BMC Medical Research Methodology

RESEARCH

Open Access

IBIS: identify biomarker-based subgroups with a Bayesian enrichment design for targeted combination therapy

Xin Chen, Jingyi Zhang, Liyun Jiang and Fangrong Yan*

Abstract

Background Combination therapies directed at multiple targets have potentially improved treatment effects for cancer patients. Compared to monotherapy, targeted combination therapy leads to an increasing number of subgroups and complicated biomarker-based efficacy profiles, making it more difficult for efficacy evaluation in clinical trials. Therefore, it is necessary to develop innovative clinical trial designs to explore the efficacy of targeted combination therapy in different subgroups and identify patients who are more likely to benefit from the investigational combination therapy.

Methods We propose a statistical tool called 'IBIS' to Identify BIomarker-based Subgroups and apply it to the enrichment design framework. The IBIS contains three main elements: subgroup division, efficacy evaluation and subgroup identification. We first enumerate all possible subgroup divisions based on biomarker levels. Then, Jensen–Shannon divergence is used to distinguish high-efficacy and low-efficacy subgroups, and Bayesian hierarchical model (BHM) is employed to borrow information within these two subsets for efficacy evaluation. Regarding subgroup identification, a hypothesis testing framework based on Bayes factors is constructed. This framework also plays a key role in go/no-go decisions and enriching specific population. Simulation studies are conducted to evaluate the proposed method.

Results The accuracy and precision of IBIS could reach a desired level in terms of estimation performance. In regard to subgroup identification and population enrichment, the proposed IBIS has superior and robust characteristics compared with traditional methods. An example of how to obtain design parameters for an adaptive enrichment design under the IBIS framework is also provided.

Conclusions IBIS has the potential to be a useful tool for biomarker-based subgroup identification and population enrichment in clinical trials of targeted combination therapy.

Keywords Biomarker, Subgroup identification, Adaptive enrichment design, Combination therapy, Bayesian hierarchical model (BHM), Two-stage design

*Correspondence:
Fangrong Yan
f.r.yan@163.com
Full list of author information is available at the end of the article

Background

In recent years, the rapid development of targeted combination therapy has brought novel treatment options for cancer patients. For example, atezolizumab (PD-L1 inhibitor) plus bevacizumab (VEGF inhibitor) could maintain clinically meaningful survival benefits compared with sorafenib in patients with unresectable hepatocellular carcinoma [1]. First-line treatment with nivolumab (PD-1 inhibitor) plus ipilimumab (CTLA4 inhibitor) resulted in a longer duration of overall survival than did chemotherapy in patients with advanced non-small-cell lung cancer [2]. A single-arm, phase Ib-II trial of pembrolizumab (PD-1 inhibitor) plus trastuzumab (HER2 inhibitor) also demonstrated activity and durable clinical benefit in patients with PD-L1-positive, trastuzumab-resistant, advanced, HER2-positive breast cancer [3]. Such combination therapies are directed at multiple therapeutic targets and may improve treatment response, prevent development of resistance, or reduce adverse events. However, the efficacy of targeted combination therapy could be heterogeneous across subgroups and is generally related to the levels of certain predictive biomarkers [4]. Traditional treatment strategy without selecting population is no longer desirable. Compared to monotherapy, subgroup identification is usually more complicated for targeted combination therapy due to the increasing number of subgroups. For example, patients treated with a PD-1 inhibitor are often divided into three subgroups, including PD-L1 less than 1, 1–49%, and 50% or greater, while patients treated with a HER2 inhibitor can be divided into HER2-positive and HER2-negative subgroups. Thus, if these two kinds of therapies are combined for treatment, there are consequently a total of six subgroups. This may result in insufficient sample sizes and slow recruitment for some subgroups, leading to efficacy evaluation challenges in clinical trials. One motivation example of this paper is an ongoing phase 1b, open-label, 2-part, multicenter, non-randomized, multiple-dose study which evaluates DS-8201a in combination with pembrolizumab in participants with advanced/metastatic breast cancer or non-small cell lung cancer (ClinicalTrials.gov Identifier: NCT04042701) [5]. DS-8201a is an anti-HER2 antibody-drug conjugate (ADC) with a novel topoisomerase I inhibitor and pembrolizumab is a PD-1 inhibitor. Therefore, it is highly likely that the efficacy of the drug combination is related to patients' expression levels of HER2 and PD-L1. The dose expansion part of the study includes breast cancer patients with both HER2-positive and HER2-low-positive, and the inclusion criteria does not limit the expression level of PD-L1. Therefore, although the primary objective of this example is not exactly to identify subgroups based on these two predictive biomarkers, we regard it as a scenario where a data-driven subgroup identification is possible.

Moreover, even if the target populations of both single drugs have been identified through historical studies, we cannot claim that the target population of the combination therapy is simply the intersection of those two populations because combination therapy can potentially enhance efficacy and reduce drug resistance by targeting multiple key pathways in a synergistic or an additive manner [6]. Unlike monotherapy, where there is usually a monotonic relationship between efficacy and biomarker level, the efficacy profiles of combination therapy across subgroups could become more complicated due to the existence of interaction effects. Therefore, it is necessary to develop new innovative clinical trial designs to explore the efficacy of targeted combination therapy in different subgroups and identify patients who are more likely to benefit from the investigational combination therapy.

Another recent change in medical practice is the increasing refinement of biomarker-based subgroup classification, representing a shift from dichotomy to multilevel classification. For example, patients with breast cancer are usually divided into HER2-positive and HER2-negative subgroups in clinical practice. For HER2-positive patients, trastuzumab-based or other HER2-targeted drug regimens are now standards of care [7]. However, recent studies have shown that HER2-low-positive and HER2-zero breast cancers, although generally classified as HER2-negative, are distinct in terms of prognosis and response to treatment [8]. Preclinical studies of DS-8201, an anti-HER2 ADC, indicate that the antitumor activity of the drug is dependent on HER2 expression level rather than on HER2 amplification [9]. An early-phase clinical trial showed that the drug has a certain effect on breast cancer patients with IHC2+ and IHC1+, while IHC1+ is generally classified as HER2-negative [10]. This evolving paradigm of subgroup classification is more consistent with the concept of precision medicine, but it may further increase the number of subgroups, causing challenges with respect to subgroup identification and efficacy evaluation. As the treatment effects may be similar in adjacent biomarker-based subgroups (e.g., IHC2+ and IHC1+), one possible strategy to address this issue is to borrow information across similar subgroups. Therefore, we propose a tool called 'IBIS' in this article to detect the potential similarities across subgroups and Identify BIomarker-based Subgroups with higher efficacy.

We also extend IBIS to an adaptive enrichment design framework to increase its applicability. The adaptive enrichment design can adjust the inclusion/exclusion criteria according to a pre-specified plan based on the results of interim analysis, allowing the flexibility to explore the efficacy of investigational drugs for different subgroups. Several adaptive enrichment designs have been proposed, some of which take into account the case of a single dichotomous biomarker [11–21]. In these trial designs, patients regardless of marker status are enrolled at stage I. Then, an interim analysis is performed to decide whether to continue enrolling the entire population or to enroll only biomarker-positive patients. Some other studies considered a more general case of nested subgroups and focused on subgroup selection, assuming multiple pre-specified subgroups with or without a priori ordering [22–26]. However, when it comes to two or more predictive biomarkers, the efficacy of the drug combination may be partially ordered. For example, we can assume the efficacy of PD-1 inhibitors in patients with PD-L1 $\geq$ 50% will not be lower than that of patients with PD-L1 < 50%; the efficacy of HER2 inhibitors in patients with HER2-positive will not be lower than that of patients with HER2-negative. However, we cannot judge whether a combination of PD-1 inhibitor and HER2 inhibitor is better for patients with PD-L1 $\geq$ 50% and HER2-negative than for patients with PD-L1 < 50% and HER2-positive. Therefore, those existing designs with or without a priori ordering for subgroups cannot completely meet the requirements of subgroup identification with multiple predictive biomarkers. To the best of our knowledge, there are limited systematic studies on the issues of biomarker-based subgroup identification and population enrichment in terms of targeted combination therapy. There are also few studies considering similarities across adjacent subgroups. Considering the increasing number of subgroups and complicated biomarker-based efficacy profiles, it is of great significance to propose new design methods to identify subgroups and enrich populations for targeted combination therapy.

Methods

IBIS design

Subgroup division

Considering a clinical trial whose primary objective is to determine whether a two-agent targeted drug combination (e.g., pembrolizumab plus trastuzumab) is effective for some specific subgroups, we first divide the subgroups into high-efficacy and low-efficacy subsets. This type of division could be an issue with the refinement of biomarker-based subgroup classification. It is assumed that two corresponding biomarkers are incorporated, denoted as *Biomarker1* and *Biomarker2* (e.g., PD-L1 and HER2). The entire population can be divided into K ordered subgroups based on *Biomarker1* or J ordered subgroups based on *Biomarker2*; thus the total number of subgroups is $K \times J$. Assuming that the efficacy of both targeted agents increases monotonically with biomarker levels, i.e., marginal monotonicities, a total of G high-efficacy subsets $\Pi_g (g = 1, \ldots, G)$ can be listed, where the G_{th} subset Π_G represents the entire population. Each high-efficacy subset has at least one high-efficacy patient subgroup, i.e., $\Pi_g \neq \emptyset$. Taking one of the simplest forms ($K = 2$, $J = 2$) as an example, there are five possible subgroup divisions ($G = 5$, Fig. 1A).

As the number of subgroups increases, the possible situations of subgroup division expand rapidly. For example, when $K = 3$ and $J = 4$, there are 34 possible divisions altogether, four among which are shown in Fig. 1B. A computer algorithm can be used to enumerate all possible divisions satisfying the marginal monotonic assumption. The algorithm is given below.

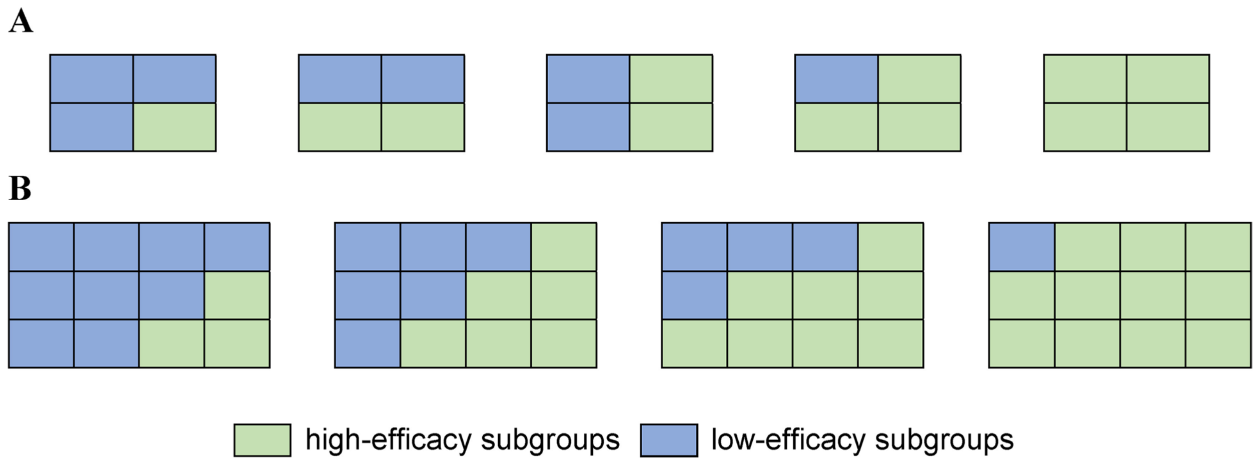

Fig. 1 Examples of subgroup divisions. The level of *Biomarker1* increases from top to bottom, and the level of *Biomarker2* increases from left to right. The high-efficacy subgroups and low-efficacy subgroups are represented in green and blue, respectively. **A** Five possible subgroup divisions when $K = 2$ and $J = 2$. **B** Four possible subgroup divisions when $K = 3$ and $J = 4$, and there are 34 divisions in total

(1) Let $r\left(k,j\right)$ denote the variable indicating the magnitude of efficacy for subgroup$\left(k,j\right)$, where $k = 1,\ldots, K$ and $j = 1,\ldots, J$. To describe the algorithm more conveniently, let $r(k,0) = r\left(0,j\right) = 0$.

(2) Use the following loop to sequentially assign values to $r\left(k,j\right)$:

```
FOR $k = 1,\ldots, K$
  FOR $j = 1,\ldots, J$
    DO $r\left(k,j\right) = \mathrm{runif}\left(\max\left(r\left(k-1,j\right),\ r\left(k,j-1\right)\right),1\right)$
   where runif$(a,b)$ denotes a uniformly distributed random number between $a$ and b, and max$(a,b)$ denotes the larger of $a$ and $b$.
```

(3) Sort $r\left(k,j\right)$ to get an ordering.

(4) Repeat steps 2 and 3 to get N orderings (N is a large number, such as 10^6). Eliminate duplicates in these orderings.

(5) Partition the orderings to obtain possible situations of the high-efficacy subset.

(6) Eliminate duplicates in the obtained divisions without regard to the orderings within subset.

Efficacy evaluation

For ease of elucidation, we first consider the scenario of a single-arm trial. Let $Y_{(k,j)}$ denote the efficacy measure outcome for patients in subgroup (k,j), which follows a one-parameter exponential family distribution, i.e., $Y_{(k,j)} \sim f(\psi_{(k,j)})$. For example, if the response rate is the efficacy endpoint, whether a patient in subgroup (k,j) responds to investigational treatment can be viewed as following a Bernoulli distribution with probability $\psi_{(k,j)}$. As the treatment effects are relatively similar within the high-efficacy and low-efficacy subsets, we transform the original parameter $\psi_{(k,j)}$ into an exchangeable parameter $\theta_{(k,j)} = h(\psi_{(k,j)})$ in preparation for borrowing information via a hierarchical model. A typical example of the transform function $h(\cdot)$ is the logit function for a binary endpoint. Table 1 shows the situations for some other commonly used endpoints.

One simple way to model the efficacy outcome is applying Bayesian hierarchical models to borrow information across all subgroups. However, when the heterogeneity across subgroups is large, using this strategy may lead to substantial bias. Therefore, it is more preferable to classify subgroups into two subsets based on accumulated data and then borrow information within each subset. Let $\Pi_{\complement g}$ (low-efficacy subset) denote the complementary set for subset Π_g (high-efficacy subset). A Bayesian hierarchical model is constructed as follows to borrow information within the high-efficacy and low-efficacy subsets:

Table 1 Sampling model, transform function and example of clinical endpoints for different types of one-parameter exponential families

$f(\psi_{(k,j)})$	$h(\psi_{(k,j)})$	Example of clinical endpoint
Normal$(\psi_{(k,j)}, \sigma^2)$	$\psi_{(k,j)}$	Percentage change in tumor size
Bernoulli$(\psi_{(k,j)})$	logit$(\psi_{(k,j)})$	Tumor response rate
Exponential$(\psi_{(k,j)})$	log$(\psi_{(k,j)})$	Progression free survival
Poisson$(\psi_{(k,j)})$	log$(\psi_{(k,j)})$	Number of tumor-involved nodes

$f(\cdot)$, sampling model; $h(\cdot)$, transform function. The nuisance parameter σ^2 is assumed to be known for a normal distribution

$$\theta_{(k,j)} \sim N\left(\theta_g, \sigma_g^2\right), \text{if } \left(k,j\right) \in \Pi_g,$$

$$\theta_{(k,j)} \sim N\left(\theta_{\complement g}, \sigma_{\complement g}^2\right), \text{if } \left(k,j\right) \in \Pi_{\complement g},$$

where θ_g and $\theta_{\complement g}$ denote the average treatment effects for high-efficacy and low-efficacy subsets, respectively. Shrinkage parameters σ_g^2 and $\sigma_{\complement g}^2$ are the intersubgroup variances of treatment effects within these two subsets, controlling the degree of information borrowing. They do not need to be specified in advance and can be data-driven. If treatment effect estimates across subgroups within a subset are similar, then the posteriors of the intersubgroup variance will be smaller, inducing a strong borrowing. If treatment effect estimates across subgroups within a subset are very different, then less borrowing will occur. Normal distributions with large variances are usually taken as the priors for θ_g and $\theta_{\complement g}$. In terms of the priors for σ_g^2 and $\sigma_{\complement g}^2$, an inverse-gamma distribution $IG(a,b)$ can be adopted. Small values of a and b are set such that the priors are vague. We constrain $\theta_g > \theta_{\complement g}$ to avoid the potential computational issue of label switching when using the Gibbs sampler to sample posteriors.

To distinguish high-efficacy and low-efficacy subgroups, we use Jensen–Shannon divergence [27] to measure the distance between the two posterior distributions of the average treatment effect θ_g and $\theta_{\complement g}$, which is also a measure of the similarity between high-efficacy subset and low-efficacy subset:

$$\mathrm{JSD}\left(\theta_g \middle\| \theta_{\complement g}\right) = \frac{1}{2} D_{KL}\left(\theta_g \middle\| \tilde{\theta}\right) + \frac{1}{2} D_{KL}\left(\theta_{\complement g} \middle\| \tilde{\theta}\right)$$

where $\tilde{\theta} = \frac{1}{2}\left(\theta_g + \theta_{\complement g}\right)$. $D_{KL}(A \| B)$ denotes the Kullback–Leibler divergence between A and B, which is defined as follows when A and B are both continuous variables:

$$D_{KL}\left(A \middle\| B\right) = \int a(x) \log\left(\frac{a(x)}{b(x)}\right) dx,$$

where $a(x)$ and $b(x)$ are the probability densities of A and B, respectively. After calculating the Jensen–Shannon divergences for all subgroup divisions, the optimal division result is defined as the division that maximizes the Jensen–Shannon divergence between θ_g and $\theta_{\complement g}$, because that is when Π_g and $\Pi_{\complement g}$ are most dissimilar. Let C_H denote the high-efficacy subset in the optimal division:

$$C_H = \underset{\Pi_g}{\operatorname{argmax}}\left\{ \mathrm{JSD}\left(\theta_g \middle\| \theta_{\complement g}\right), g = 1, \ldots, G-1 \right\}.$$

Based on this optimal division, the posterior distribution of the treatment effect for each subgroup $\theta_{(k,j)}$ can be obtained by applying the aforementioned Bayesian hierarchical model. We use the Jensen–Shannon divergence here because it is based on the well-known Kullback–Leibler divergence, and it has the property of symmetry. Some other measures of distance between distributions, such as the Hellinger distance [28], may also be applicable.

The model introduced above can be easily extended to randomized controlled trials (RCTs). The treatment effects of the investigational drug and the control intervention for subgroup (k,j) are $\theta^T_{(k,j)}$ and $\theta^C_{(k,j)}$, respectively. Therefore, the effect size of interest $\theta_{(k,j)} = \theta^T_{(k,j)} - \theta^C_{(k,j)}$, and the above statistical model can still be applied.

Subgroup identification

The objective of subgroup identification is to find a collection of subgroups with clinically meaningful treatment effects. Suppose the hypothesis is as follows:

$$H_{0(k,j)} : \theta_{(k,j)} \leq \theta_0, H_{1(k,j)} : \theta_{(k,j)} > \theta_0 \text{ for } k = 1, \ldots, K, j = 1, \ldots, J.$$

In single-arm trials, θ_0 represents the minimum acceptable treatment effect, which is usually equal to the efficacy of the existing standard of care. In randomized controlled trials, θ_0 is the superiority margin and is usually taken as 0.

Bayes factors are used to test the hypothesis in IBIS. If the Bayes factor $BF_{(k,j)}$ corresponding to subgroup (k,j) is sufficiently large, i.e.,

$$BF_{(k,j)} = \frac{\Pr\left(H_{1(k,j)}|D\right)/\Pr\left(H_{0(k,j)}|D\right)}{\Pr\left(H_{1(k,j)}\right)/\Pr\left(H_{0(k,j)}\right)} > BF_{E(k,j)},$$

the investigational combination therapy is considered to be effective for subgroup (k,j), where D denotes the accumulated trial data and $BF_{E(k,j)}$ is a pre-specified threshold for (k,j). Under the Bayesian paradigm, we assign each of the hypotheses a prior probability of being true, as denoted by $\Pr(H_{0(k,j)})$ and $\Pr(H_{1(k,j)})$. Correspondingly, the posterior probabilities are denoted as $\Pr(H_{0(k,j)}|D)$ and $\Pr(H_{1(k,j)}|D)$. To satisfy the assumption that efficacy increases monotonically with biomarker levels, once the investigational therapy is considered effective for subgroup $\left(\tilde{k}, \tilde{j}\right)$, it is also deemed effective for subgroups in subset $C : \left\{k \geq \tilde{k}, j \geq \tilde{j}\right\}$.

We do not directly judge whether the investigational therapy is effective for C_H (i.e., making inference on θ_g at the subset level) because this is a kind of 'statistical' subgroup division, not a 'clinical' one. In addition, such a subset can always be identified as mentioned before. If the therapy is ineffective for all subgroups, making such judgments will inevitably lead to type I errors; if the therapy is effective for all subgroups, it will exclude some subgroups and result in type II errors. The main purpose for subgroup division in IBIS is to reduce the bias of estimation generated by BHM when there is heterogeneity across subgroups rather than directly inferring and making decisions based on the division results.

The reason why we use the Bayes factors rather than directly using the posterior probability $\Pr(H_{1(k,j)}|D)$ or the posterior odds $\Pr(H_{1(k,j)}|D)/\Pr(H_{0(k,j)}|D)$ to make decisions is mainly that when the intervals corresponding to the null and alternative hypotheses are not the same length, such as a point null hypothesis, the posterior probability or the posterior odds may not reflect what we really want to quantify. Suppose that there is a null hypothesis $H_{0(k,j)}:\theta_{(k,j)}=\theta_0$ and an alternative hypothesis $H_{1(k,j)}:\theta_{(k,j)}\neq\theta_0$. For posterior odds $\Pr(H_{1(k,j)}|D)/\Pr(H_{0(k,j)}|D)$, it is always going to get an infinity result which cannot used for decision-making. However, in terms of Bayes factor, the $\Pr(H_1|D)/\Pr(H_1)$ is approximately equal to 1. So we can get the Bayes factor by calculating $\Pr(H_0)/\Pr(H_0|D)$, i.e., the ratio of the prior density and the posterior density. Therefore, Bayes factors can increase the flexibility in formulating relevant hypotheses. On the other hand, although the the posterior probability or the posterior odds value can represent the strength of evidence in favour of the alternative hypothesis, the Bayes factor is a more commonly used Bayesian solution to the hypothesis testing problems.

Extension to adaptive enrichment design

Along the way of subgroup identification, IBIS can be extended to the following multi-stage adaptive enrichment design. Suppose there are a total of I analyses, including $I-1$ interim analyses and one final analysis. Patients with any biomarker status can be enrolled in the initial stage of the trial. In the $_{th}(i=1,\ldots,I-1)$ interim analysis, a go/no-go decision is made based on the Bayes factors, specifically as follows:

1. If $BF_{(k,j)} > BF^{(i)}_{E(k,j)}$, the investigational therapy is considered to be effective for subgroup (k,j);
2. If $BF_{(k,j)} \leq BF^{(i)}_{P(k,j)}$, the investigational therapy is considered to be ineffective for subgroup (k,j);
3. If $BF_{(k,j)}$ is between these two thresholds, the investigational therapy is considered promising for subgroup (k,j).

Thresholds $BF^{(i)}_{E(k,j)}$ and $BF^{(i)}_{P(k,j)}$ are important design parameters that need to be calibrated carefully, and the calibration strategy will be covered later. After the interim analysis, only promising subgroups will be enrolled in the next stage. The enrollment of the other two kinds of subgroups will be stopped early for efficacy or futility. The whole trial will be stopped early if there are no promising subgroups. Note that we can define $BF^{(i)}_{E(k,j)} = \infty(i = 1, \ldots, I-1)$ and thus prevent early stopping for efficacy. At the final analysis, if $BF_{(k,j)} > BF^{(I)}_{E(k,j)}$, the investigational therapy is considered to be effective for subgroup (k,j); otherwise, the investigational therapy is considered to be ineffective for subgroup (k,j). It is logistically and operationally intractable to change enrollment criteria too often in one clinical trial and a two-stage design is often recommended for the adaptive enrichment design. Figure 2 shows the schema of a two-stage adaptive enrichment design based on IBIS.

We also considered the monotonic relationship between drug efficacy and biomarker levels here. Specifically, once the investigational therapy is considered promising for subgroup $\left(\tilde{k}, \tilde{j}\right)$, it is also deemed promising for subgroups which were initially judged ineffective in subset $C : \left\{k \geq \tilde{k}, j \geq \tilde{j}\right\}$; once the investigational therapy is considered effective for subgroup $\left(\tilde{k}, \tilde{j}\right)$, it is also deemed effective for subgroups which were initially judged ineffective or promising in subset $C : \left\{k \geq \tilde{k}, j \geq \tilde{j}\right\}$.

Since the proposed adaptive enrichment design incorporates the strategy of early stopping for futility and efficacy, it may face challenges including type I error inflation and reduced statistical power. Therefore, it is critical to determine reasonable design parameters in the planning stage. Usually, two key metrics, family-wise error rate (FWER) and conjunctive power [29], are the main concern; these metrics are defined as follows:

$$FWER = \Pr\left(Reject\ at\ least\ one\ H_{0(k,j)} | H_{0(k,j)}\ is\ true\right)$$

$$Conjunctive\ Power = \Pr\left(Reject\ all\ H_{0(k,j)} | H_{1(k,j)}\ is\ true\right)$$

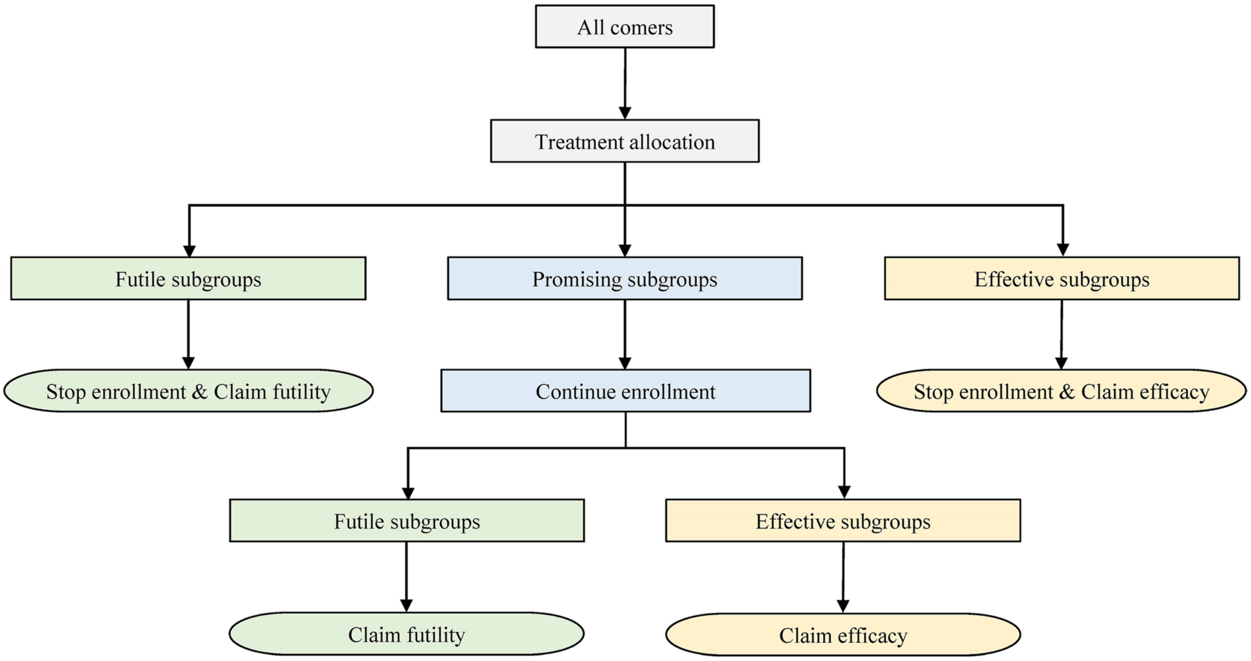

Fig. 2 Schema of a two-stage adaptive enrichment design based on IBIS

The general strategy of calibrating design parameters is to make the design achieve satisfactory FWER, conjunctive power and expected sample size in several typical scenarios by simulation.

Simulation study

Evaluation of the estimation performance

We conduct computer simulations to evaluate the estimation performance of IBIS. Consider a one-stage trial where the primary efficacy endpoint is the ratio of tumor size at 1 month after treatment to that at baseline. After transformation to the log scale, this ratio is assumed to be a normally distributed continuous variable. The smaller the log of the ratio, the greater the benefit to patients. To be consistent with the hypothesis testing framework described before, we add a minus sign to the transformed endpoint. Using statistical notation, if $Y_{(k,j)}$ denotes the ratio of tumor size for subgroup (k,j), then $-\log\left(Y_{(k,j)}\right) \sim N\left(\theta_{(k,j)}, \sigma^2_{(k,j)}\right)$. There are three and four levels for *Biomarker1* and *Biomarker2*, respectively, so the entire population can be divided into 12 subgroups. The minimum acceptable treatment effect $\theta_0 = 0$ and a clinically meaningful treatment effect is equal to 1. A total of eight scenarios are incorporated in the simulation, including the global null (scenario 1), global alternative (scenario 2), good nugget (scenario 3), bad nugget (scenario 4), mostly null (scenario 5), mostly alternative (scenario 6), half alternative (scenario 7) and linear (scenario 8) scenarios. The detailed treatment effect $\theta_{(k,j)}$ for each subgroup can be found in Table 2. It is assumed that all $\sigma^2_{(k,j)}$ equal 1, and the sample size of each subgroup is 10. The priors are set as follows: $\theta_g \sim N(1, 10^3)$, $\theta_{\mathtt{C}g} \sim N(0, 10^3)$, $\sigma^2_g \sim IG\left(10^{-3}, 10^{-3}\right)$ and $\sigma^2_{\mathtt{C}g} \sim IG\left(10^{-3}, 10^{-3}\right)$.

The metrics for evaluating the estimation performance include the mean squared error (MSE), bias and average width of the 95% equal-tailed credible interval for the posterior distribution of $\theta_{(k,j)}$. The MSE is defined as the average squared difference between the estimated values and the actual value of $\theta_{(k,j)}$. The bias is defined as the expected difference between the estimated values and the actual value of $\theta_{(k,j)}$. For simplification, we omit the subscript (k,j) when it does not cause ambiguity:

$$MSE = E\left\{\left(\hat{\theta} - \theta^*\right)^2\right\} \approx \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \left(\hat{\theta}_i - \theta^*\right)^2,$$

$$Bias = E\left(\hat{\theta} - \theta^*\right) \approx \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \left(\hat{\theta}_i - \theta^*\right),$$

where $\hat{\theta}$ and θ^* denote the estimated value and the actual value of the treatment effect, respectively. $\hat{\theta}_i$ denotes the estimated value of the treatment effect for the i_{th} simulated trial, where we use the posterior mean as the estimate. The total number of simulated trials, denoted as n_{sim}, is 10,000 here. The average width of the 95% equal-tailed credible interval for the posterior distribution is defined as follows:

Table 2 Eight scenarios of treatment effects for patients with different biomarker levels

	Biomarker2							
Level	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Biomarker1	*Scenario 1*				*Scenario 2*			
1	0	0	0	0	**1**	**1**	**1**	**1**
2	0	0	0	0	**1**	**1**	**1**	**1**
3	0	0	0	0	**1**	**1**	**1**	**1**
	Scenario 3				*Scenario 4*			
1	0	0	0	0	0	**1**	**1**	**1**
2	0	0	0	0	**1**	**1**	**1**	**1**
3	0	0	0	**1**	**1**	**1**	**1**	**1**
	Scenario 5				*Scenario 6*			
1	0	0	0	0	0	0	**1**	**1**
2	0	0	**1**	**1**	0	0	**1**	**1**
3	0	0	**1**	**1**	**1**	**1**	**1**	**1**
	Scenario 7				*Scenario 8*			
1	0	0	0	**1**	0	0.25	0.5	**1**
2	0	0	**1**	**1**	0.25	0.5	**1**	**1.25**
3	0	**1**	**1**	**1**	0.5	**1**	**1.25**	**1.5**

Subgroups with clinically meaningful treatment effect are shown in boldface

$$Width = E\{l(0.975,\theta) - l(0.025,\theta)\} \approx \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \left\{ l\left(0.975,\theta_i\right) - l\left(0.025,\theta_i\right) \right\}$$

where $l(q,\theta) = \{l : \Pr(\theta \leq l) = q\}$, which denotes a quantile function. θ_i represents the posterior distribution of the treatment effect for the i_{th} simulated trial.

Estimation methods used for comparison include independent analysis and BHM. In the independent analysis, the parameter estimate is the sample mean $\overline{Y}$. Analogous to the 95% credible interval described above, the precision of the estimate is measured by the width of the 95% confidence interval under the frequentist statistics, i.e., $\overline{Y} \pm t_{1-\alpha/2}(n-1) \times \frac{s}{\sqrt{n}}$, where $t_{1-\alpha/2}(n-1)$ is the $1-\alpha/2$ quantile of the t distribution with degrees of freedom equal to $n-1$ ($\alpha = 0.05$ here). The sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$ and n is the sample size of the subgroup under evaluation. In BHM, all $\theta_{(k,j)}$ are treated as exchangeable, i.e., borrowing information is conducted across all subgroups. We force $\theta_{(k,j)} \sim N\left(\widetilde{\theta}, \widetilde{\sigma}^2\right)$ with the priors of $\widetilde{\theta}$ and $\widetilde{\sigma}^2$ equal to $N(0,10^3)$ and $IG(10^{-3},10^{-3})$, respectively.

Evaluation of the operating characteristics

The operating characteristics of IBIS on subgroup identification are also evaluated by computer simulation. The simulation settings are the same as those of the above evaluation of the estimation performance. The evaluation metrics include FWER and conjunctive power. The methods used for comparison include BHM, independent analysis and another frequentist subgroup identification method, denoted as 'Freq' here. Suppose there is only one subgroup, then with a null hypothesis $H_0 : \theta \leq 0$, an expected treatment effect to be 1 and a standard deviation to be 1, a study with 10 participants has approximately 90% power to reject the null hypothesis using t-test at the 5% significance level. Therefore, it can be expected that if the FWER is controlled when there are 12 subgroups (each with 10 participants), then the conjunctive power will decrease a lot, i.e., resulting in poor performance on subgroup identification. In this case, we presume that using IBIS to borrow information across subgroups may improve the accuracy of subgroup identification.

The decision-making process of BHM is the same as that of IBIS. With the vague prior distributions we set for parameters in IBIS and BHM method, there is no preference for the null hypothesis or the alternative hypothesis with regard to any one subgroup, i.e., $\Pr(H_{0(k,j)}) = \Pr(H_{1(k,j)}) = 0.5$, in the simulation of BHM and IBIS. The decision-making process of independent analysis is nearly the same as that of the IBIS, and the only change is to replace the Bayes factor with a decision-making based on a t-test. If the following inequation

$$t_{(k,j)} = \frac{\sqrt{n_{(k,j)}}\left(\overline{Y}_{(k,j)} - \theta_0\right)}{s_{(k,j)}} > t_{E(k,j)}$$

is satisfied, then the investigational combination therapy is considered effective for subgroup (k,j). $n_{(k,j)}$, $\overline{Y}_{(k,j)}$ and $s_{(k,j)}$ denote the sample size, sample mean and sample standard deviation of subgroup (k,j), respectively.

In the Freq method, decision-making is based on Lai et al. [24], who first divide the subgroups into two subsets and then make inferences for the subsets separately. Different from using Jensen–Shannon divergence, the high-efficacy subset is determined by selecting the largest t-test statistic, which is defined as follows:

$$C_H = \underset{\Pi_g}{\operatorname{argmax}} \left\{ t_{\Pi_g}, g = 1, \ldots, G-1 \right\}.$$

where t_{Π_g} is the t-test statistic for subset Π_g. Then, t tests are performed for C_H and its complementary set, respectively. If the test statistic is greater than a pre-specified threshold, the investigational combination therapy is judged to be effective for that subset; otherwise, it will be considered ineffective.

Let the test thresholds for each subgroup in the same method be equal to facilitate evaluation, although adjustments can be made in practice considering the anticipated efficacy and prevalence of each subgroup. In the Freq method, the test thresholds for high-efficacy and low-efficacy subsets are equal. To make the four methods comparable, we calibrate the thresholds to enable their FWER in the global null scenario (scenario 1) to be controlled within 0.1 or 0.05. Specifically, we performed a series of simulation studies for each method under the null scenario over a grid of the thresholds (i.e., the thresholds of Bayes factors for IBIS and BHM; the thresholds of t-test statistics for independent analysis and the Freq method). Then, for each method, the minimum threshold with simulated FWER less than or equal to 0.1 (or 0.05) would be determined as the design parameter.

Regarding the adaptive enrichment design, the parameter calibration is much more complicated. A general strategy is to make the design achieve an acceptable FWER, conjunctive power and expected sample size in some typical scenarios by simulation with a limited maximum sample size. This may be computationally expensive, which is the sacrifice for adaptability and flexibility.

Consider a two-stage trial with a sample size of five for each subgroup in the first stage. If the investigational therapy is promising for a subgroup, five more patients will be enrolled in the second stage. To preliminarily conduct the parameter calibration, let the decision thresholds for each subgroup be the same, i.e., $BF_{E(k,j)}^{(1)} = BF_{E(k,j)}^{(2)} = BF_E$ and $BF_{P(k,j)}^{(1)} = BF_P$. Specifying loose decision thresholds may increase power for high-efficacy subgroups, but may also inflate type I errors for low-efficacy subgroups. Therefore, we define a decision score function to comprehensively measure FWER, conjunctive power and expected sample size:

$$Score = Power - \beta_1 \cdot FWER + \beta_2 \cdot \left(1 - \frac{EN}{N_{max}}\right)$$

where EN denotes the expected sample size and N_{max} denotes the maximum sample size, which is equal to 120 here. The above function shows that the loss of increasing one unit FWER can offset the benefit of increasing β_1 unit conjunctive power; the loss of increasing one unit EN can offset the benefit of increasing β_2/N_{max} unit conjunctive power. For example, setting $\beta_1 = 1$ and $\beta_2 = 0.5$ means that a 1% increase of the FWER is enough to offset a 1% increase of power. At the same time, an increase of one expected sample size could offset either a 0.4% (i.e., β_2/N_{max}) increase of power or a 0.4% decrease of the FWER. Such a decision score function can be interpreted as the tradeoff between FWER, conjunctive power and expected sample size. If a large value is set for β_1, the design will favor a stricter control for FWER. The larger the β_2, the more inclined we are to control the cost of the current trial by reducing the expected sample size. How to choose β_1 and β_2 is a key and difficult problem, and is mainly determined by the potential losses caused by type I and type II errors. The type I error results in the loss of future clinical research with ineffective investigational therapeutics, while the type II error indicates the unavailability of effective treatments for some patients and the loss of marketing revenues. How do sponsors view these potential losses and the cost of the current trial will play a decisive role in determining β_1 and β_2. With $\beta_1 = 1$ and $\beta_2 = 0.5$, scenario 8 is taken as an example here to explain how to obtain design parameters for such an adaptive enrichment design.

Results

Accuracy and precision

The simulation results of the estimation performance are shown in Figs. 3, 4 and 5 and could reflect the accuracy and precision of IBIS. It can be seen that IBIS has the lowest MSE overall, with only few scenarios where the MSEs for few subgroups are slightly higher than that of independent analysis (Fig. 3). Specifically, in subgroup (3,4) of scenario 3 (the good nugget), the MSE of IBIS is relatively high. This is because the treatment effect for this subgroup is so different from other subgroups, and IBIS tends to analyze it alone or just combine it with few adjacent subgroups, resulting in a relatively large variance of the estimate, which can be demonstrated by its wide 95% CI as well (Fig. 5). In this nugget scenario, BHM inevitably produces a large estimation bias (Fig. 4), which in turn leads to a much higher MSE. Similar results also arise in subgroup (1,1) of scenario 4. The estimate of the independent analysis for each subgroup is unbiased (Fig. 4), so the MSE is consistent with the variance of the sample mean, i.e., standard error, which is a constant of 0.1 (Fig. 3). The MSE of BHM is the smallest in homogeneous scenarios (scenarios 1 and 2), but BHM is not robust in heterogeneous scenarios, especially in scenarios 3 and 4, where the MSEs for nugget subgroups exceed 0.4. This is already more than four times that of the other two methods. The bias of BHM is also large in heterogeneous scenarios, and the absolute values of the bias for some subgroups are higher than 0.2 (Fig. 4), while the IBIS is much more robust. Notably, although BHM borrows information across all subgroups, the 95% CIs for its estimates are generally wider in heterogeneous scenarios than that of IBIS. This is because the variance across subgroups $\overset{\sim}{\sigma}^2$ estimated by BHM is large due to heterogeneous subgroups, resulting in limited information borrowing. In contrast, the subgroup division in IBIS makes the similarity within each subset higher and thus leads to more information borrowing.

In summary, from the three metrics (MSE, bias and 95% CI width), the estimation performance of the IBIS is superior and robust under the pre-specified scenarios. The proposed method is especially suitable for scenarios where subgroups should be divided into two subsets.

Operating characteristics of subgroup identification

The simulation results of the operating characteristics are presented in Table 3. In the global alternative scenario (scenario 2), BHM can identify all subgroups correctly with a 100% probability, indicating a strong borrowing strength. The Freq method achieves nearly 90% conjunctive power (hereinafter referred to as power), which reflects the advantage of partial combination. The power

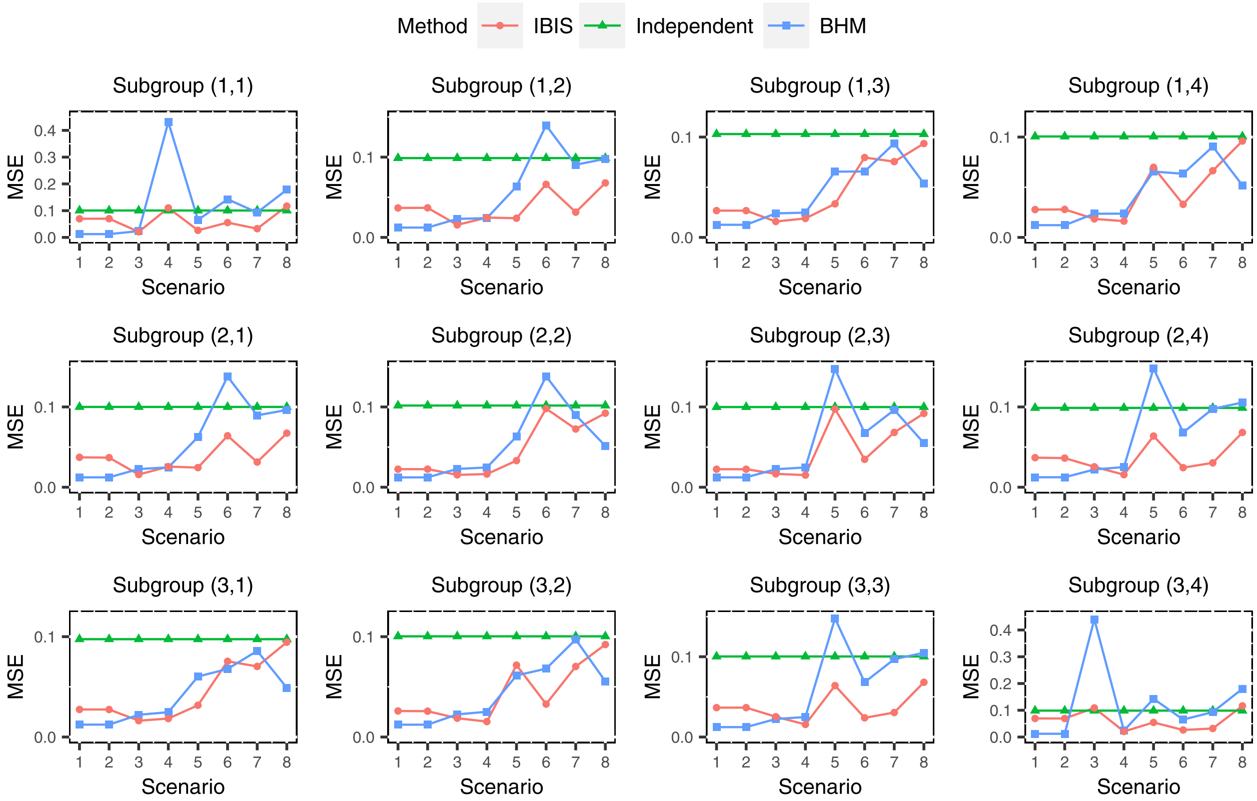

Fig. 3 Simulated mean squared error (MSE) of treatment effect estimates for 12 subgroups under eight scenarios using 'IBIS', 'Independent Analysis' and 'BHM'

produced by IBIS is slightly lower than that of the Freq method but is much higher than that of the independent analysis.

When the subgroups are heterogeneous (scenarios 3–8), the IBIS outperforms the other three methods, evidenced by a better balance between FWER and power. Compared with independent analysis, the power of the IBIS is much higher. When the investigational therapy is effective for several subgroups (scenarios 4, 6, 7, and 8), it is difficult for independent analysis to identify all alternative subgroups correctly at the same time. The FWER of IBIS is slightly higher than that of independent analysis in most scenarios, except in scenario 3, where information borrowing across low-efficacy subgroups reduces the type I error rates. However, such type I error inflation is relatively limited, for example, approximately 10–20% in scenarios 5–7. At this time, the number of null subgroups is small, and the borrowing strength between subgroups is not as strong as that in scenarios 1–4. Therefore, there are some misjudgments for null subgroups. There is a consensus that controlling type I error rates is critical in clinical trials. But it should also be recognized that if the investigational drug is effective for a large portion of the population (e.g., scenario 5), it may be inappropriate to borrow no information at all. In scenarios 5–7, although FWER inflates, the proposed IBIS design misjudge only one or two null subgroups in most cases (see section A of the Additional file 1 for the the detailed simulated probabilities of misjudging null and alternative subgroups in each scenario). Taking scenario 5 in simulation 1 as an example, there are a total of 8 subgroups with low efficacy to treatment, and the family-wise type I error rate is 15.86%, higher than 10%. But most of the error cases are misjudging one (9.52%) or two subgroups (4.54%). In these cases, the investigational drug combination are effective in 2/3 of the identified subgroups and are ineffective in 1/3 of the identified subgroups. Therefore, given that some marketed drugs are not effective for all patients, this kind of type I error inflation is acceptable to some extent. We believe that strictly controlling the type I error of one or two subgroups at the expense of power gains of most subgroups is not desirable for subgroup identification in an exploratory trial.

The power produced by BHM is high in most scenarios, except in scenario 3, where there is only one good nugget subgroup. However, BHM also tends to overestimate the efficacy of investigational therapy for low-efficacy subgroups, resulting in unacceptable FWERs.

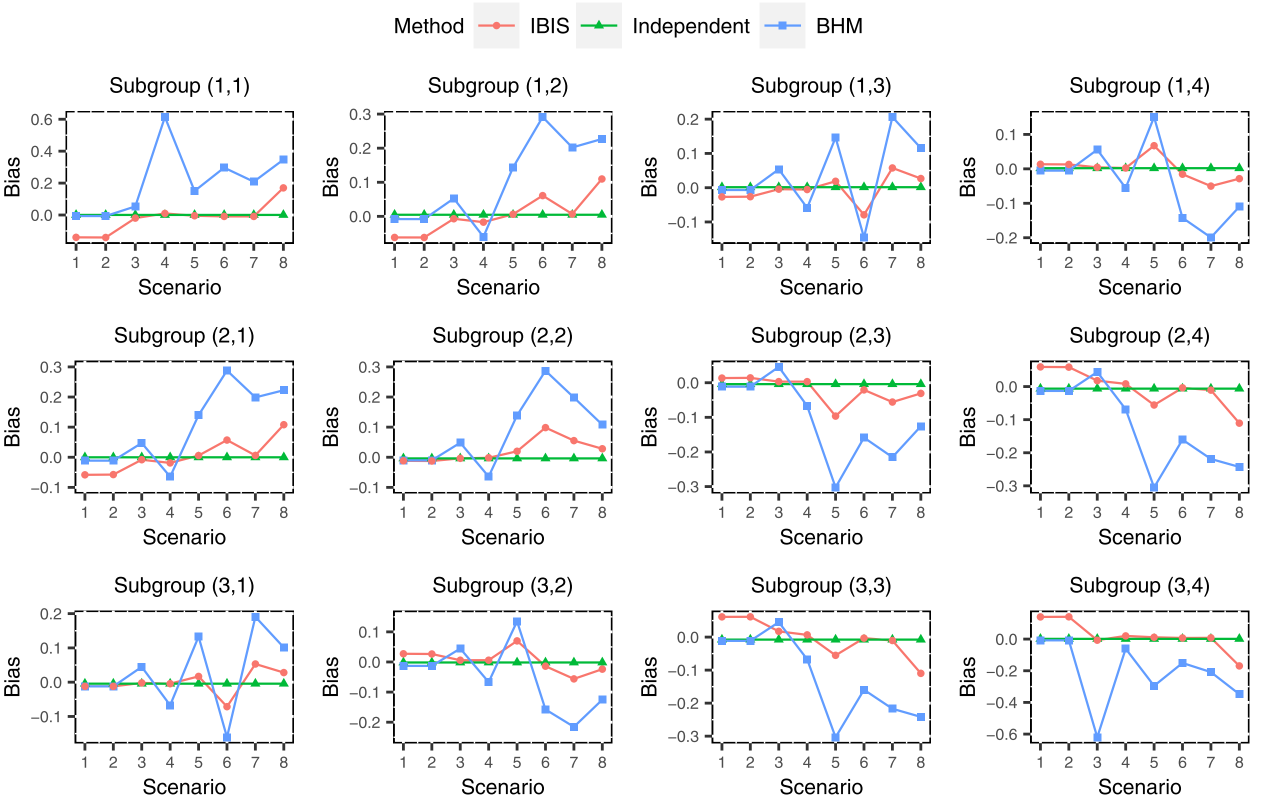

Fig. 4 Simulated bias of treatment effect estimates for 12 subgroups under eight scenarios using 'IBIS', 'Independent Analysis' and 'BHM'

The operating characteristics of the Freq method are relatively robust compared with independent analysis and BHM, but are inferior to the proposed IBIS in general. This is due to its strategy of partial combination. For example, if Freq mistakenly classifies an alternative subgroup into a low-efficacy subset, it is likely to accept the null hypothesis in subsequent inferences and conclude that the investigational therapy is ineffective for all subgroups in the low-efficacy subset. In contrast, the IBIS method, although also making subgroup divisions, only borrows information within each subset, and the inferences are still based on each subgroup. Therefore, the probabilities of making such incorrect decisions are smaller in most scenarios.

Additional sensitivity analyses evaluating the prior settings for the shrinkage parameters are performed and the results are shown in section B of the Additional file 1. The results show that the used inverse-gamma priors are robust in our simulation study. But it should be noted that there may be no optimal choice for the prior distributions. The priors we used here may be not applicable in other studies. The appropriate prior distributions for a particular trial must be determined by the cooperation of clinicians and statisticians at the time of trial design. See section B of the Additional file 1 for more details and remarks.

An example of how to obtain design parameters for a multi-stage adaptive enrichment design

Regarding the adaptive enrichment design, we evaluate the performance of IBIS under different values of BF_E and BF_P for scenario 8 and draw corresponding heatmaps for FWER, conjunctive power, expected sample size and decision score (Fig. 6). The corresponding heatmaps for other scenarios are shown in section C of the Additional file 1. With the help of such heatmaps, decision-makers can more intuitively observe the changes of the above metrics along with the decision thresholds and thus make trade-offs between these metrics.

It should be noted that the restriction $BF^{(1)}_{E(k,j)} = BF^{(2)}_{E(k,j)} = BF_E$ and $BF^{(1)}_{P(k,j)} = BF_P$ is set mainly because we want to first have a general overview of the operating characteristics with different combinations of decision thresholds. After determining the range of decision thresholds that meet the requirements (in the example above, when BF_P is from 5.0 to 12.5 and BF_E is from 75 to 150, the decision score is relatively high), we can further adjust such thresholds

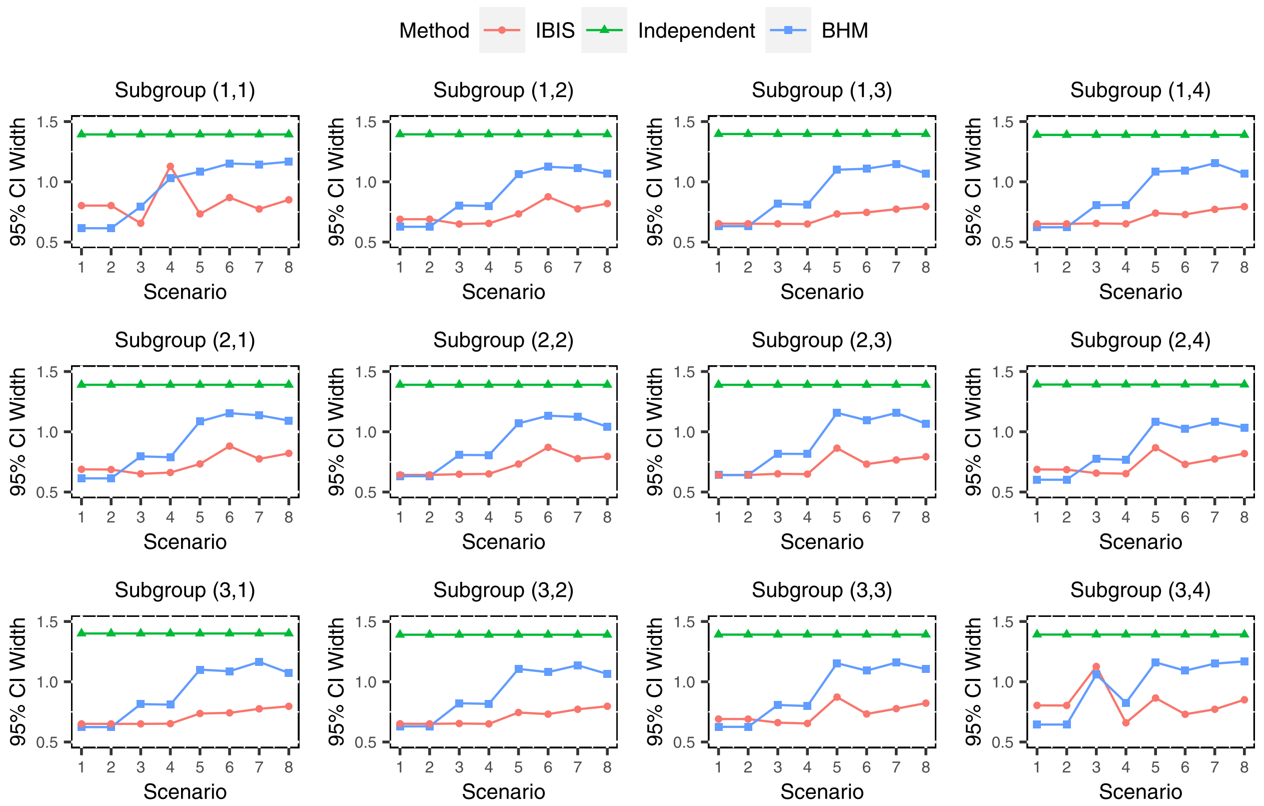

Fig. 5 Simulated average width of 95% CI of treatment effect estimates for 12 subgroups under eight scenarios using 'IBIS', 'Independent Analysis' and 'BHM'. CI means credible interval in 'IBIS' and 'BHM' method, while it means confidence interval in 'Independent Analysis'

according to the anticipated efficacy and prevalence of each subgroup.

In addition to FWER, conjunctive power and expected sample size, some other metrics can also be considered, such as the disjunctive power, which is defined as follows:

$$\textit{Disjunctive Power} = \Pr\left(\textit{Reject at least one } H_{0(k,j)} | H_{1(k,j)} \textit{ is true}\right)$$

This metric makes sense, especially when the sample size is limited. On the other hand, the operating characteristics of the design in other scenarios should be also taken into account (see section C of the Additional file 1). In conclusion, the composition of the decision score function and the weights of different metrics need to be discussed sufficiently by researchers, sponsors and biostatisticians. We should set up reasonable scenarios based on existing medical research data case-by-case and take into account the requirements of various stakeholders when calibrating parameters of multi-stage adaptive enrichment designs.

Discussion

Compared with monotherapy, targeted combination therapy leads to an increasing number of subgroups and complicated biomarker-based efficacy profiles. Therefore, we propose a statistical tool called IBIS in this article and incorporate it into the adaptive enrichment design framework. The IBIS contains three main elements: subgroup division, efficacy evaluation and subgroup identification. We first enumerate all possible subgroup divisions based on biomarker levels. Then, Jensen–Shannon divergence is used to distinguish high-efficacy and low-efficacy subgroups, and BHM is employed to borrow information within these two subsets for efficacy evaluation. Regarding subgroup identification, a hypothesis testing framework based on Bayes factors, which also plays a key role in go/no-go decisions and enriching specific subgroups in subsequent stage of trial, is constructed. Simulation studies show that, compared with some traditional methods, our proposed IBIS has superior and robust operating characteristics, and has the potential to be a useful tool for subgroup identification and population enrichment in clinical trials of targeted combination therapy.

As the combination of two targeted agents is a promising therapeutic approach at present, we take trial designs that consider two ordinal biomarkers as examples in this paper. The proposed design can be naturally extended for subgroup identification where multiple biomarkers are incorporated. We can still enumerate all possible

Table 3 Operating characteristics of four methods under eight scenarios

Method	Simulation 1		Simulation 2	
	FWER	Conjunctive Power	FWER	Conjunctive Power
	Scenario 1		*Scenario* 1	
IBIS	0.0971	/	0.0490	/
Independent	0.1000		0.0496	
BHM	0.0991		0.0500	
Freq	0.0985		0.0499	
	Scenario 2		*Scenario* 2	
IBIS	/	0.8602	/	0.8092
Independent		0.6089		0.4664
BHM		1.0000		1.0000
Freq		0.8961		0.8787
	Scenario 3		*Scenario* 3	
IBIS	0.0455	0.8984	0.0265	0.8193
Independent	0.0916	0.6447	0.0449	0.4910
BHM	0.3093	0.6027	0.1760	0.4587
Freq	0.2149	0.8609	0.1717	0.7582
	Scenario 4		*Scenario* 4	
IBIS	0.0578	0.9541	0.0399	0.9341
Independent	0.0079	0.3748	0.0045	0.2251
BHM	0.7718	0.9980	0.6805	0.9964
Freq	0.1072	0.8101	0.0866	0.7900
	Scenario 5		*Scenario* 5	
IBIS	0.1586	0.8774	0.1106	0.8393
Independent	0.0693	0.6200	0.0343	0.4743
BHM	0.6445	0.9768	0.4459	0.9407
Freq	0.1772	0.8480	0.1647	0.8375
	Scenario 6		*Scenario* 6	
IBIS	0.1178	0.8092	0.0838	0.7779
Independent	0.0349	0.3766	0.0175	0.2201
BHM	0.7028	0.9891	0.5440	0.9686
Freq	0.1153	0.7656	0.0973	0.7508
	Scenario 7		*Scenario* 7	
IBIS	0.1877	0.8108	0.1508	0.7742
Independent	0.0525	0.2507	0.0248	0.1185
BHM	0.6700	0.9688	0.4945	0.9184
Freq	0.1860	0.7138	0.1712	0.7012
	Scenario 8		*Scenario* 8	
IBIS	0.0843	0.8617	0.0386	0.7637
Independent	0.0079	0.4113	0.0045	0.2171
BHM	0.4294	0.9982	0.3169	0.9937
Freq	0.3961	0.9030	0.3030	0.8597

Simulations 1 & 2 correspond to different degrees of control over FWER in the global null scenario (scenario 1)

For global null scenario (scenario 1), calculating the conjunctive power makes no sense. Similarly, for global alternative scenario (scenario 2), calculating the FWER makes no sense

subgroup divisions, find the optimal division, and then use the BHM to borrow information within each subset. In terms of continuous biomarkers, if we cannot pre-define subgroups and the study objective is to find arbitrary biomarker cutoffs, the proposed design cannot work. If the subgroups can be pre-defined, e.g., be divided by quartiles, then the proposed design can still be applied.

Another feasible monitoring method for interim analysis is to calculate a predictive probability that $BF_{(k,j)}$ exceed a success threshold at the end of the trial. Each time an interim analysis is performed, this predictive probability is compared with pre-specified cutoffs to decide the target population in the subsequent stage. Some researches, e.g., Lee and Liu (2008) [30], found that comparable operating characteristics can be obtained by taking both the predictive probability and the posterior probability approaches. This conclusion should also be generalized to our proposed design as the predictive probability approach does not include more patients' data compared with the posterior probability approach. One advantage of the predictive probability approach is that it can give the probability of finally obtaining a positive result at interim analysis. However, it also requires more computational expense when predicting future data.

Some points need to be paid attention to when applying IBIS. For example, when the number of subgroups is small, it may be inappropriate to directly use the inverse gamma distribution with small shape and scale parameters as a prior for the variances across subgroups [31]. Some weakly informative priors can be applied instead. The prevalence of each subgroup also influences trial design and conduction. Enrollment of patients may be difficult when the prevalence of certain subgroups is low, and we cannot accumulate enough evidence to demonstrate how well the investigational therapy works on those subgroups. On the other hand, the treatment parameters of subgroups with smaller sample sizes tend to have stronger shrinkage in BHM [32], potentially causing misleading interpretations of trial results. When faced with such a problem, our suggestion is to first consider whether the anticipated efficacy for the rare subgroups is close to that of the adjacent subgroups based on clinical experience. If they are close enough, pooled analysis is suggested; if it is uncertain whether they are close, an independent exploratory analysis for the rare subgroups is recommended rather than a BHM-based analysis.

Regarding the choice between single-arm trial and RCT, we must admit that RCT is a better choice when cost, ethics and some other aspects allow, due to its advantages of eliminating selection bias and minimizing confoundings. However, RCT also has the limitations of high cost, long period and difficult implementation, which may not be appropriate if such a subgroup identification or enrichment trial is positioned as an exploratory

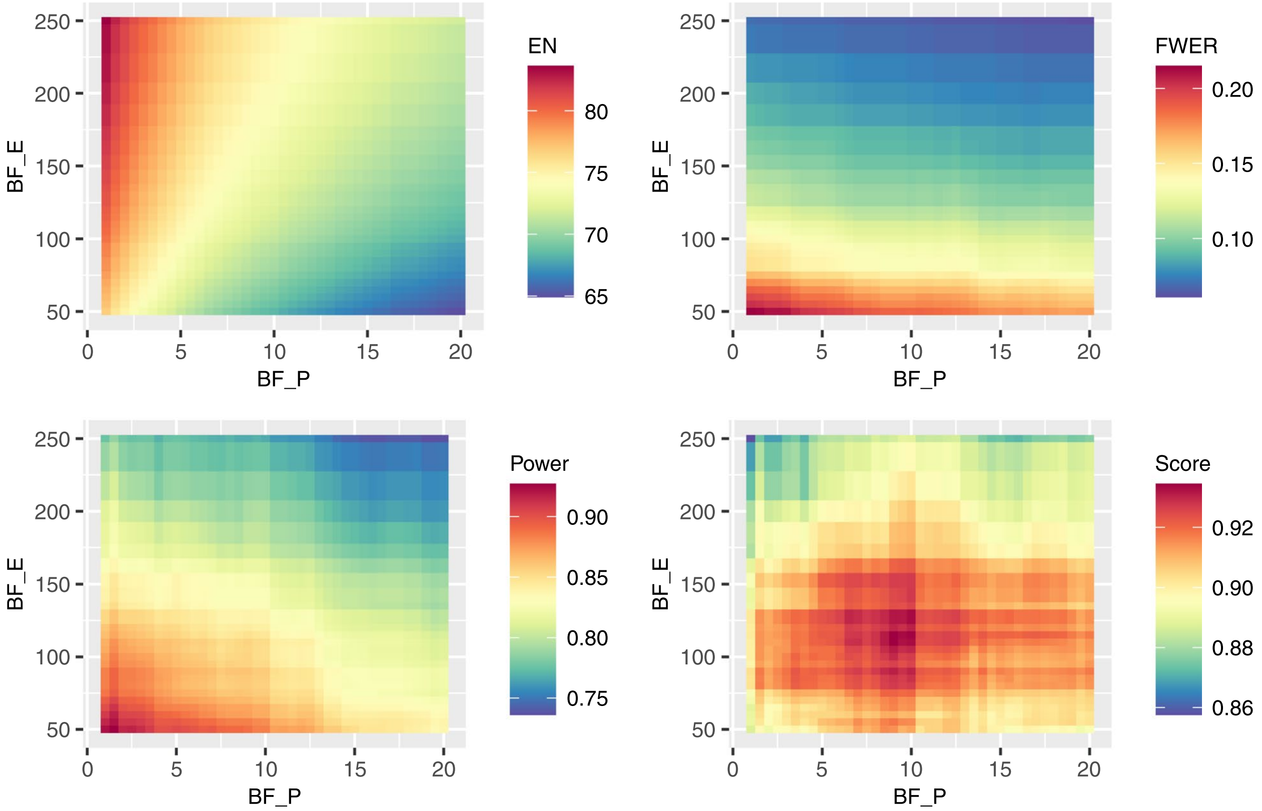

Fig. 6 Operating characteristics of the two-stage adaptive enrichment design with varying decision thresholds under scenario 8. EN, expected sample size; FWER, family-wise type I error rate; Power, conjunctive power; Score, values of decision score

trial. One strategy that deserves further exploration is the seamless transition design from open-label single-arm to randomized double-arm clinical trials.

There are also some other areas where this research can be extended. For example, the sample size re-estimation based on the Bayes factors can be considered in the interim analysis. Historical information on monotherapy can be incorporated to determine the priors in Bayes factors. There is also a class of adaptive enrichment designs that take into account situations where subgroups are not pre-defined [33–42]. Therefore, in addition to multilevel biomarkers, we can further consider how to determine the cutoffs of continuous biomarkers, or include other important covariates in the model, such as age, Eastern Cooperative Oncology Group Performance Status (ECOG PS), key laboratory testing results, etc., which is more in line with the concept of precision medicine.

Conclusions

IBIS has superior and robust operating characteristics in terms of subgroup identification and population enrichment. It has the potential to be a useful tool for biomarker-based subgroup identification in clinical trials of targeted antitumor combination therapy.

Abbreviations

BHM	Bayesian hierarchical model
ADC	antibody-drug conjugate
RCT	randomized controlled trial
FWER	family-wise error rate
MSE	mean squared error
ECOG PS	Eastern Cooperative Oncology Group Performance Status

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-01877-w.

Additional file 1. Supplemental simulation results.

Acknowledgements

We would like to thank AJE (www.aje.cn) for English language editing.

Authors' contributions

XC, JZ and FY contributed to the conception and design of the work. XC finished the code. XC, LJ and FY interpreted the results. XC and FY drafted the manuscript, and all authors read, edited, and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [No.81973145, No. 82273735] and Key R&D Program of Jiangsu Province (Social Development) (BE2020694).

Availability of data and materials

The R code for simulation during the current study are available in the github repository, https://github.com/cccc633/IBIS.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, Nanjing, China.

Received: 13 October 2022 Accepted: 24 February 2023
Published online: 20 March 2023

References

1. Cheng AL, Qin S, Ikeda M, Galle PR, Ducreux M, Kim TY, et al. Updated efficacy and safety data from IMbrave150: Atezolizumab plus bevacizumab vs. sorafenib for unresectable hepatocellular carcinoma. J Hepatol. 2022;76(4):862–73.
2. Hellmann MD, Paz-Ares L, Bernabe Caro R, Zurawski B, Kim SW, Carcereny Costa E, et al. Nivolumab plus Ipilimumab in advanced non-small-cell lung Cancer. N Engl J Med. 2019;381(21):2020–31.
3. Loi S, Giobbie-Hurder A, Gombos A, Bachelot T, Hui R, Curigliano G, et al. Pembrolizumab plus trastuzumab in trastuzumab-resistant, advanced, HER2-positive breast cancer (PANACEA): a single-arm, multicentre, phase 1b-2 trial. Lancet Oncol. 2019;20(3):371–82.
4. Oldenhuis CN, Oosting SF, Gietema JA, de Vries EG. Prognostic versus predictive value of biomarkers in oncology. Eur J Cancer. 2008;44(7):946–53.
5. Borghaei H, Besse B, Bardia A, Mazieres J, Popat S, Augustine B, et al. Trastuzumab deruxtecan (T-DXd; DS-8201) in combination with pembrolizumab in patients with advanced/metastatic breast or non-small cell lung cancer (NSCLC): a phase Ib, multicenter, study. J Clin Oncol 2020;38(15):suppl.TPS1100.
6. Bayat Mokhtari R, Homayouni TS, Baluch N, Morgatskaya E, Kumar S, Das B, et al. Combination therapy in combating cancer. Oncotarget. 2017;8(23):38022–43.
7. Choong GM, Cullen GD, O'Sullivan CC. Evolving standards of care and new challenges in the management of HER2-positive breast cancer. CA Cancer J Clin. 2020;70(5):355–74.
8. Denkert C, Seither F, Schneeweiss A, Link T, Blohmer JU, Just M, et al. Clinical and molecular characteristics of HER2-low-positive breast cancer: pooled analysis of individual patient data from four prospective, neoadjuvant clinical trials. Lancet Oncol. 2021;22(8):1151–61.
9. Takegawa N, Tsurutani J, Kawakami H, Yonesaka K, Kato R, Haratani K, et al. [fam-] trastuzumab deruxtecan, antitumor activity is dependent on HER2 expression level rather than on HER2 amplification. Int J Cancer. 2019;145(12):3414–24.
10. Modi S, Park H, Murthy RK, Iwata H, Tamura K, Tsurutani J, et al. Antitumor activity and safety of Trastuzumab Deruxtecan in patients with HER2-low-expressing advanced breast Cancer: results from a phase Ib study. J Clin Oncol. 2020;38(17):1887–96.
11. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. Stat Med. 2009;28(10):1445–63.
12. Gotte H, Donica M, Mordenti G. Improving probabilities of correct interim decision in population enrichment designs. J Biopharm Stat. 2015;25(5):1020–38.
13. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharm Stat. 2011;10(4):347–56.
14. Liu A, Liu C, Li Q, Yu KF, Yuan VW. A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. Clin Trials. 2010;7(5):537–45.
15. Mehta C, Schafer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. Stat Med. 2014;33(26):4515–31.
16. Rosenblum M, Luber B, Thompson RE, Hanley D. Group sequential designs with prospectively planned rules for subpopulation enrichment. Stat Med. 2016;35(21):3776–91.
17. Rosenblum M, van der Laan MJ. Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. Biometrika. 2011;98(4):845–60.
18. Sinha AK, Moye L III, Piller LB, Yamal JM, Barcenas CH, Lin J, et al. Adaptive group-sequential design with population enrichment in phase 3 randomized controlled trials with two binary co-primary endpoints. Stat Med. 2019;38(21):3985–96.
19. Uozumi R, Hamada C. Interim decision-making strategies in adaptive designs for population selection using time-to-event endpoints. J Biopharm Stat. 2017;27(1):84–100.
20. Wang SJ, O'Neill RT, Hung HM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat. 2007;6(3):227–44.
21. Wu LW, Li Q, Liu MY, Lin JC. Incorporating surrogate information for adaptive subgroup enrichment design with sample size re-estimation. Stat Biopharm Res. 2022;14(4):493–504.
22. Chiu YD, Koenig F, Posch M, Jaki T. Design and estimation in clinical trials with subpopulation selection. Stat Med. 2018;37(29):4335–52.
23. Lai TL, Lavori PW, Liao OY. Adaptive choice of patient subgroup for comparing two treatments. Contemp Clin Trials. 2014;39(2):191–200.
24. Lai TL, Lavori PW, Tsang KW. Adaptive enrichment designs for confirmatory trials. Stat Med. 2019;38(4):613–24.
25. Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. Stat Med. 2013;32(16):2695–714.
26. Wang SJ, Hung HM, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. Biom J. 2009;51(2):358–74.
27. Fuglede B, Topsoe F. Jensen-Shannon divergence and Hilbert space embedding. International Symposium onInformation Theory, 2004. 2004;ISIT 2004. Proceedings:31.
28. Kailath T. The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans Commun Technol. 1967;15(1):52–60.
29. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. BMC Med Res Methodol. 2019;19(1):129.
30. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. Clin Trials. 2008;5(2):93–106.
31. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on an article by Browne and Draper). Bayesian Anal. 2006;1(3):515–33.
32. Berry SM, Broglio KR, Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. Clin Trials. 2013;10(5):720–34.
33. Diao G, Dong J, Zeng D, Ke C, Rong A, Ibrahim JG. Biomarker threshold adaptive designs for survival endpoints. J Biopharm Stat. 2018;28(6):1038–54.
34. Hui J, Guo W. Optimal biomarker cutoff identification and validation. Stat Biosci. 2022;14:352–62.
35. Johnston SE, Lipkovich I, Dmitrienko A, Zhao YD. A two-stage adaptive clinical trial design with data-driven subgroup identification at interim analysis. Pharm Stat. 2022;21(5):1090–108.
36. Li J, Zhao L, Tian L, Cai T, Claggett B, Callegaro A, et al. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. Biometrics. 2016;72(3):877–87.
37. Park Y, Liu S. A randomized group sequential enrichment design for immunotherapy and targeted therapy. Contemp Clin Trials. 2022;116:106742.
38. Simon N, Simon R. Adaptive enrichment designs for clinical trials. Biostatistics. 2013;14(4):613–25.
39. Simon N, Simon R. Using Bayesian modeling in frequentist adaptive enrichment designs. Biostatistics. 2018;19(1):27–41.

40. Spencer AV, Harbron C, Mander A, Wason J, Peers I. An adaptive design for updating the threshold value of a continuous biomarker. Stat Med. 2016;35(27):4909–23.
41. Xu Y, Constantine F, Yuan Y, Pritchett YL. ASIED: a Bayesian adaptive subgroup-identification enrichment design. J Biopharm Stat. 2020;30(4):623–38.
42. Zhang Z, Chen R, Soon G, Zhang H. Treatment evaluation for a data-driven subgroup in adaptive enrichment designs of clinical trials. Stat Med. 2018;37(1):1–11.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.