Cabarrou *et al.*
BMC Medical Research Methodology (2022) 22:278
https://doi.org/10.1186/s12874-022-01748-w

BMC Medical Research Methodology

RESEARCH

Open Access

A stratified adaptive two-stage design with co-primary endpoints for phase II clinical oncology trials

Bastien Cabarrou[1], Eve Leconte[2], Patrick Sfumato[3], Jean-Marie Boher[3,4] and Thomas Filleron[1*]

Abstract

Background: Given the inherent challenges of conducting randomized phase III trials in older cancer patients, single-arm phase II trials which assess the feasibility of a treatment that has already been shown to be effective in a younger population may provide a compelling alternative. Such an approach would need to evaluate treatment feasibility based on a composite endpoint that combines multiple clinical dimensions and to stratify older patients as fit or frail to account for the heterogeneity of the study population to recommend an appropriate treatment approach. In this context, stratified adaptive two-stage designs for binary or composite endpoints, initially developed for biomarker studies, allow to include two subgroups whilst maintaining competitive statistical performances. In practice, heterogeneity may indeed affect more than one dimension and incorporating co-primary endpoints, which independently assess each individual clinical dimension, would therefore appear quite pertinent. The current paper presents a novel phase II design for co-primary endpoints which takes into account the heterogeneity of a population.

Methods: We developed a stratified adaptive Bryant & Day design based on the Jones et al. and Parashar et al. algorithm. This two-stage design allows to jointly assess two dimensions (e.g. activity and toxicity) in two different subgroups. The operating characteristics of this new design were evaluated using examples and simulation comparisons with the Bryant & Day design in the context where the study population is stratified according to a pre-defined criterion.

Results: Simulation results demonstrated that the new design minimized the expected and maximum sample sizes as compared to parallel Bryant & Day designs (one in each subgroup), whilst controlling type I error rates and maintaining a competitive statistical power as well as a high probability of detecting heterogeneity.

Conclusions: In a heterogeneous population, this two-stage stratified adaptive phase II design provides a useful alternative to classical one and allows to identify a subgroup of interest without dramatically increasing sample size. As heterogeneity is not limited to older populations, this new design may also be relevant to other study populations such as children or adolescents and young adults or the development of targeted therapies based on a biomarker.

Keywords: Phase II clinical oncology trials, Heterogeneity, Adaptive stratified design, Co-primary endpoints

Background

The main objective of a phase II oncology trial is to assess the anti-tumoral activity of an experimental treatment. If promising results are obtained, the phase II is followed by a phase III trial to evaluate the effectiveness of an experimental treatment compared to a standard

*Correspondence: filleron.thomas@iuct-oncopole.fr

[1] Biostatistics & Health Data Science Unit, Institut Claudius Regaud - IUCT-O, 1 avenue Irène Joliot-Curie, 31059 Cedex 9 Toulouse, France
Full list of author information is available at the end of the article

treatment. Older patients are vastly underrepresented in phase III clinical trials and the problem of recruiting older people has been largely documented in the literature. The most common barriers cited were: stringent eligibility criteria, oncologists concerns for toxicity, patients and family refusal [1]. Given the challenges of conducting randomized phase III trials in older patients, several authors have previously suggested conducting single-arm phase II trials to assess the feasibility of a treatment that has been shown to be effective in a younger population [2, 3]. Indeed, perhaps more importantly than in any other population, cancer care should not compromise quality of life or autonomy [2, 3]. Treatment feasibility can be evaluated with a composite endpoint combining multiple clinical dimensions (e.g. activity, toxicity, quality of life, etc.). The treatment may be considered feasible if it fulfills some or all components of the composite endpoint. Another conundrum is to take into account the heterogeneity of this population and stratifying older patients as fit or frail is crucial to recommend an appropriate treatment approach [4]. Classical phase II designs for binary or composite endpoints [5–7] do not deal with this heterogeneity and can lead to erroneous conclusions in an unselected population, while a specific subgroup of less frail (or less fit) patients might benefit (or not) from the new therapeutic. Stratified adaptive two-stage designs for binary or composite endpoints, which allow the inclusion of two subgroups and identify one of interest at the end of the first or the second stage, have recently been proposed [8–10]. Initially developed for biomarker studies, these types of approaches can also be applied to geriatric clinical oncology trials and allow to minimize the sample size whilst maintaining a competitive statistical performance that is comparable to classical approaches [11]. These stratified adaptive designs have been developed for binary or composite endpoints and they take into account the heterogeneity of a population when considering a single or combined clinical dimensions where each of them theoretically carries the same clinical importance. However, depending on the clinical context, the impact on autonomy or quality of life may take precedence over anti-tumoral activity in treatment decision-making. Moreover, interpretation may be difficult if there are divergent results for each clinical dimension separate. Thus, the use of co-primary endpoints that assess each clinical dimension independently appears more relevant in this light [12]. Several designs that deal with these types of endpoints have been proposed, but the most widely used is the one developed by Bryant and Day [13]. To the best of our knowledge the current literature does not include any reports of phase II designs for co-primary endpoints that account for heterogeneity. The current paper therefore details a stratified adaptive Bryant & Day (SABD) design based on the algorithm developed by Jones et al. [8] and Parashar et al. [10] (Methods section). The operating characteristics of the novel design are then evaluated using examples and simulation comparisons with the Bryant & Day (BD) design (Results section).

Methods

Bryant & Day (BD) design

The BD design can be considered as a two-stage Simon optimal design [6] which considers two dimensions as co-primary endpoints, namely activity and toxicity. The BD design, where X_{R1} and X_{T1} represent the number of responses and non-toxicities observed at the end of the first stage and X_R and X_T the total number of responses and non-toxicities observed at the end of the second stage, is shown in Fig. 1.

After the inclusion of N_1 patients, the study will be stopped for futility if an insufficient number of responses or non-toxicities are observed (i.e. $X_{R1} < k_{R1}$ or $X_{T1} < k_{T1}$). The experimental treatment will be considered as promising (i.e. «go-decision») if a sufficient number of responses and non-toxicities are observed in the interim (i.e. $X_{R1} \geq k_{R1}$ and $X_{T1} \geq k_{T1}$) and in the final (i.e. $X_R \geq k_R$ and $X_T \geq k_T$) analysis.

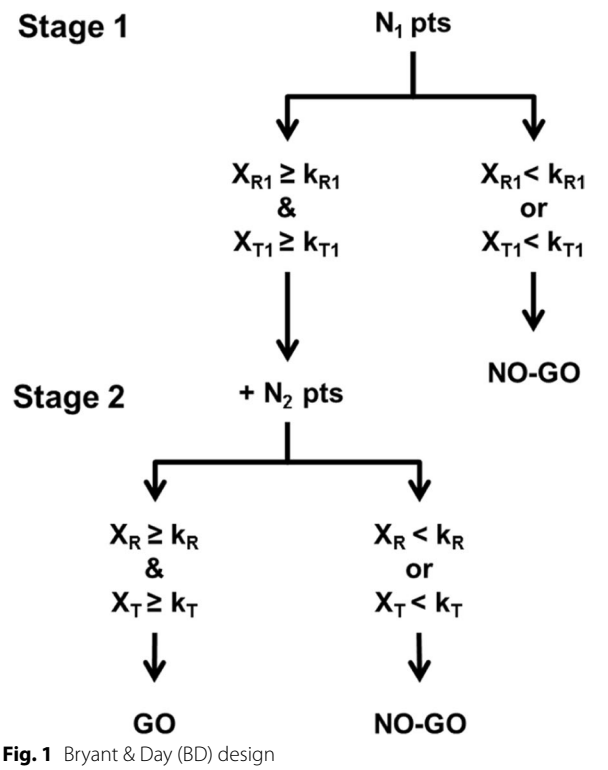

Fig. 1 Bryant & Day (BD) design

Unacceptable and acceptable rates for each dimension are denoted as follows, with p_R and p_T respectively representing the response rate and the non-toxicity rate:

- p_{R0}: unacceptable response rate
- p_{R1}: acceptable response rate
- p_{T0}: unacceptable non-toxicity rate
- p_{T1}: acceptable non-toxicity rate

Given the two-dimensional nature of the endpoint, the null and alternative hypotheses are areas and defined by H_0: $\{p_R \leq p_{R0}$ or $p_T \leq p_{T0}\}$ and H_1: $\{p_R > p_{R0}$ and $p_T > p_{T0}\}$, respectively. Four particular hypotheses corresponding to four possible states are considered:

- H_{00}: $\{p_R = p_{R0}$ and $p_T = p_{T0}\}$
- H_{01}: $\{p_R = p_{R0}$ and $p_T = p_{T1}\}$
- H_{10}: $\{p_R = p_{R1}$ and $p_T = p_{T0}\}$
- H_{11}: $\{p_R = p_{R1}$ and $p_T = p_{T1}\}$

There are four associated error rates:

- α: is the probability of considering the treatment as promising in the case where true response and non-toxicity rates are considered as unacceptable (i.e. under H_{00}),
- α_R: is the probability of considering the treatment as promising in the case where true response and non-toxicity rates are considered as unacceptable and acceptable, respectively (i.e. under H_{01}),
- α_T: is the probability of considering the treatment as promising in the case where true response and non-toxicity rates are considered as acceptable and unacceptable, respectively (i.e. under H_{10}),
- β: is the probability of considering the treatment as insufficiently promising in the case where true response and non-toxicity rates are considered as acceptable (i.e. under H_{11}).

Sample sizes of stage 1 and 2 (N_1 and N_2) and stopping boundaries (k_{R1}, k_{T1}, k_R and k_T) are determined from the specified values for p_{R0}, p_{T0}, p_{R1} and p_{T1} and the type I (α_R and α_T) and type II (β) error rates. The optimal design is defined as the one that minimizes the maximum expected sample size (ESS) under H_{10} or H_{01} (i.e. max{ESS under H_{10}, ESS under H_{01}}) whilst controlling for α_R, α_T and β.

Stratified Adaptive Bryant & Day (SABD) design

To take into account population heterogeneity, we developed a SABD design based on the Jones et al. [8] and Parashar et al. [10] algorithm. As compared to these designs that have been developed for binary or composite endpoints, this novel two-stage design allows to jointly assess two clinical dimensions (e.g. activity and toxicity) through co-primary endpoints in two different subgroups and to identify one of interest at the end of the first or the second stage. In the context of a geriatric clinical oncology trial for example, this allows patients to be stratified, according to a geriatric criterion, into frail and fit subgroups. To simplify the notation, these two subgroups will be defined as negative (‹‹-››) and positive (‹‹+››) subgroups respectively. The two-stage algorithm proposed by Jones et al. and Parashar et al., presented in Fig. 2, relies on an assumption of hierarchy between the subgroups as the true response and non-toxicity rates will always be equal or higher in the positive subgroup than in the negative subgroup. This implies that, according to the preliminary results observed at the end of the first stage, enrollment continues in an unselected population if promising results are observed in the negative subgroup, or in the positive subgroup (i.e. enrichment) if promising results are observed in this subgroup only.

Based on this algorithm and adapted from the BD design to consider two co-primary endpoints, we proposed the SABD design presented in Fig. 3.

The study begins with the inclusion of N_1^- and N_1^+ patients in the negative and positive subgroup, respectively. According to the results observed at the end of the first stage, enrollment will be stopped for futility if an insufficient number of responses or non-toxicities are observed in the two subgroups (i.e. ($X_{R1}^- < k_{R1}^-$ or $X_{T1}^- < k_{T1}^-$) and ($X_{R1}^+ < k_{R1}^+$ or $X_{T1}^+ < k_{T1}^+$)). Otherwise, enrollment will continue in the unselected population if a sufficient number of responses and non-toxicities are observed in the negative subgroup (i.e. $X_{R1}^- \geq k_{R1}^-$ and $X_{T1}^- \geq k_{T1}^-$). If a sufficient number of responses and non-toxicities are only observed in the positive subgroup (i.e. ($X_{R1}^- < k_{R1}^-$ or $X_{T1}^- < k_{T1}^-$) and ($X_{R1}^+ \geq k_{R1}^+$ and $X_{T1}^+ \geq k_{T1}^+$)) then enrollment will continue in this subgroup only. At the end of the second stage, the experimental treatment may be considered as promising (i.e. ‹‹go-decision››) in the two subgroups (i.e. S1) or in the positive subgroup only (i.e. S2 or S3).

Hypotheses

Similarly to the BD design, our SABD design assumes that the co-primary endpoints are independent in the two subgroups. If p_R^-, p_R^+, p_T^- and p_T^+ respectively correspond to the true response and non-toxicity rates in the negative and positive subgroups, the unacceptable and acceptable rates for each endpoint and subgroup may then be expressed as follows:

- p_{R0}^- and p_{R0}^+: unacceptable response rates in the negative and positive subgroups,

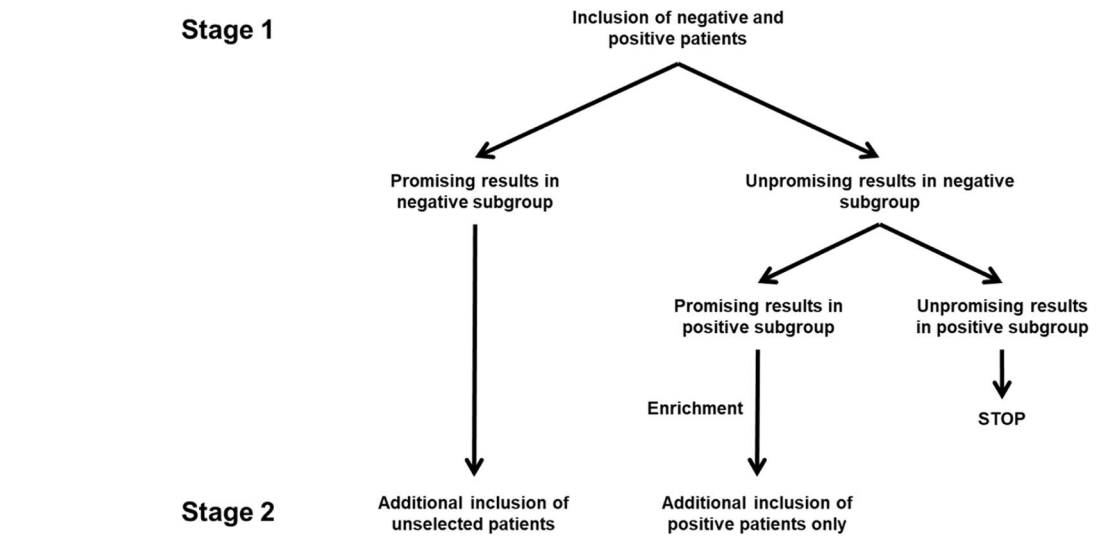

Fig. 2 Jones et al. and Parashar et al. algorithm

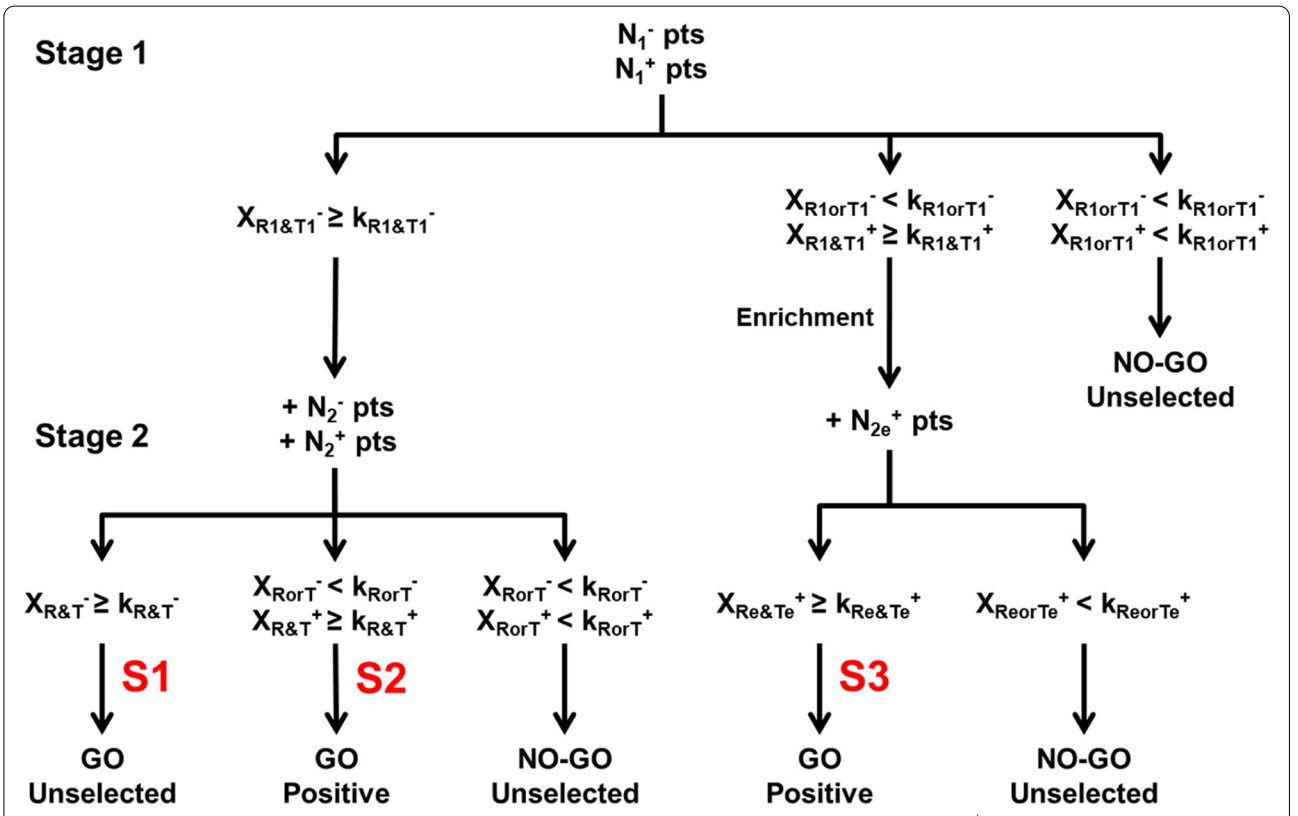

Fig. 3 Stratified Adaptive Bryant & Day (SABD) design. *e*: «enrichment». +: «inclusion of additional». N_1^- and N_1^+: number of patients to be included at the first stage. N_2^-, N_2^+ and N_{2e}^+: number of patients to be included at the second stage. $N^- = N_1^- + N_2^-$, $N^+ = N_1^+ + N_2^+$ and $N_e^+ = N_1^+ + N_{2e}^+$: total number of patients to be included. k_{R1}^-, k_{T1}^-, k_{R1}^+ and k_{T1}^+: first stage stopping boundaries. k_R^-, k_T^-, k_R^+, k_T^+, k_{Re}^+ and k_{Te}^+: second stage stopping boundaries. X_{R1}^-, X_{T1}^-, X_{R1}^+ and X_{T1}^+: number of responses and non-toxicities observed during the first stage. X_{R2}^-, X_{T2}^- X_{R2}^+, X_{T2}^+, X_{R2e}^+ and X_{T2e}^+: number of responses and non-toxicities observed during the second stage. $X_R^- = X_{R1}^- + X_{R2}^-$, $X_T^- = X_{T1}^- + X_{T2}^-$, $X_R^+ = X_{R1}^+ + X_{R2}^+$, $X_T^+ = X_{T1}^+ + X_{T2}^+$, $X_{Re}^+ = X_{R1}^+ + X_{R2e}^+$ and $X_{Te}^+ = X_{T1}^+ + X_{T2e}^+$: total number of responses and non-toxicities observed

- p_{R1}^{-} and p_{R1}^{+}: acceptable response rates in the negative and positive subgroups,
- p_{T0}^{-} and p_{T0}^{+}: unacceptable non-toxicity rates in the negative and positive subgroups,
- p_{T1}^{-} and p_{T1}^{+}: acceptable non-toxicity rates in the negative and positive subgroups.

It is assumed that the null hypothesis is identical between subgroups for the co-primary endpoints (i.e. $p_{R0}^{-}=p_{R0}^{+}$ and $p_{T0}^{-}=p_{T0}^{+}$). Null and alternative hypotheses in both subgroups are therefore defined as follows:

- $H_0^{-(+)}$: $\{p_R^{-(+)} \leq p_{R0}^{-(+)}$ or $p_T^{-(+)} \leq p_{T0}^{-(+)}\}$
- $H_1^{-(+)}$: $\{p_R^{-(+)} > p_{R0}^{-(+)}$ and $p_T^{-(+)} > p_{T0}^{-(+)}\}$

Four particular hypotheses in both subgroups are considered:

- $H_{00}^{-(+)}$: $\{p_R^{-(+)} = p_{R0}^{-(+)}$ and $p_T^{-(+)} = p_{T0}^{-(+)}\}$
- $H_{10}^{-(+)}$: $\{p_R^{-(+)} = p_{R1}^{-(+)}$ and $p_T^{-(+)} = p_{T0}^{-(+)}\}$
- $H_{01}^{-(+)}$: $\{p_R^{-(+)} = p_{R0}^{-(+)}$ and $p_T^{-(+)} = p_{T1}^{-(+)}\}$
- $H_{11}^{-(+)}$: $\{p_R^{-(+)} = p_{R1}^{-(+)}$ and $p_T^{-(+)} = p_{T1}^{-(+)}\}$

Probability of rejecting null hypotheses

There are three possible scenarios where the experimental treatment is considered as promising in the unselected population or only in the positive subgroup (i.e. ‹‹go-decision››). These scenarios correspond to S1, S2 and S3 presented in Fig. 3.

The probability of considering the experimental treatment as promising in the unselected population (i.e. reject H_0^{-} and H_0^{+}) according to S1 is defined as:

$$P\left(S1|p_R^-,p_T^-\right) = P(X_{R1}^- + X_{R2}^- \geq k_R^- \, and \, X_{R1}^- \geq k_{R1}^-) \times P(X_{T1}^- + X_{T2}^- \geq k_T^- \, and \, X_{T1}^- \geq k_{T1}^-)$$

According to the hypothesis of hierarchy between subgroups, the probability of considering the experimental treatment as promising depends on the true response and non-toxicity rate in the negative subgroup only.

The probability of considering the experimental treatment as promising in the positive subgroup only (i.e. reject H_0^{+}) according to S2 is defined as:

$$P\left(S2\middle|p_R^-,\ p_T^-,\ p_R^+,\ p_T^+\right) = P(X_R^+ \geq k_R^+) \times P(X_T^+ \geq k_T^+) \times \left[P\left((X_{R1}^- + X_{R2}^- < k_R^- \, and \, X_{R1}^- \geq k_{R1}^- \, and \, X_{T1}^- \geq k_{T1}^-\right) \, or \, \left(X_{T1}^- + X_{T2}^- < k_T^- \, and \, X_{R1}^- \geq k_{R1}^- \, and \, X_{T1}^- \geq k_{T1}^-\right))\right]$$

The probability of considering the experimental treatment as promising in the positive subgroup only (i.e. reject H_0^{+}) according to S3 is defined as:

$$P\left(S3|p_R^-,p_T^-,p_R^+,p_T^+\right) = P\left(X_{R1}^+ + X_{R2e}^+ \geq k_{Re}^+ \, and \, X_{R1}^+ \geq k_{R1}^+\right) \times P\left(X_{T1}^+ + X_{T2e}^+ \geq k_{Te}^+ \, and \, X_{T1}^+ \geq k_{T1}^+\right) \times P(X_{R1}^- < k_{R1}^- \, or \, X_{T1}^- < k_{T1}^-)$$

To compute probabilities of rejecting null hypotheses, it is assumed that the number of responses and non-toxicities follow a binomial distribution, $B(N,p)$, with parameters N and p defined in Table 1.

Type I errors

Similarly to the BD design, three type I errors may be considered for the SABD design. The overall type I error rate α corresponds to the probability of considering the treatment as promising in the unselected population or in the positive subgroup in the case where true response and non-toxicity rates are considered as unacceptable in the two subgroups (i.e. under H_{00}^{-} and H_{00}^{+}). It is defined as:

$$\alpha = P\left(S1|p_{R0}^-,p_{T0}^-\right) + P\left(S2|p_{R0}^-,p_{T0}^-,p_{R0}^+,p_{T0}^+\right) + P\left(S3|p_{R0}^-,p_{T0}^-,p_{R0}^+,p_{T0}^+\right)$$

Type I error rate α_R corresponds to the probability of considering the treatment as promising in the unselected population or in the positive subgroup in the case where true response and non-toxicity rates are considered as unacceptable and acceptable, respectively, in the two subgroups (i.e. under H_{01}^{-} and H_{01}^{+}). It is defined as:

$$\alpha_R = P\left(S1|p_{R0}^-,p_{T1}^-\right) + P\left(S2|p_{R0}^-,p_{T1}^-,p_{R0}^+,p_{T1}^+\right) + P\left(S3|p_{R0}^-,p_{T1}^-,p_{R0}^+,p_{T1}^+\right)$$

Type I error rate α_T corresponds to the probability of considering the treatment as promising in the unselected population or in the positive subgroup in the case where true response and non-toxicity rates are considered as acceptable and unacceptable, respectively, in the two subgroups (i.e. under H_{10}^{-} and H_{10}^{+}). It is defined as:

$$\alpha_T = P\left(S1|p_{R1}^-,p_{T0}^-\right) + P\left(S2|p_{R1}^-,p_{T0}^-,p_{R1}^+,p_{T0}^+\right) + P\left(S3|p_{R1}^-,p_{T0}^-,p_{R1}^+,p_{T0}^+\right)$$

Statistical power

The probability of considering the treatment as promising in the unselected population in the case where

Table 1 Parameters of binomial distributions

	Response	Non-toxicity
Negative	$X_{R1}^{-}\sim B(N_1^{-},p_R^{-})$	$X_{T1}^{-}\sim B(N_1^{-},p_T^{-})$
	$X_{R2}^{-}\sim B(N_2^{-},p_R^{-})$	$X_{T2}^{-}\sim B(N_2^{-},p_T^{-})$
Positive	$X_{R1}^{+}\sim B(N_1^{+},p_R^{+})$	$X_{T1}^{+}\sim B(N_1^{+},p_T^{+})$
	$X_R^{+}\sim B(N^{+},p_R^{+})$	$X_T^{+}\sim B(N^{+},p_T^{+})$
	$X_{R2e}^{+}\sim B(N_{2e}^{+},p_R^{+})$	$X_{T2e}^{+}\sim B(N_{2e}^{+},p_T^{+})$

$\alpha_T/2$ to adjust for multiplicity). To delineate the parameter search space, the maximum sample size is set at $2\times N^{-}$.

Probability of Early termination (PET)

The study will stop for futility if an insufficient number of responses or non-toxicities are observed in both groups in the interim analysis. The PET is determined as follows:

$$PET\left(p_R^{-},p_T^{-},p_R^{+},p_T^{+}\right)=P\left(X_{R1}^{-}<k_{R1}^{-}\ or\ X_{T1}^{-}<k_{T1}^{-}\right)\times P\left(X_{R1}^{+}<k_{R1}^{+}\ or\ X_{T1}^{+}<k_{T1}^{+}\right)$$

true response and non-toxicity rates are considered as acceptable in the negative subgroup, and therefore in the positive subgroup by the assumption of hierarchy, corresponds to $P(S1|p_{R1}^{-},p_{T1}^{-})$. The probability of considering the treatment as promising in the positive subgroup in the case where true response and non-toxicity rates are considered as acceptable in the positive subgroup only corresponds to $P(S2|p_{R0}^{-},p_{T0}^{-},p_{R1}^{+},p_{T1}^{+})+P(S3|p_{R0}^{-},p_{T0}^{-},p_{R1}^{+},p_{T1}^{+})$. As proposed by Parashar et al. [10], the overall power is defined by the minimum of these two probabilities:

$$power=1-\beta=\min\{P\left(S1|p_{R1}^{-},p_{T1}^{-}\right),P\left(S2|p_{R0}^{-},p_{T0}^{-},p_{R1}^{+},p_{T1}^{+}\right)+P\left(S3|p_{R0}^{-},p_{T0}^{-},p_{R1}^{+},p_{T1}^{+}\right)\}$$

Expected sample size (ESS) and optimal design

A minimum of $N_1^{-}+N_1^{+}$ patients need to be included. According to the number of responses and non-toxicities observed in the interim analysis, three scenarios are considered: none or $N_2^{-}+N_2^{+}$ or N_{2e}^{+} additional patients will need to be included at the second stage. The ESS is determined as follows:

$$ESS\left(p_R^{-},p_T^{-},p_R^{+},p_T^{+}\right)=N_1^{-}+N_1^{+}+\left(N_2^{-}+N_2^{+}\right)\times P\left(X_{R1}^{-}\geq k_{R1}^{-}\right)\times P\left(X_{T1}^{-}\geq k_{T1}^{-}\right)+N_{2e}^{+}\times P\left(X_{R1}^{+}\geq k_{R1}^{+}\right)\times P\left(X_{T1}^{+}\geq k_{T1}^{+}\right)\times P\left(X_{R1}^{-}<k_{R1}^{-}\ or\ X_{T1}^{-}<k_{T1}^{-}\right)$$

As proposed by Parashar et al. [10], the optimal design (k_{R1}^{-}, k_{T1}^{-}, k_{R1}^{+}, k_{T1}^{+}, N_1^{-}, N_1^{+}, k_{Re}^{+}, k_{Te}^{+}, N_e^{+}, k_R^{-}, k_T^{-}, k_R^{+}, k_T^{+}, N^{-}, N^{+}) is defined as the one that minimizes the maximum ESS under (H_{01}^{-},H_{01}^{+}) or (H_{10}^{-},H_{10}^{+}) (i.e. $\max\{ESS(p_{R0}^{-},p_{T1}^{-},p_{R0}^{+},p_{T1}^{+}),ESS(p_{R1}^{-},p_{T0}^{-},p_{R1}^{+},p_{T0}^{+})\}$) while controlling type I (α_R and α_T) and type II (β) error rates. To determine the optimal design, 15 parameters need to be estimated. To reduce the computational burden, a similar approach to the one proposed by Jones et al. [8] is used. Parameters (N_1^{-}, N^{-}, k_{R1}^{-}, k_{T1}^{-}, k_R^{-}, k_T^{-}) and (N_1^{+}, k_{R1}^{+}, k_{T1}^{+}) are derived from the BD design with (p_{R0}^{-}, p_{T0}^{-}, p_{R1}^{-}, p_{T1}^{-}, $\alpha_R/2$, $\alpha_T/2$, β) and (p_{R0}^{+}, p_{T0}^{+}, p_{R1}^{+}, p_{T1}^{+}, $\alpha_R/2$, $\alpha_T/2$, β), respectively (type I error rates are set at $\alpha_R/2$ and

Results

Examples of SABD design

Three examples of the SABD design are considered. In the first example, hypotheses are based on the GERICO10 phase II trial which aimed to evaluate the feasibility of a chemotherapy treatment with docetaxel-prednisone in patients age 75 or older, classified as vulnerable or frail according to the International Society of Geriatric Oncology criteria, with castration-resistant metastatic prostate cancer [14]. Same hypotheses are defined for the two co-primary endpoints in the two subgroups ($p_{R0}^{-(+)}=p_{T0}^{-(+)}=0.70$ and $p_{R1}^{-(+)}=p_{T1}^{-(+)}=0.90$). In the second example, different hypotheses are defined between the two co-primary endpoints in the two subgroups ($p_{R0}^{-(+)}=0.30$, $p_{T0}^{-(+)}=0.60$, $p_{R1}^{-(+)}=0.60$ and $p_{T1}^{-(+)}=0.90$). In the third example, different hypotheses are defined between the two co-primary endpoints and between the two subgroups for non-toxicity ($p_{R0}^{-(+)}=0.10$, $p_{T0}^{-(+)}=0.60$, $p_{R1}^{-(+)}=0.40$, $p_{T1}^{-}=0.80$ and $p_{T1}^{+}=0.90$). Type I error rates (α_R and α_T) and overall power (1-β) are set at 10% and 80%, respectively. The hypotheses, parameters and operating characteristics for the three examples are summarized in Table 2.

In the first example, a maximum of 67 patients need to be included and the interim analysis is performed after the enrollment of 10 patients into each subgroup. According to the number of responses and non-toxicities observed at the end of the first stage, three scenarios are possible: the study is stopped for futility if at most 7 responses or non-toxicities are observed in the negative and positive subgroups; enrollment continues in an unselected population with the recruitment of additional 25 ($N_2^{-}=N^{-}$-N_1^{-}) and 22 ($N_2^{+}=N^{+}$-N_1^{+}) patients in the negative and positive subgroups, respectively, if at

Table 2 Examples of stratified adaptive Bryant & Day (SABD) design (ESS_{RiTj} and PET_{RiTj} correspond to *ESS* (p_{Ri}^-,p_{Tj}^-,p_{Ri}^+,p_{Tj}^+) and *PET* (p_{Ri}^-,p_{Tj}^-,p_{Ri}^+,p_{Tj}^+), respectively)

Hypotheses				Parameters								Operating characteristics		
p_{R0}	p_{R1}^- p_{R1}^+	p_{T0}	p_{T1}^- p_{T1}^+	k_{R1}^- k_{R1}^+	k_{T1}^- k_{T1}^+	N_1^- N_1^+	k_{Re}^+ k_{Te}^+	N_e^+	k_R^- k_R^+	k_T^- k_T^+	N^- N^+	Attained α_R / α_T / 1- β	max (ESS_{R0T1},ESS_{R1T0})	min (PET_{R0T1},PET_{R1T0})
0.70	0.90 0.90	0.70	0.90 0.90	8 8	8 8	10 10	29 29	35	29 27	29 27	35 32	0.094 / 0.094 / 0.800	42.5	0.415
0.30	0.60 0.60	0.60	0.90 0.90	4 4	7 7	9 9	10 16	21	11 8	18 13	23 16	0.094 / 0.093 / 0.800	25.7	0.554
0.10	0.40 0.40	0.60	0.80 0.90	3 2	12 7	17 9	4 13	16	7 3	26 9	35 10	0.087 / 0.096 / 0.800	32.1	0.580

least 8 responses and non-toxicites are observed in the negative subgroup; enrollment continues in the positive subgroup only with the recruitment of additional 25 patients (enrichment: $N_{2e}^+ = N_e^+ - N_1^+$) if at most 7 responses and non-toxicities are observed in the negative subgroup and at least 8 responses and non-toxicites are observed in the positive subgroup. At the end of the second stage after the enrollment of 35 and 32 patients in the negative and positive subgroups, respectively, a «go-decision» is declared in the unselected population or in the positive subgroup only if at least 29 responses and 27 non-toxicites are observed in the negative or in the positive subgroup only, respectively. After the enrollment of 35 patients in the positive subgroup (enrichment), a «go-decision» is declared in the positive subgroup only if at least 29 responses and 29 non-toxicites are observed. The ESS and the PET for insufficient activity and/or excessive toxicity equate to 42.5 patients and 41.5%, respectively.

In the second example, a maximum of 39 patients need to be included and the interim analysis is performed after the enrollment of 9 patients into each subgroup. According to the number of responses and non-toxicities observed at the end of the first stage, three scenarios are possible: the study is stopped for futility; enrollment continues in an unselected population with the recruitment of additional 14 ($N_2^- = N^- - N_1^-$) and 7 ($N_2^+ = N^+ - N_1^+$) patients in the negative and positive subgroups, respectively; enrollment continues in the positive subgroup only with the recruitment of additional 12 patients (enrichment: $N_{2e}^+ = N_e^+ - N_1^+$). The ESS and the PET for insufficient activity and/or excessive toxicity equate to 25.7 patients and 55.4%, respectively.

In the third example, a maximum of 45 patients need to be included and the interim analysis is performed after the enrollment of 17 and 9 patients in the negative and positive subgroups, respectively. According to the number of responses and non-toxicities observed at the end of the first stage, three scenarios are possible: the study is stopped for futility; enrollment continues in an unselected population with the recruitment of additional 18 ($N_2^- = N^- - N_1^-$) and 1 ($N_2^+ = N^+ - N_1^+$) patients in the negative and positive subgroups, respectively; enrollment continues in the positive subgroup only with the recruitment of additional 7 patients (enrichment: $N_{2e}^+ = N_e^+ - N_1^+$). The ESS and the PET for insufficient activity and/or excessive toxicity equate to 32.1 patients and 58.0%, respectively.

A selection of SABD designs with pre-specified hypotheses are detailed in Supplementary Table 1.

An optimal SABD design requires a total of 15 parameters to be estimated. This involves a very large number of combinations and therefore necessitates an extensive computational effort when using standard software. For example, the computation time needed to determine an optimal SABD design can vary from a few minutes or hours to several weeks, depending on the hypotheses, using R software (https://cran.r-project.org/).

Simulation studies

Simulations were carried out to investigate the operating characteristics of the SABD design and to compare to a parallel BD design (i.e. two parallel studies with one BD design in each subgroup). Three case studies corresponding to the three examples presented in previous section were considered. Type I error rate and power, for the SABD design, were set at 10% and 80%, respectively. In the parallel BD design, adjustment for multiplicity was performed to achieve an adequate overall type I error rate and sufficient statistical power to draw meaningful conclusions in the unselected population or only in the positive subgroup. Type I error rate and power were therefore set at 5% (i.e. $\alpha_R/2$ and $\alpha_T/2$) and 90% (i.e. $1-\beta/2$) in each subgroup for parallel BD design, respectively. Four scenarios were considered:

- Scenario 1A: simulations were performed under $H_{01}^{-(+)}$ ($p_R^{-(+)}=p_{R0}^{-(+)}$ and $p_T^{-(+)}=p_{T1}^{-(+)}$) to assess type I error rate α_R and PET.
- Scenario 1B: simulations were performed under $H_{10}^{-(+)}$ ($p_R^{-(+)}=p_{R1}^{-(+)}$ and $p_T^{-(+)}=p_{T0}^{-(+)}$) to assess type I error rate α_T and PET.
- Scenario 2: simulations were performed under H_{00}^{-} ($p_R^{-}=p_{R0}^{-}$ and $p_T^{-}=p_{T0}^{-}$) and H_{11}^{+} ($p_R^{+}=p_{R1}^{+}$ and $p_T^{+}=p_{T1}^{+}$) to evaluate the probability of detecting heterogeneity at the first stage (i.e. stop enrollment for futility in the negative subgroup) and the probability of considering the treatment as promising in the positive subgroup (i.e. reject H_0^{+}).
- Scenario 3: simulations were performed under $H_{11}^{-(+)}$ ($p_R^{-(+)}=p_{R1}^{-(+)}$ and $p_T^{-(+)}=p_{T1}^{-(+)}$) to evaluate the probability of considering the treatment as promising in the unselected population (i.e. reject H_0^{-} and H_0^{+}).

For each case study and scenario, 100 000 hypothetical trials were simulated. The number of responses and non-toxicities were randomly generated using binomial distributions $B(N,p)$ with N corresponding to the number of patients presented in Table 2 (N_1^{-}, N_1^{+}, N^{-}—N_1^{-}, N^{+}—N_1^{+} and N_e^{+}—N_1^{+}) and p corresponding to the true response and non-toxicity rates defined above (p_R^{-}, p_T^{-}, p_R^{+} and p_T^{+}).

The ESSs were also estimated for each case study and scenario. Simulation results are presented in Table 3.

In all three case studies, the maximum sample size was larger with the parallel BD design with respectively 88, 62 and 79 patients compared with 67, 39 and 45 patients for the SABD.

Scenarios 1A and 1B, the SABD gave the smallest ESS with a maximum of 42.4, 25.7 and 32.0 patients compared to the parallel BD with a maximum of 52.8, 35.1 and 44.8 patients for the three case studies, respectively. The probability of rejecting H_0^{-} or H_0^{+} (i.e. type I error rates α_R and α_T) was approximately 10% for each design and case study (except in scenarios 1A and 1B for case study 2 and 3 with the parallel BD, respectively). The PET varied between 41 and 46% for case study 1 and was higher when using the SABD with a minimum of 55.5% and 58.2% compared to the parallel BD with a minimum of 40.8% and 49.3% for the case studies 2 and 3, respectively.

In scenario 2, for each case study, the probability of rejecting H_0^{+} is higher when using the parallel BD (approximately 90%) compared to the SABD (approximately 80%). The probability of detecting heterogeneity at the first stage was at least 80% for each design and case study, except for the SABD in case study 1 (73.9%). The SABD gave the smallest ESS with respectively 50.6, 31.0 and 34.9 patients compared to the parallel BD with 63.4, 42.4 and 45.7 patients for the three case studies.

In scenario 3, the probability of rejecting H_0^{-} and H_0^{+} was approximately 80% for each design and case study. The ESS was lower for the three case studies when using the SABD, with 63.5, 37.4 and 43.5 patients compared to 85.1, 59.2 and 75.8 patients for the parallel BD, respectively.

Discussion

The stratified adaptive phase II design developed and presented in this paper takes into account the heterogeneity of a population when considering co-primary endpoints. The SABD design based on the Jones et al. [8] and Parashar et al. [10] algorithm, allows to include two pre-defined subgroups and to identify whether the therapeutic benefits one of these subgroups at the end of the first or the second stage. Different hypotheses can be defined between the subgroups and/or co-primary endpoints. We used three case studies to simulate different scenarios and investigate the operating characteristics of the SABD approach. The results demonstrate good statistical performances for the SABD when compared to the parallel BD (one BD for each subgroup). The SABD indeed allows to reduce the number of patients exposed to an insufficiently active or overly toxic treatment (scenarios 1A and 1B). The ESS required to reach an adequate statistical power to draw meaningful conclusions in the unselected population is also lower compared to the parallel BD (scenario 3). The same trend is observed in scenario 2 but the parallel BD yields a higher statistical power to conclude to the feasibility of the treatment in the positive subgroup only (i.e. «go-decision»). If there was heterogeneity between the two subgroups, the probability of detecting it at the first stage was generally at least 80%. To account for multiplicity and obtain an adequate overall type I error rate of 10%, α_T and α_R were set at 5% for each BD. In case study 2 and 3, optimal BD designs were determined using binomial probabilities with α_T and α_R less than 3.5%. This could explain the lower type I error rate observed in scenarios 1A and 1B for the parallel BD, compared to the SABD.

Given that the endpoint was two-dimensional, alternative case studies or scenarios may also be considered. It would, for instance, be interesting to investigate the statistical performance of the SABD design when heterogeneity only affects one dimension.

Similarly to the BD design, the SABD design assumes that the co-primary endpoints are independent. An alternative to the BD design which pre-defines the association between co-primary endpoints has also been developed [15]. Such an extension of the SABD design to correlated endpoints implies, among other things, to consider a bivariate binomial distribution with a correlation between

Table 3 Simulation results

			Parallel BD	SABD
Case study 1				
		Maximum sample size	88	67
1A	$p_R^- = 0.7/p_T^- = 0.9$ $p_R^+ = 0.7/p_T^+ = 0.9$	Expected sample size	52.8	42.4
		Probability of early termination	0.458	0.416
		Probability of rejecting H_0^- or H_0^+	0.092	0.094
1B	$p_R^- = 0.9/p_T^- = 0.7$ $p_R^+ = 0.9/p_T^+ = 0.7$	Expected sample size	52.8	42.4
		Probability of early termination	0.457	0.416
		Probability of rejecting H_0^- or H_0^+	0.092	0.094
2	$p_R^- = 0.7/p_T^- = 0.7$ $p_R^+ = 0.9/p_T^+ = 0.9$	Expected sample size	63.4	50.6
		Probability of rejecting H_0^+	0.906	0.801
		Probability of detecting heterogeneity (1st stage)	0.840	0.739
3	$p_R^- = 0.9/p_T^- = 0.9$ $p_R^+ = 0.9/p_T^+ = 0.9$	Expected sample size	85.1	63.5
		Probability of rejecting H_0^- and H_0^+	0.826	0.800
Case study 2				
		Maximum sample size	62	39
1A	$p_R^- = 0.3/p_T^- = 0.9$ $p_R^+ = 0.3/p_T^+ = 0.9$	Expected sample size	34.5	25.7
		Probability of early termination	0.430	0.555
		Probability of rejecting H_0^- or H_0^+	0.083	0.093
1B	$p_R^- = 0.6/p_T^- = 0.6$ $p_R^+ = 0.6/p_T^+ = 0.6$	Expected sample size	35.1	24.3
		Probability of early termination	0.408	0.627
		Probability of rejecting H_0^- or H_0^+	0.054	0.092
2	$p_R^- = 0.3/p_T^- = 0.6$ $p_R^+ = 0.6/p_T^+ = 0.9$	Expected sample size	42.4	31.0
		Probability of rejecting H_0^+	0.905	0.802
		Probability of detecting heterogeneity (1st stage)	0.807	0.799
3	$p_R^- = 0.6/p_T^- = 0.9$ $p_R^+ = 0.6/p_T^+ = 0.9$	Expected sample size	59.2	37.4
		Probability of rejecting H_0^- and H_0^+	0.821	0.800
Case study 3				
		Maximum sample size	79	45
1A	$p_R^- = 0.1/p_T^- = 0.8$ $p_R^+ = 0.1/p_T^+ = 0.9$	Expected sample size	43.7	31.2
		Probability of early termination	0.511	0.620
		Probability of rejecting H_0^- or H_0^+	0.055	0.086
1B	$p_R^- = 0.4/p_T^- = 0.6$ $p_R^+ = 0.4/p_T^+ = 0.6$	Expected sample size	44.8	32.0
		Probability of early termination	0.493	0.582
		Probability of rejecting H_0^- or H_0^+	0.083	0.096
2	$p_R^- = 0.1/p_T^- = 0.6$ $p_R^+ = 0.4/p_T^+ = 0.9$	Expected sample size	45.7	34.9
		Probability of rejecting H_0^+	0.904	0.802
		Probability of detecting heterogeneity (1st stage)	0.864	0.825
3	$p_R^- = 0.4/p_T^- = 0.8$ $p_R^+ = 0.4/p_T^+ = 0.9$	Expected sample size	75.8	43.5
		Probability of rejecting H_0^- and H_0^+	0.821	0.801

the two co-primary endpoints but also between the two subgroups. This merits further investigation. A simulation study assessing the impact of an erroneous assumption of this pre-defined association however recommends using the BD design. Indeed, incorrectly assuming independence of endpoints only slightly increases the type I and II error rates. This is in contrast to wrongly defining the level of correlation between co-primary endpoints which results in a significant loss of statistical power and an increase in the type I error rate [16]. Future studies will be required to investigate the impact of wrongly assuming independence of co-primary endpoints on the performance of stratified design approaches.

The Jones et al. [8] and Parashar et al. [10] algorithm assumes that there is a hierarchy between subgroups, such that the true response and non-toxicity rate will

always be higher in the positive subgroup. This may lead to the results of the positive subgroup having no impact on the outcome of the study if promising results are observed in the negative subgroup in the interim and the final analysis. Indeed, if this hierarchy assumption is incorrect, enrollment of an unselected population may be continued even though promising results are only observed in the negative subgroup at the interim analysis. In this scenario, an additional type I error may occur by declaring a «go-decision» in the unselected population in the case where true response and non-toxicity rates are considered as acceptable in the negative subgroup only (i.e. wrongly reject H_0^+). Zang & Yuan proposed a reverse approach to address this shortfall [17]. The trial is initially only conducted in the positive subgroup and then in the negative subgroup if promising results are observed in the positive subgroup. An alternative two-stage approach has also been published by Dutton & Holmes [18]. In this approach, futility is first tested in the unselected population and then in the negative or positive subgroup depending on whether or not promising results are observed. The impact of an assumption-based error in relation to hierarchy remains to be evaluated and deserves further investigation.

An optimal SABD design requires a total of 15 parameters to be estimated. A similar approach to the one proposed by Jones et al. [8], which is described in «Expected sample size (ESS) and optimal design» section, was used to reduce the number of parameters that needed to be determined and thus also reduce the computational burden. Further work is required to provide technical solutions and to determine optimal designs over the 15-dimensional parameter space.

Conclusions

The SABD design allows to independently assess two dimensions through co-primary endpoints in a heterogeneous population without dramatically increasing the sample size. This is particularly useful for geriatric clinical oncology trials as it allows to stratify the population according to a geriatric criterion and to identify a subgroup of interest that has an acceptable and clinically relevant benefit-risk ratio at the end of the first or the second stage. As population heterogeneity is not limited to older populations, the SABD design may also be applicable to other study populations such as children or adolescents and young adults [19]. Children populations are heterogeneous particularly in terms of age, with tolerance of a treatment potentially dependent on these aspects [20]. Our novel SABD approach may also be envisaged for phase II trials of targeted therapies based on a biomarker (positive versus negative) to select the appropriate study population for the subsequent phase III trial.

Abbreviations

BD: Bryant & Day; SABD: Stratified Adaptive Bryant & Day; ESS: Expected Sample Size; PET: Probability of Early Termination.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01748-w.

Additional file 1: Supplementary Table 1. Stratified adaptive Bryant & Day (SABD) designs with $\alpha_R = 0.1$, $\alpha_T = 0.1$, $\beta = 0.2$ (ESS_{RiTj} and PET_{RiTj} correspond to $ESS(p_{Ri}^-, p_{Tj}^-, p_{Ri}^+, p_{Tj}^+)$ and $PET(p_{Ri}^-, p_{Tj}^-, p_{Ri}^+, p_{Tj}^+)$, respectively).

Acknowledgements

The authors would like to thank 'La Ligue Nationale Contre le Cancer, France' (Comité des Pyrénées-Orientales, Comité de la Meuse, Comité du Maine-et-Loire) for their financial support and Petra Neufing, native speaker, for her assistance with the English proofreading.

Authors' contributions

BC, EL and TF developed the novel design and wrote the manuscript. PS and JMB reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by a grant from 'La Ligue Nationale Contre le Cancer, France' (PI: Thomas Filleron).

Availability of data and materials

The data generated and used during this study are available from the corresponding author on reasonable request.
The R program implementing the proposed SABD design is available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Biostatistics & Health Data Science Unit, Institut Claudius Regaud - IUCT-O, 1 avenue Irène Joliot-Curie, 31059 Cedex 9 Toulouse, France. [2]Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France. [3]Biostatistics Unit, Institut Paoli-Calmettes, Marseille, France. [4]Aix Marseille Université, INSERM, IRD, SESSTIM, Marseille, France.

Received: 26 April 2022 Accepted: 4 October 2022
Published online: 26 October 2022

References

1. Sedrak MS, Mohile SG, Sun V, Sun C-L, Chen BT, Li D, et al. Barriers to clinical trial enrollment of older adults with cancer: A qualitative study of the perceptions of community and academic oncologists. J Geriatr Oncol. 2020;11:327–34.
2. Wildiers H, Mauer M, Pallis A, Hurria A, Mohile SG, Luciani A, et al. End points and trial design in geriatric oncology research: a joint European

organisation for research and treatment of cancer–Alliance for Clinical Trials in Oncology-International Society Of Geriatric Oncology position article. J Clin Oncol. 2013;31:3711–8.
3. Cabarrou B, Mourey L, Dalenc F, Balardy L, Kanoun D, Roché H, et al. Methodology of phase II clinical trials in metastatic elderly breast cancer: a literature review. Breast Cancer Res Treat. 2017. https://doi.org/10.1007/s10549-017-4278-5.
4. Ferrat E, Paillaud E, Caillet P, Laurent M, Tournigand C, Lagrange J-L, et al. Performance of Four Frailty Classifications in Older Patients With Cancer: Prospective Elderly Cancer Patients Cohort Study. J Clin Oncol. 2017;35:766–77.
5. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics. 1982;38:143–51.
6. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10:1–10.
7. A'Hern RP. Widening eligibility to phase II trials: constant arcsine difference phase II trials. Control Clin Trials. 2004;25:251–64.
8. Jones CL, Holmgren E. An adaptive Simon two-stage design for phase 2 studies of targeted therapies. Contemp Clin Trials. 2007;28:654–61.
9. Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. How a new stratified adaptive phase II design could improve targeting population. Stat Med. 2011;30:1555–62.
10. Parashar D, Bowden J, Starr C, Wernisch L, Mander A. An optimal stratified Simon two-stage design. Pharm Stat. 2016;15:333–40.
11. Cabarrou B, Sfumato P, Mourey L, Leconte E, Balardy L, Martinez A, et al. Addressing heterogeneity in the design of phase II clinical trials in geriatric oncology. Eur J Cancer. 2018;103:120–6.
12 Sedrak MS, Freedman RA, Cohen HJ, Muss HB, Jatoi A, Klepin HD, et al. Older adult participation in cancer clinical trials: A systematic review of barriers and interventions. CA Cancer J Clin. 2021;71:78–92.
13. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995;51:1372–83.
14 Mourey L, Sevin E, Latorzeff I, Houede N, Meunier J, Priou F, et al. Is docetaxel-prednisone (DP) feasible in frail elderly (age 75 or older) patients with castration-resistant metastatic prostate cancer (CRMPC)? GERICO10-GETUG P03 trial: A trial from elderly and genitourinary oncology UNICANCER groups. JCO. 2012;30(5_suppl):93–93.
15. Conaway MR, Petroni GR. Designs for phase II trials allowing for a trade-off between response and toxicity. Biometrics. 1996;52:1375–86.
16. Tournoux C, De Rycke Y, Médioni J, Asselain B. Methods of joint evaluation of efficacy and toxicity in phase II clinical trials. Contemp Clin Trials. 2007;28:514–24.
17. Zang Y, Yuan Y. Optimal sequential enrichment designs for phase II clinical trials. Stat Med. 2017;36:54–66.
18. Dutton P, Holmes J. Single arm two-stage studies: Improved designs for molecularly targeted agents. Pharm Stat. 2018;17:761–9.
19. Sposto R, Gaynon PS. An adjustment for patient heterogeneity in the design of two-stage phase II trials. Stat Med. 2009;28:2566–79.
20. Paoletti X, Geoerger B, Doz F, Baruchel A, Lokiec F, Le Tourneau C. A comparative analysis of paediatric dose-finding trials of molecularly targeted agent with adults' trials. Eur J Cancer. 2013;49:2392–402.

Publisher's Note