.N. Harper, M.L. Ricks, Optimal adaptive designs for binary response trials, Biometrics 57 (2001) 909–913.

[25] M a. Moatti, S. Chevret, S. Zohar, W.F. Rosenberger, A bayesian hybrid adaptive randomisation design for clinical trials with survival outcomes, Methods Inf. Med. 55 (2016) 4–13.

[26] X. Gu, N. Chen, C. Wei, S. Liu, V.A. Papadimitrakopoulou, R.S. Herbst, J.J. Lee, Bayesian two-stage biomarker-based adaptive design for targeted therapy development, Stat Biosci 8 (2016) 99–128.

[27] P. Thall, J. Wathen, Practical bayesian adaptive randomisation in clinical trials, Eur. J. Cancer 43 (2007) 859–866.

[28] X. Zhou, S. Liu, E.S. Kim, R.S. Herbst, J.J. Lee, Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine, Clin. Trials 5 (2008) 181–193.

[29] V. Vinnat, S. Chevret, Enrichment bayesian design for randomized clinical trials using categorical biomarkers and a binary outcome, BMC Med. Res. Methodol. 22 (2022) 1–15.

[30] F. Harrell, C. Lindsell, Statistical design and analysis plan for sequential parallel-group rct for covid-19. http://hbiostat.org/proj/covid19/bayesplan.html.

[31] F. Hu, L.X. Zhang, Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials, Ann. Stat. 32 (2004) 268–301.

[32] Y. Jeon, F. Hu, Optimal adaptive designs for binary response trials with three treatments, Stat. Biopharm. Res. 2 (2010) 310–318.

[33] F. Hu, W.F. Rosenberger, Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons, J. Am. Stat. Assoc. 98 (2003) 671–678.

[34] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020. URL: https://www.R-project.org/.

[35] Y.S. Su, M. Yajima, R2jags: using r to run 'jags, URL, https://CRAN.R-project.org/package=R2jags. r package version 0.5-7, 2015.

[36] J.C. Eickhoff, K. Kim, J. Beach, J.M. Kolesar, J.R. Gee, A bayesian adaptive design with biomarkers for targeted therapies, Clin. Trials 7 (2010) 546–556, https://doi.org/10.1177/1740774510372657.

[37] C. Hu, J.J. Dignam, Biomarker-driven oncology clinical trials: key design elements, types, features, and practical considerations, JCO Precision Oncol. 3 (2019) 1–12, https://doi.org/10.1200/PO.19.00086.

[38] J.J. Lee, X. Gu, S. Liu, Bayesian adaptive randomization designs for targeted agent development, Clin. Trials 7 (2010) 584–596.

[39] S.J. Mandrekar, D.J. Sargent, Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges, J. Clin. Oncol. 27 (2009) 4027–4034, https://doi.org/10.1200/JCO.2009.22.3701.

[40] L. Trippa, B.M. Alexander, Bayesian baskets: a novel design for biomarker-based clinical trials, J. Clin. Oncol. 35 (2017), https://doi.org/10.1200/JCO.2016.68.2864. JCO.2016.68.2864.

[41] Y. Du, X. Wang, J.J. Lee, Simulation study for evaluating the performance of response-adaptive randomization, Contemp. Clin. Trials 40 (2015) 15–25.

[42] M. A, F. Rm, C. Av, Histology-agnostic drugs: a paradigm shift-a narrative review, Adv. Ther. (2022), https://doi.org/10.1007/s12325-022-02362-4.

[43] R. Micheletti, C. Pagnoux, R. Tamura, P. Grayson, C. McAlear, R. Borchin, J. Krischer, P. Merkel, Vasculitis clinical research consortium. protocol for a randomized multicenter study for isolated skin vasculitis (aramis) comparing the efficacy of three drugs: azathioprine, colchicine, and dapsone, Trials 21 (2020) 362, 110.1186/s13063-020-04285-3.

[44] B. Freidlin, L.M. McShane, E.L. Korn, Randomized clinical trials with biomarkers: design issues, J. Natl. Cancer Inst. 102 (2010) 152–160.

[45] T. Ondra, S. Jobjörnsson, R.A. Beckman, C.F. Burman, F. König, N. Stallard, M. Posch, Optimized adaptive enrichment designs, Stat. Methods Med. Res. 28 (2019), 20962111, https://doi.org/10.1177/0962280217747312.pMID:29254436.

[46] O. Wang, J. Zhou, T. Wang, S.L. George, On enrichment strategies for biomarker stratified clinical trials, J. Biopharm. Stat. 28 (2018) 292–308.

[47] T. Wang, X. Wang, H. Zhou, J. Cai, S.L. George, Auxiliary variable–enriched biomarker-stratified design, Stat. Med. 37 (2018) 4610–4635.

[48] D.A. Berry, Adaptive clinical trials: the promise and the caution, J. Clin. Oncol. 29 (2011) 606–609, https://doi.org/10.1200/JCO.2010.32.2685.