Statistical Papers (2023) 64:1305–1328
https://doi.org/10.1007/s00362-023-01433-0

REGULAR ARTICLE

New insights into adaptive enrichment designs

Alessandro Baldi Antognini[1] · **Rosamarie Frieri**[1] · **Maroussa Zagoraiou**[1]

Received: 31 December 2022 / Revised: 28 February 2023 / Published online: 31 March 2023

Abstract
The transition towards personalized medicine is happening and the new experimental framework is raising several challenges, from a clinical, ethical, logistical, regulatory, and statistical perspective. To face these challenges, innovative study designs with increasing complexity have been proposed. In particular, adaptive enrichment designs are becoming more attractive for their flexibility. However, these procedures rely on an increasing number of parameters that are unknown at the planning stage of the clinical trial, so the study design requires particular care. This review is dedicated to adaptive enrichment studies with a focus on design aspects. While many papers deal with methods for the analysis, the sample size determination and the optimal allocation problem have been overlooked. We discuss the multiple aspects involved in adaptive enrichment designs that contribute to their advantages and disadvantages. The decision-making process of whether or not it is worth enriching should be driven by clinical and ethical considerations as well as scientific and statistical concerns.

Keywords Continuous biomarker · Personalized medicine · Predictive biomarker · Stratified medicine · Subgroup identification

1 Introduction

Randomized clinical trials are the gold standard for the evaluation of new drugs. Traditionally, later-phase trials are designed to enroll a large group of patients with the intention to treat a broad population. The underlying assumption is that the effect of the treatment is homogeneous across the diseased patients, indeed these trials are aimed

✉ Rosamarie Frieri
rosamarie.frieri2@unibo.it

Alessandro Baldi Antognini
a.baldi@unibo.it

Maroussa Zagoraiou
maroussa.zagoraiou@unibo.it

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

at estimating and testing the average treatment effect in the overall population. Then, eventually, post-hoc subset analyses are performed to discover those who particularly benefit from the drug.

Unfortunately, in several situations clinical trials carried out on an unselected population may not be suitable. For instance, since anticancer agents are molecularly targeted therapies, only a subgroup of patients might benefit (Freidlin and Korn 2014). Thanks to the increasing understanding of disease biology and the mechanism of action of new agents, it is now consolidating the idea that a disease is not generally homogeneous. Then, pharmaceutical developers are moving away from the idea that all patients can be successfully treated with the same therapy. Moreover, advances in biomedicine are leading to an increasing knowledge of the reasons for patients' heterogeneity in response to treatments, for instance, due to immunological, molecular, cellular, or genetic differences (Wang et al. 2009; Stallard et al. 2014). Thus, recently, almost all branches of medicine are moving towards stratified and personalized medicine, but a special interest is in the oncology area, also for minimizing toxicity (Simon and Simon 2013; Stallard et al. 2014; Maitournam and Simon 2005). In particular, stratified medicine is aimed at using different treatments for different subgroups of the patient population; the extreme is personalized medicine, in which patients receive individually tailored treatment regimes.

Reasons for targeted clinical trials. In general, the motivations for a targeted clinical trial may be manifold and not mutually exclusive: some patients may be not or poorly compliant, some patients may introduce too much variability in the study (a heterogeneous population may increase variability in the response not drug-related), and other patients may be inherently unsuitable for the treatment, because unresponsive or because the treatment causes side effects (Temple 1994; FDA 2019). When those patients are the minority, they will only have a minor impact on the study, while when they represent a more consistent fraction of the enrolled population, the study's success may be compromised. Excluding those patients usually makes a treatment effect much easier to demonstrate, if exists. While this approach affects the generalizability of the study results, i.e. making it impossible to draw inferential conclusions on the whole population, a larger clinical trial in an unselected population i) can be more expensive, ii) may expose subjects to treatments that may be ineffective or even harmful and iii) may decrease the efficiency of the trial (especially if the benefiting subset of patients is small), iv) may lead to the misleading conclusion that the drug is effective on the whole population while there still exist subgroups of poor respondents. It is clear that, in the presence of heterogeneity, the estimated average treatment effect in the overall population is diluted by those not responding, leading to possible erroneous conclusions on a treatment that may be truly beneficial for some groups of patients and/or truly ineffective or dangerous for others. In these cases "we have a scientific and ethical obligation to identify these subgroups" (Magnusson and Turnbull 2013).

Issues of subgroup analysis. The heterogeneity in the response of patients to therapies has long been recognized as an issue in clinical development and, traditionally, post-hoc subgroup analysis has been implemented. However, subgroup analysis can be a "dangerous exercise" and, to obtain reliable results, the trial sample size must be large enough (see e.g. Maitournam and Simon 2005; Foster et al. 2011; Lipkovich et al.

2017 and references therein). A standard approach for subgroup analysis is to test the treatment by covariate interaction, but when the test is not planned in advance, it may be seriously underpowered (see also Wang et al. 2007). Besides the inherent possible concern of subgroup analysis, these methodologies do not provide definitive evidence of treatment effectiveness on subgroups, so a new confirmatory trial should be carried out to target the subgroup in which the new treatment seems to be effective.

Biomarkers. In precision medicine, a biomarker (i.e., a biological marker) is defined as a *"characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"* (Atkinson et al. 2001). This is a basic definition but, according to the context, specific definitions may exist (Califf 2018). Biomarkers are often used as a means to classify patients into subgroups. Following a broad classification, biomarkers can be prognostic and/or predictive. Prognostic biomarkers are those associated with the disease outcome, regardless of the treatment received, namely they identify subpopulations based on the outcome of interest. A biomarker is predictive when it is associated with drug response, namely when it separates subpopulations based on the outcome of interest given the received treatment. In practice, from a mathematical viewpoint, a biomarker can be treated as a covariate. A review on general biomarker-guided trials is in Antoniou et al. (2016).

Enrichment strategies. When there is some evidence that the effect of a treatment may differ in certain subpopulations and biomarkers identifying the potential benefitting subgroups are available, enrichment designs may be useful (Rosenblum and Van Der Laan 2011; Mandrekar and Sargent 2009b; Maitournam and Simon 2005) and a successful enrichment maneuver may increase the power of the study with a reduced sample size and trial's duration (Freidlin and Simon 2005; Jiang et al. 2007; Simon 2015; Magnusson and Turnbull 2013). Temple (2010) reviews the concepts of practical enrichment (to decrease variability), prognostic enrichment (enrolling more individuals more likely to experience the event of interest/disease-related condition) and predictive enrichment (accruing more individuals more likely to respond to a drug treatment). Especially predictive enrichment, i.e. enrichment based on a predictive biomarker, supports the transition towards personalized medicine (Temple 2010). The use of enrichment designs has been supported by the Food and Drug Administration (FDA 2020): *"enrichment strategies can increase the ability of a trial to detect an effect of the investigational drug"* and, in a dedicated document (FDA 2019), it is presented as a strategy to increase the efficiency of the drug development process and to support precision medicine. According to this document, enrichment is the *"prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is, in fact, present) is more likely than it would be in an unselected population"*. Similar strategies can be applied for safety assessments.

Trying to trace back the origin of enrichment, probably these kinds of designs found their origin in oncology. The literature (see e.g. Fedorov and Liu 2007 and references therein) seems in agreement to award Amery and Dony (1975) as one of the firsts that used this strategy. In their study, patients non-responding to the treatment are then excluded in a second study phase. Even if subpopulation enrichment had been employed in practice, the first papers that more formally discuss the enrichment

practice have been found in the 1990s (e.g. Hallstrom et al. 1991; Temple 1994; Davis et al. 1995). Russek-Coen and Simon (1997) considered the gender by treatment interaction. Most studies neglect such interaction and their aim is to design a study that takes into account that each gender may experience a different response to the treatment. In their two-stage design comparing a treatment to a control, after stage 1, the test for gender by treatment interaction is performed: if non-significant the trial is stopped and the overall treatment effect is computed, while if the test is significant the treatment effect is estimated separately for gender (after collecting further data in a second stage). Despite simple, this scheme has been adapted -with variations- in many of the enrichment trials up to these days. In the same year, we read in Follmann (1997) that *"if during the course of a fixed sample size trial substantial differences in response between strata become apparent, it may be tempting to quit randomizing patients from unresponsive strata and only randomize in the responsive strata. Intuitively, if we can accurately identify responsive and unresponsive strata midway through the trial, and subsequently randomize only from the responsive subgroups, then we should have a more powerful trial. Another reason for dropping a stratum occurs when the treatment shows substantial harm*".

Enrichment design based on biomarkers. The use of biomarkers to drive the design requires a deep understanding of the relationship between treatments and biomarkers and it is necessary to test the biomarker validity and utility (see, e.g., Freidlin et al. 2010; Antoniou et al. 2016). The process of developing and validating a biomarker is complex and requires considerable time and resources. Additionally, the biomarker role may change within the trial; for instance, it can be a stratification factor at earlier stages and then it can be used as a tool for inclusion/exclusion criteria (Ondra et al. 2016). Thus, more complex designs are required and this is stimulating research in several disciplines and discussion among statisticians, clinicians, and trial managers (see e.g. Antoniou et al. 2019).

Essentially, in enrichment designs, patients are screened for their biomarker profile, and then only those with or without certain characteristics are included in the trial. The enrichment can be based on a dichotomous or, in general, on a categorical covariate (e.g., sex, presence of genetic marker(s), a concomitant illness, mutation status) or a continuous one (e.g., age, blood pressure) that is often broken down into several categories (FDA 2019). The simplest scenario is when a binary predictive biomarker (or a binary classifier based on continuous and/or multiple biomarkers) can be identified at the planning stage. In this case, it is common to refer to the two ensuing subpopulations as *biomarker-positive* ($\mathcal{B}^+$) and *biomarker-negative* ($\mathcal{B}^-$) patients. In the case of substantial evidence that $\mathcal{B}^+$ patients benefit from the treatment, a traditional clinical trial enrolling only $\mathcal{B}^+$ may be appropriate and the design and analysis of the study are straightforward. However, regulatory guidelines for enrichment strategies (FDA 2019) have warned investigators that it is necessary to collect data also on the $\mathcal{B}^-$ patients, unless there is sufficient evidence of their unresponsiveness to the drug. In practice, it is frequently hard/impossible to have such a characterization of the population before phase III. Multiple sources of uncertainty may be present, such as incomplete information on the biomarker's prevalence or related to the type and strength of the biomarker-response relationship (FDA 2019). In addition, even if

some knowledge/biological rationale/experimental evidence for candidate biomarkers is available, it is still unclear which is the best way to use them or how they could be combined at this stage of clinical development. Moreover, when the candidate biomarker is measured on a continuous scale, the optimal cutpoint to discriminate among $\mathcal{B}^+$ and $\mathcal{B}^-$ patients is generally unknown (Jiang et al. 2007; Simon and Simon 2013; Spencer et al. 2016; Frieri et al. 2022). Here, the selection of a patients' subpopulation corresponds to the setting of a biomarker threshold discriminating $\mathcal{B}^+$ and $\mathcal{B}^-$. The decision rule to identify such cutoff changes according to the experimental setting, but it is reasonable to desire that $\mathcal{B}^+$ group should be as large as possible, to include the largest possible subset of benefitting subjects in future studies.
The objectives of enrichment designs are usually within the following:

- identification of the target population (i.e., $\mathcal{B}^+$);
- biomarker validation, namely estimation and testing the effect of the biomarker on the outcome;
- estimation of treatment effect in the whole population and/or in $\mathcal{B}^+$;
- test whether the experimental arm is better than the control in the full population and/or in $\mathcal{B}^+$.

Adaptive enrichment. To avoid an inappropriate enrichment of the population, with the risk of restricting too weakly or too hardly the enrollment in the study, adaptive enrichment designs are being increasingly used (FDA 2019). By using the information accrued on the marker performance during the course of the study, these designs are aimed at finding the trade-off between restricting the enrollment to increase the power of detecting a drug effect and being not too restrictive for avoiding a too small target population. With a smaller sample size, the accuracy of the estimators may deteriorate and the recruitment period may become prohibitive. The analysis of this kind of adaptive trial needs to take into account the bias arising from the use of the same data for both the subpopulation selection and the final analysis. Many authors, as well as FDA guidelines, pointed out the importance of controlling type I error rate in enrichment designs: because the modifications in the enrollment criteria depend on the outcome, standard methods of analysis do not not guarantee the control of type I errors.

Several additional statistical and logistical issues are added in the presence of a continuous candidate predictive biomarker. For instance, FDA (2019) mentions the problem of sample size determination to detect a treatment effect with a given power in $\mathcal{B}^+$ or in the overall population: this can be a difficult issue when there is uncertainty about the biomarker cutpoint identifying $\mathcal{B}^+$ and $\mathcal{B}^-$ and about the prevalence of $\mathcal{B}^+$ (i.e., the proportion of biomarker-positive patients on the whole population). In this framework, the choice of the biomarker threshold becomes part of the study's objectives.

Further discussion on the appropriateness of the enrichment maneuver and motivating examples can be found, for instance, in Mandrekar and Sargent (2009a) and Stallard et al. (2014), in the context of real clinical examples.

Notice that the wording "adaptive enrichment" may seem redundant as, intuitively, the idea of the enrichment is itself an adaptation. However, the reference literature adopts the terminology "adaptive enrichment design" to refer to a study in which

modifications of important aspects are allowed during the course of the trial based on the accrued information (e.g. refinement of the subpopulation, treatment assignments, study duration, and endpoints). For consistency reasons, we follow this nomenclature. In contrast, an example of a non-adaptive enrichment design could be a clinical trial in which only biomarker-positive patients are randomized (maybe based on considerations coming from previous knowledge) while biomarker-negative ones are excluded from the study (see, e.g., Antoniou et al. (2017)).

Within the paper the term biomarker may be used in a broader sense, intending, for example, a demographic characteristic or a biological characteristic. We basically restrict our thoughts to frequentist methodologies, though other strategies have been adopted. In Sect. 2 we discuss adaptive enrichment designs, while in Sect. 3 we introduce the notation and a general framework for enrichment studies. In Sect. 4 we review designs with two pre-defined subgroups (Sect. 4.1) and we discuss the case of quantitative biomarkers (Sect. 4.2). Section 5 is dedicated to an extensive discussion on open issues and directions for future research.

2 Adaptive enrichment designs

In adaptive enrichment designs, as the trial evolves the accrued information is used to better characterize the target population and, eventually, to restrict the enrollment for modifying the population under investigation. This protects against the possibility of making wrong decisions on the enrollment criteria (or using only a subset of data for the analysis) with the advantage of potentially having an improvement in power without increasing the sample size, making these designs appealing to both patients and sponsors (see e.g. Simon 2015).
The typical adaptive enrichment designs scheme is articulated in the following steps:

1. initial recruitment from the entire population meeting standard broad eligibility criteria;
2. interim analysis (one or more) to assess efficacy/safety in the entire population and in one (or more) subgroup(s);
3. design the rest of the trial based on the interim analysis.

Interim analysis could lead to the following actions:

action A): recruitment continues from the entire population (with possible sample size re-estimation);
action B): recruitment is restricted to one (or more) subgroup (enrichment design);
action C): trial stops for futility or efficacy;
action D): recruitment weighted towards one (or more) subgroup;
action E): stop randomization to one or more treatments and/or start randomizing subjects to one or more new arms.

These interim decisions may not be mutually exclusive; for example, sample size re-estimation can be matched with action B or C. Indeed, the sample size is often calculated at the beginning of the trial based on limited information. Thus, stage size can be adjusted based on interim data which could provide important information

(e.g., on the biomarker prevalence). Moreover, in many enrichment designs the total sample size is fixed: by stopping the recruitment from one or more groups at the interim analysis, a sample size re-allocation rule is performed as the recruitment will only involve the remaining patients. Note also that actions D and E involve some enrichment, but these designs can be thought of as a part of master protocols (see Sect. 5.3).

Adaptive enrichment designs were introduced when the overall population is partitioned into two subgroups specified in advance and most of the literature, even recently, has been focused on this framework (Jennison and Turnbull 2007; Wang et al. 2009; Rosenblum and Van Der Laan 2011; Rosenblum et al. 2020; Ballarini et al. 2021; Wu et al. 2022). This basically resembles the case in which there is a single dichotomous biomarker that identifies $\mathcal{B}^+$ and $\mathcal{B}^-$. Following the steps 1, 2, 3 above, they are adaptive in the sense that the inclusion criteria of patients are modified after the interim look at the data. A similar setting is when the population is partitioned into more than two groups, for instance by a polytomous biomarker or by a continuous one broken down into several predetermined categories.

As regards continuous biomarkers, several cutpoints can be a priori fixed, going back to the case of predefined subgroups; however, even if the cutoffs can be settled following a biological rationale, this procedure could clearly lead to several downsides (see e.g. FDA 2019; Simon 2015). There is relatively limited work providing insights into how these groups should be identified for continuous biomarkers, but particular care and attention are required for the threshold determination and validation (Spencer et al. 2016). Frequently, $\mathcal{B}^+$ and $\mathcal{B}^-$ are identified by data coming from sources other than clinical evidence: for instance, they could be elicited from animal models or defined by taking a convenient value as a cutoff (Jiang et al. 2007). However, recently, a methodological effort has been done to fill this gap. While some works are still based on categorizing a continuous biomarker by using a set of candidate cutpoints (e.g. quantiles/percentiles as in Jiang et al. 2007; Spencer et al. 2016; Stallard 2023), others treat the biomarker directly on a continuous scale (e.g. Diao et al. 2018; Lin et al. 2021; Frieri et al. 2022). These designs have the efficiency of enrichment without the need of dealing with predefined subgroups of patients.

3 General framework for enrichment designs

From now on, we restrict our attention to two-stage designs with a single predictive biomarker. Clearly, the proposed framework could be extended to the case of multiple stages and multiple biomarkers as well.

3.1 Notation and model

Consider a clinical experiment aimed at comparing an experimental treatment (T) versus a control (C). Let us denote by Y the patient outcome and, without loss of generality, assume that larger values of the outcomes are preferred. For instance, the

outcome may be binary, count, time-to-event, etc. Let also denote by $X \in \mathcal{X}$ a predictive biomarker and let $\delta \in \{T, C\}$ be the treatment indicator. A general approach consists of modeling $E[Y|X, \delta]$, for instance with linear regression. Via the biomarker X, the full population $\mathcal{F}$ can be partitioned into J disjoint subgroups $G_1, \ldots, G_J$, namely $\mathcal{F} = \bigcup_{j=1}^{J} G_j$ with $\bigcap_{j=1}^{J} G_j = \varnothing$. Possible other subpopulation(s) of interest could be built as appropriate by defining one (or more) subset(s) I of $\{1, \ldots, J\}$ such that $\mathcal{S}_I = \bigcup_{j \in I} G_j$. At the end of most of the enrichment designs, the interest is in the partition $\mathcal{F} = \mathcal{B}^+ \cup \mathcal{B}^-$ (with $\mathcal{B}^+$ being the subgroup to which the enrollment is restricted in the case of enrichment). The group $\mathcal{B}^+$ could be a single G_j or a composite subpopulation. In practice, the set $\{1, \ldots, J\}$ is partitioned into two disjoint sets, say I^* and its complement $(I^*)^c$ so that $\mathcal{B}^+ = \mathcal{S}_{I^*}$ and $\mathcal{B}^- = \mathcal{S}_{(I^*)^c}$. When X is qualitative, $\mathcal{X}$ is a finite (possibly unordered) set while if X is quantitative then $\mathcal{X} \subseteq \mathbb{R}$ and the definition of the benefitting subpopulation could be done via a biomarker threshold x^*, e.g. $\mathcal{B}^+ = \mathcal{S}_{I^*} = \{i \in \mathcal{F} : X_i \leq x^*\}$ (note that when X is quantitative, often the J nested subpopulations $\bigcup_{h=1}^{j} G_h$ for $j = 1, \ldots, J$ are of interest).

In particular,

- when $\mathcal{X} = \{x_1, \ldots, x_J\}$ the population is naturally divided into J subgroups with $G_j = \{i \in \mathcal{F} : X_i = x_j\}$. The binary biomarker is a particular case with $\mathcal{X} = \{x_1, x_2\}$, and the two biomarker levels naturally define the two subpopulations G_1 and G_2, that can be directly taken as the $\mathcal{B}^+$ and the $\mathcal{B}^-$ groups.
- When $\mathcal{X} \subseteq \mathbb{R}$, one or more cutoffs $x_1^*, \ldots x_J^*$ have to be identified to define $G_1 = \{i \in \mathcal{F} : X_i \leq x_1^*\}$, $G_j = \{i \in \mathcal{F} : x_j^* < X_i \leq x_{j+1}^*\}$ $j = 1, \ldots, J-1$ and $G_J = \{i \in \mathcal{F} : X_i > x_J^*\}$.

Let us denote by N the total sample size and by n_j and $p_j = n_j/N$ the size and the prevalence of subgroup j for $j = 1, \ldots, J$, where clearly $\sum_{j=1}^{J} n_j = N$. We also denote by $n^{(k)}$ stage k sample size, with $k = 1, 2$, such that $n^{(1)} + n^{(2)} = N$. The treatment effect (computed with respect to the control) in the overall population is denoted by $\Delta_{\mathcal{F}}$. Usually $\Delta_{\mathcal{F}} = E[Y|T] - E[Y|C]$ but it can also be the difference in hazard functions in T and C or defined on other scales like, e.g., log hazard ratio or log odds ratio. Analogously, the treatment effect in G_j is denoted by Δ_j and Δ_I is the treatment effect for any other subpopulation $\mathcal{S}_I$ with $I \subseteq \{1, \ldots, J\}$. The treatment effect can be both an absolute or a standardized measure. The hypothesis of interests at interim analysis or at the end could be $H_{0\mathcal{F}} : \Delta_{\mathcal{F}} = (\leq)0$ and/or $H_{0j} : \Delta_j = (\leq)0$ or other hypotheses on $\mathcal{S}_I$, i.e. $H_{0I} : \Delta_I = (\leq)0$. Based on the chosen hypothesis and the treatment effect measure, we can define the test statistics as functions of the sample data. In particular, $Z_{\mathcal{F}}$ is used to test $H_{0\mathcal{F}}$, Z_j for H_{0j} and Z_I for H_{0I}. The superscript (k) is used to indicate the stage to which the test statistic is referred, with $k = 1, 2$.

3.2 Interim decision making for enrichment designs

At the interim analysis, the decision-making is based on the collected data: this is a decision rule that maps stage 1 data into a decision concerning stage 2 enrollment (e.g.

Rosenblum and Van Der Laan 2011). In the most used framework, this decision rule consists of identifying the set $I^* \subseteq \{1, \ldots, J\}$ that defines the group $\mathcal{S}_{I^*} = \mathcal{B}^+$ to be enrolled in stage 2. Some identification rules are based on the estimated treatment effect, others on test statistics, and others are more complicated. Some of the most used rules in enrichment strategies to identify I^* are listed below under a "the larger the better" scenario.

ID1) Set $I^* = \{j : \Delta_j > \gamma\}$, namely select all subgroups where the treatment effect is larger than a minimal clinically significant difference γ (e.g. Freidlin and Simon 2005; Renfro et al 2014; Lin et al. 2021; Frieri et al. 2022).

ID2) Set $I^* = \arg\max_{j \in J} \Delta_j$, i.e., select the subgroup with maximum treatment effect (e.g. Stallard 2023).

ID3) Set $I^* = \arg\max_{j \in J} Z_j$, namely select the subgroup where the chosen test statistic Z_j is maximized.

ID4) Set $I^* = \{j : Z_j > c\}$ for a positive threshold c (see e.g. Magnusson and Turnbull 2013 and Kelly et al. 2005).

ID5) Set $I^* = \arg\max_{j \in J} n_j \Delta_j$, namely select the subgroup that maximizes the overall benefit (Stallard 2023). In the same spirit, another rule consists of taking $I^* = \arg\max_{j \in J} p_j^w \Delta_j$, where $w \in [0, 1]$ is a weight; this rule is related to the Kullback–Leibler information. In practice, the selection is based on a utility function that provides a trade-off between the size of the subgroup and its treatment effect (see also Lai et al. 2014; Joshi et al. 2020).

ID6) Set I^* based on a conditional power metric defined as the probability of obtaining a successful outcome at the end of the study given the interim data, e.g., the probability of rejecting the null hypothesis at the final analysis given the interim data (see e.g. Wu et al. 2022; Johnston et al. 2022).

ID7) Select the J^* best subgroups (with J^* pre-specified) in terms of Δ_j or Z_j (e.g. Friede et al. 2020).

Other options involve more complex and less interpretable functions of stage 1 data. Subgroup selection methods as described in Lipkovich et al. (2017) can be considered too. The choice of the identification rule can also involve clinical and economical considerations, as discussed in Sect. 5. All the above-mentioned rules generally depend on unknown population parameters that have to be estimated with the available data.

We share the opinion that selecting the subpopulation with the highest observed treatment effect (rule ID2) may be very limiting, especially with a monotonic treatment-biomarker relationship. While it is true that the interim decision should be based on a trade-off between the size of the subpopulation and the magnitude of the treatment effect, attention should be paid to the scale with which the benefitting subpopulation size is accounted for in rule ID5, as this could lead to too conservative decisions.

In the case of a binary biomarker, $I^* = \{1\}$ or $\{2\}$ so $\mathcal{B}^+$ coincides with G_1 or G_2. In combination with the previous identification rules, the following decision rules have been considered (with eff and saf denoting an efficacy and a safety boundary, respectively).

R1) If $\Delta_{I^*} < 0$ stop for futility, if $\Delta_{I^*} \geq 0 \cap \Delta_{\mathcal{F}} < 0$ recruit only $\mathcal{B}^+$, while if $\Delta_{I^*} \geq 0 \cap \Delta_{\mathcal{F}} \geq 0$ recruit from $\mathcal{F}$;

R2) stop recruiting from $\mathcal{B}^-$ if $Z^{(1)}_{(I^*)^c} \leq eff \leq 0$ or if $Z^{(1)}_{(I^*)^c} \geq saf$.

Usually, the interim decision rule is based on the primary outcome. However, especially in some contexts such as survival analysis, the use of secondary or surrogate endpoints may be helpful. For instance, Wu et al. (2022) proposed an adaptive enrichment design that combines the information from the primary and surrogate endpoints at the interim analysis, also allowing sample size re-estimation. They considered normal outcomes with predefined $\mathcal{B}^+$ and $\mathcal{B}^-$ groups and their decisions are based on conditional power.

Finally, notice that when multiple biomarkers are considered with possible interactions among them, the region of $\mathcal{X}$ guaranteeing enhanced treatment effect may be complex and the corresponding subpopulation would be defined by multiple constraints. In such a case there may not exist a single cutpoint to discriminate $\mathcal{B}^+$ and $\mathcal{B}^-$. A discussion on possible extensions to multiple biomarkers can be found in Sect. 5.2.

3.3 Analysis at the end of the study

At the end of stage 1, when the enrichment is not performed and the study is continued on the overall population in stage 2 (action A), the final analysis includes the whole sample of patients. Whereas, when the enrichment at the second stage occurred by enrolling $\mathcal{B}^+$ subjects (action B), the following possibilities have been employed for the final analysis: use only stage 2 data (which can be seen as a new study as in Renfro et al 2014 or Frieri et al. 2022), or combining data from the two stages as, for instance, in Simon and Simon (2013) or Wang et al. (2007). For this latter case, the key point is how to combine data from the two stages for a valid inference. The problem arises since adaptation is based on data-dependent choices and (often) the same data used for interim decisions, are included in the final analysis.

From an estimation viewpoint, statisticians' challenge has been to adjust from possible bias in the estimated treatment effect for the enriched population. Various methodologies based on bootstrap or cross-validation have been employed (see e.g. Simon and Simon 2017; Zhang et al. 2017).

Two of the most common methods for combining data from the two stages are the conditional error function approach (see e.g. Placzek and Friede 2019) and the combination function approach (see e.g. Bauer and Kohne 1994). A third method of combining data is to use adaptive likelihood ratio tests with a model that accounts for the dependencies caused by using stage 1 data for both the enrichment and the final analysis (see Flournoy and Tarima 2023; Tarima and Flournoy 2022). The authors proved that for models in one-parameter exponential family, likelihood ratio tests for natural parameters are uniformly most powerful conditional on the interim decision. While this latter method does not necessarily requires asymptotics, other methodologies are based on asymptotic considerations (Rosenblum and Van Der Laan 2011; Lin et al. 2021). Finally, other approaches used Bayesian techniques or a decision-theoretic framework (for instance Ondra et al. 2016, 2019; Ballarini et al. 2021).

In addition, especially in a confirmatory setting, the interest lies in testing and, since patients' subpopulations are usually multiple, potentially several hypotheses are tested (at the interim analysis and at the end of the study) and type I error rate may be inflated and more difficult to compute. Another related issue could be the low power to detect

a treatment effect in subpopulations with low prevalence. Thus, the authors' endeavor has been focused on developing procedures for multiple testing that guarantee the strong control of the familywise type I error rate (FWER), as a standard requirement of most clinical trials. When testing J null hypothesis H_{0j} for $j = 1, \ldots J$, FWER is the probability of rejecting any true simple hypothesis H_{0j} and it is controlled in the strong sense if FWER is lower than a given significance level α (see e.g. Hochberg and Tamhane 1987). In such a way the type I error for the test of the particular hypothesis H_{0j^*} is not greater than α. Note that, even when a single hypothesis is tested at the end, this could be because other hypotheses have been dropped after the interim analysis of the previous stage. Thus, it is necessary to account for this bias: the multiple testing problem arises from the fact that potentially any hypothesis is tested at the end and the probability of erroneously rejecting such hypotheses should be controlled.

One of the most used procedures to control FWER in the enrichment context is the closed testing procedure by Marcus et al. (1976). Let $H_{\mathcal{K}} = \cap_{j \in \mathcal{K}} H_{0j}$ be the intersection hypothesis with $\mathcal{K} \subseteq \{1, \ldots, J\}$. Then hypothesis H_{0j} can be rejected if and only if $H_{\mathcal{K}}$ is rejected at level α for all the subsets $\mathcal{K}$ that contain j. Reviews on testing treatment effect in subgroups as a multiple testing problem can be found in Jennison and Turnbull (2007); Stallard et al. (2014).

Finally, as noted by several authors (e.g. Stallard 2023) with nested subpopulations the test statistics are correlated. However, in this case, the analysis of data is comparable to the analysis of a group sequential trial, so similar methods can be used (see also Sect. 5.1).

4 Adaptive enrichment designs based on a single biomarker

Unless otherwise stated, all the designs presented in this section follow the steps of the typical adaptive enrichment design presented in Sect. 2.

4.1 The case of two predefined subgroups

This section is dedicated to reviewing some enrichment designs in which $\mathcal{B}^+$ and $\mathcal{B}^-$ groups are known before the study starts. This experimental scenario may correspond to an enrichment based on a dichotomous marker or on a higher dimensional classifier known in advance, e.g. developed based on scientific knowledge and previous studies.

Wang et al. (2007) were one of the firsts introducing adaptive enrichment designs. Without assuming a particular parametric model, rule R2 was applied to decide at the interim analysis whether continue recruiting $\mathcal{F}$ or just recruit $\mathcal{B}^+$ at stage 2 (action A or B). At the end of stage 2, the hypothesis of interest is $H_{0\mathcal{F}}$ (or H_{0I^*}) and the test statistic is a weighted average of $Z_{\mathcal{F}}^{(1)}$ and $Z_{\mathcal{F}}^{(2)}$ (or $Z_{I^*}^{(1)}$ and $Z_{I^*}^{(2)}$, respectively). Then the stage-wise p-values are computed and combined with a combination approach for the final test. The sample size is planned according to a non-adaptive approach, thus N is set to detect a minimum significant treatment effect in $\mathcal{F}$. Thus, N is fixed and it is not influenced by the interim decision and early stopping is not considered in their design (neither for efficacy nor for futility). In a later paper (Wang et al. 2009),

the authors deal with an adaptive enrichment design that accommodates sample size changes based on a conditional power metric computed at the interim analysis.

Rosenblum and Van Der Laan (2011) considered a very general data-generating distribution with limited support and finite variance and suggested a design in which actions A or B are taken according to ID3-ID4. Still equal sample sizes for T and C, as well as for G_1 and G_2 are set and N is pre-specified. Their interest is in testing $H_{0\mathcal{F}}$, H_{01} and H_{02} and, at the end of the trial, the appropriate null hypothesis (on the basis of the interim decision) is tested with a weighted combination of the test statistic computed at stage 1 and stage 2. The rejection rule is based on a threshold for this final test statistic that guarantees strong control of the asymptotic FWER at a given significance level (found as a solution of a numerical optimization problem). Surprisingly, under some assumptions, such a threshold is given by a standard normal cumulative distribution function and it coincides with the threshold used in a standard single-stage fixed design (namely, asymptotically, there is no type I error inflation). See also Stallard et al. (2014) for a further discussion and comparisons.

A different kind of approach can be found in Liu et al. (2010) and Yang et al. (2015). The enrollment is first restricted to $\mathcal{B}^+$ patients and then, if the interim data show promising results (in terms of magnitude of the estimated treatment effect or power), some $\mathcal{B}^-$ patients are enrolled to then assess the overall treatment effect.

Recently, Rosenblum et al. (2020) proposed a design in which at interim analysis the possible actions are A, B, or D. Their design is based on minimizing the total expected sample size under constraints on power and type I error rate. Due to the non-convexity of their optimization problem, it is computationally unfeasible to be solved directly and they address it by a sparse linear program, assuming equal allocation to T and C. The final analysis involves multiple testing procedures.

4.2 Enrichment designs with a quantitative biomarker

Here the objective is to find a single threshold discriminating between $\mathcal{B}^+$ and $\mathcal{B}^-$. The assumption of most papers is that the biomarker-response relationship is monotonically increasing (decreasing). In this case, the effect of the treatment is expected to increase as the levels of the biomarker grow, then $\mathcal{B}^+$ is the subpopulation having biomarker greater than a threshold x^*. As noted by many authors, the treatment effect may not be monotone, leading to a more complex definition of $\mathcal{B}^+$. For example, when the biomarker-response relationship is U-shaped it would be probably convenient to select two thresholds to discriminate between $\mathcal{B}^+$ and $\mathcal{B}^-$ (with a similar situation if the curve has a reversed U shape).

4.2.1 Enriching from candidate cutpoints

The first proposal aimed at both subgroup identification and testing the treatment effect is the Adaptive Signature Design by Freidlin and Simon (2005), later generalized by Jiang et al. (2007). Under time-to-event outcomes modeled by proportional hazards, Jiang et al. (2007) proposed two procedures intending to i) select the biomarker cutoff among a set of candidate thresholds and ii) assess whether T is better than C in $\mathcal{F}$

or $\mathcal{B}^+$ at the end of the study. These are not enrichment designs (they don't include a stage for enrichment) but are the first attempts to incorporate biomarker information with unknown discriminant into the study design. Sample size considerations are also included.

Later, Magnusson and Turnbull (2013) proposed a group sequential enrichment design with subgroup identification in K stages for normal outcomes. They consider J nested subpopulations and set K lower bounds and upper bounds $(l_k, u_k), k = 1, \ldots, K$ for stopping rules. After stage 1 the decision rule is aimed at dropping subpopulations of non-responsive patients: the enrollment is stopped from G_j for all the j s.t. $Z_j^{(1)} \leq l_1$. The remaining indexes, say I^*, define the subpopulation $\mathcal{S}_{I^*}$ to enroll in the $K-1$ subsequent stages. If $Z_{I^*}^{(1)} \geq u_1$ the trial stops with rejection of H_{0I^*} and $\mathcal{B}^+ = \mathcal{S}_{I^*}$. Otherwise, the trial is taken to the second stage (or stopped for futility). In the subsequent stages patients from $\mathcal{S}_{I^*}$ are enrolled and, at each stage k, if the test statistic exceeds u_k then H_{0I^*} is rejected and $\mathcal{B}^+ = \mathcal{S}_{I^*}$, while if the test statistic is lower than l_k, the trial is stopped for futility; otherwise the trial proceeds to stage $k+1$. Here, the authors assume the decision rule R2 based on ID4 and trial termination is guaranteed since the authors set $l_K = u_K$. Their methodology is focused on identifying l_k and u_k for $k = 1, \ldots K$ such that FWER is controlled in the strong sense at a given significance level. They also deal with the problem of correcting bias in point estimation via bootstrap. A possible limitation of this approach is that the subpopulation is not further adjusted after stage 1 (and each interim decision is based only on data from the previous stage).

A similar setting is considered in Lai et al. (2014) and Lai et al. (2019), who proposed a three-stage group sequential design with subgroup identification. After the first stage, if $H_{0\mathcal{F}}$ is rejected then the trial stops for efficacy (action C) and $\mathcal{B}^+ = \mathcal{F}$. Otherwise, the trial evolves to stage 2 by accruing patients from $\mathcal{F}$ (if the test statistics meet some futility boundaries) or from subgroup $\mathcal{S}_{I^*}$, associated with the largest value of the generalized likelihood ratio statistic (this corresponds to select the subgroup with the largest estimated Kullback–Leibler divergence, i.e., ID5). Then $H_{0\mathcal{F}}$ or H_{0I^*} is tested and, with similar considerations to the ones made at the first interim analysis, the trial is taken or not to stage 3. In such final stage the trial is continued on $\mathcal{F}$ or $\mathcal{S}_{I^*}$. Again, the aim is to find the efficacy/futility boundaries to which the test statistics have to be compared in order to guarantee FWER with a closing testing principle.

In the oncology area, Renfro et al (2014) considered survival outcomes and a continuous biomarker with a monotonic increasing relationship with the outcome. At the interim analysis, a set of candidate thresholds is considered and, for each cut-point, a Cox proportional hazard model is fitted taking the progression-free survival as the outcome. The threshold for which the strongest treatment-by-biomarker effect is observed is used to find I^*. If the trial is not stopped for futility, stage 2 is performed and, at the end, if the enrichment occurred the final test on $\mathcal{S}_{I^*}$ is limited to stage 2 data, otherwise the two stages' data are considered together. The allocation probabilities are 2/3 to T and 1/3 to C and $n^{(1)}$ is computed to have a minimum power for the treatment by biomarker interaction test.

An approach inspired by group sequential trials is in Graf et al. (2019), which considered nested subpopulations arising from several candidate cutpoints of a continuous biomarker and they tested the resulting nested hypothesis. This framework corresponds to a classical group sequential trial in which, at each interim analysis, a null hypothesis is tested. In such a case the test statistic is compared to adjusted critical boundaries that account for the correlation among the test statistics and guarantee control of the FWER. Such critical boundaries can be employed in the enrichment design but rely on the assumption of homoscedasticity across each subgroup. Via simulations, the authors discussed possible FWER inflation when this assumption does not hold, also proposing possible alternative tests.

With time-to-event endpoints, Kimani et al. (2020) were interested in estimation at the end of the study, when interim actions are B or C. The total population consists of subgroups determined by J candidate threshold a priori settled. They derive asymptotically unbiased estimators for the log-hazard ratio and interval estimators at the final analysis (with data from both stages) that are appropriate for some interim selection rules.

Stallard (2023), under the monotonicity assumption proposed a design with the objective of assessing whether T is superior to C in $\mathcal{B}^+$, combining the data from both stages. From a continuous biomarker, some candidate thresholds are settled after observing stage 1 data. Then, the subpopulations are taken as nested and, following action B at the interim, I^* can be defined by rules ID2, ID3 or ID5. At stage 2 the recruitment is restricted to $\mathcal{S}_{I^*} = \mathcal{B}^+$ and, at the end $H_{0I^*} : \Delta_{I^*} \leq 0$ is tested. Stage 1 p-values allowing for correction due to subgroup selection are computed based on multivariate normal distributions (or on Brownian motion approximation). Then, data from the two stages are combined using a combination test and the stagewise p-values are combined as proposed by Bauer and Kohne (1994). The discussion papers on Stallard (2023)'s contain several insights and suggestions, as well as possible alternative methods. For instance, the Tarima and Flournoy (2022) strategy does not require the monotonicity assumption (see also Flournoy and Tarima 2023).

Again in the framework of nested subpopulations, Placzek and Friede (2022) considered homoscedastic normal outcomes for T and C with the goal of testing $H_{0\mathcal{F}}$ and H_{0I} for all the nested sets I. In order to control FWER in the strong sense, the closed testing procedure is used, while the conditional error function is adopted to combine data from both stages. For interim analysis, the authors use ID4 and incorporated a blinded sample size recalculation with equal allocations to T and C. Based on their simulations, the optimal time for interim analysis is around 40%-50% of N in order to maximize the power of the conditional error function approach (see also Placzek and Friede (2019)).

Recently, Johnston et al. (2022) introduced a design with survival outcomes and four predictive biomarkers that are converted to categorical values by means of quartiles. The interim analysis is planned after 40% of the expected number of events occurred, regardless of other considerations (i.e., $n^{(1)}$ is fixed). At an interim stop, an event count re-estimation (to increase the number of events in the trial for guaranteeing power) and subgroup identification methods are performed. They consider various identification algorithms (like recursive partitioning and penalized regression methods) based on rules ID2 and ID3, obtaining a target population $\mathcal{B}^+$. Then, based on an approximation

of the conditional power metric with the log-rank statistics, the possible decisions are: continue the trial in $\mathcal{F}$ (action A), continue the trial in $\mathcal{B}^+$ (action B), or stop the study (action C). For the final analysis, combination tests and closed testing procedures are employed.

4.2.2 Enriching from estimated cutpoints

Simon and Simon (2013) proposed adaptive enrichment designs in which inclusion/exclusion criteria are sequentially updated based on previous data. They considered more than one biomarker whose information is combined via a binary classifier. The decision rule can be based on a binary classifier or on the cutpoint that maximizes the log-likelihood function. The main problem with this design is that, at the end of the study, the single null global hypothesis of whether exists a benefitting subpopulation is tested by combining data from all the stages via a weighted average of stage-wise test statistics. Thus, if rejected, we can state that there exists a benefitting subset of patients, but it is unknown how to identify $\mathcal{B}^+$ (Simon 2015; Simon and Simon 2017; Johnston et al. 2022). The indicated subpopulation may be the one defined by the enrollment criterion of the final stage of the trial, but a formal test on the treatment effect on $\mathcal{B}^+$ is missing. Later, Simon and Simon (2017) considered bootstrap methods to correct the bias in the estimation of the treatment effect in $\mathcal{B}^+$ defined by their last stage eligibility and explored the conditions under which, when the overall null hypothesis is rejected, a significant treatment effect in $\mathcal{B}^+$ can be claimed.

In the context of early-stage clinical trials (like phase II oncological trials), Spencer et al. (2016) presented a biomarker-adaptive threshold design for a single-arm trial with the objective of determining whether a subpopulation with a clinically relevant response rate exists. With a continuous biomarker and under monotonicity, for the enrichment strategy, the threshold is selected from a set of candidate cutpoints, but it is then estimated on a continuous range. In addition, stopping for futility is possible if no threshold is found. In the interim analysis, the biomarker threshold is selected following several steps based on a beta-binomial prediction model and predicted power. At the end of the study, data are combined to test the hypothesis that the response rate in $\mathcal{B}^+$ is greater than a clinically significant value. No design considerations are provided and $n^{(1)}$ and $n^{(2)}$ are fixed in advance.

Zhang et al. (2017) focused on the estimation problem for binary outcomes. With a fixed sample size, they use cross-validation and bootstrap to adjust the bias in the estimator of treatment effect in $\mathcal{B}^+$.

With time-to-event endpoints described by a Cox regression model, Diao et al. (2018) presented an adaptive enrichment design with fixed N. The design considers a single continuous biomarker with a monotonic relationship with the outcome. However, for threshold determination, they propose to use a grid search method. To select the cutoff, two decision rules are considered: one is based on minimizing the regression coefficient of the model and the other is based on maximizing the absolute value of the treatment by biomarker interaction effect (as in Renfro et al 2014). In the end, they test the null hypothesis that there is no difference between the hazard functions in T and C. Note that, by this approach, benefitting patients might also be in the $\mathcal{B}^-$ group.

Lin et al. (2021) adopted a bivariate normal model for the pair (Y, X) in a single-arm trial (the treatment effect is computed with respect to a historical control). At an interim decision, the biomarker threshold to discriminate between $\mathcal{B}^+$ and $\mathcal{B}^-$ is estimated directly on a continuous scale with rule ID1. In their framework, the correlation outcome-biomarker is an explicit gauge of the predictive nature of the biomarker. At the end of the study, the interest is in testing the treatment effect in $\mathcal{B}^+$ and whether the correlation equals 0 or it is positive. By letting $n^{(2)}$ tend to infinity they derive the asymptotic distribution of the maximum likelihood estimators of the treatment effect and the correlation. The target population for future use is identified by the maximum likelihood estimator of the threshold at the end of the study.

Under the bivariate normal model and the monotonicity assumption, Frieri et al. (2022) used rule ID1 for comparing T and C and $\mathcal{B}^+$ is identified with a threshold estimated on a continuous scale. In their design, stage 1 is specifically designed to estimate the threshold, while stage 2 is like a phase III confirmatory trial only on $\mathcal{B}^+$. This approach allows us to link the predictive strength of the biomarker to sample size considerations. Adopting equal allocation, first $n^{(2)}$ is computed to ensure a target power for the final analysis and then $n^{(1)}$ is chosen to efficiently estimate the biomarker cutoff at the interim analysis (imposing a bound on the mean squared error of the estimate). Before stage 2 starts, $n^{(2)}$ is re-estimated based on stage 1 data. They question whether enrichment is worthwhile when there are not enough resources for the threshold estimation at stage 1, including various simulation studies.

5 Discussion

5.1 Related methods: subgroup identification/treatment selection

There is a large body of literature on subgroup identification and subgroup selection, which is often not integrated with the enrichment design literature. Potentially, many subgroup selection methods could be incorporated into adaptive enrichment designs to drive the decision rule at interim analysis and to identify the subpopulation of interest for the subsequent stages. However, subgroup selection rules are usually complex and rely on many biomarkers/covariates: this might make the interim decision hard to interpret, with increasing difficulties in drawing design considerations. In addition, as recently addressed by Cai et al. (2022), many identification methods are aimed at finding the subgroup with the highest treatment effect, which could be in contrast to the spirit of enrichment designs. The reference paper reviewing some of the most common subgroup identification methods is the tutorial by Lipkovich et al. (2017). It is also worth noticing that often similar methodologies for subgroup selection are used for treatment selection since the problem of identifying a responsive subgroup of patients has analogies with the problem of selecting a treatment among multiple candidates (see e.g. Jennison and Turnbull 2007; Stallard et al. 2014; Wassmer and Dragalin 2015; Friede et al. 2020). Indeed, some authors adapted methodologies for treatment selection to determine subpopulations of patients (see e.g. Magnusson and Turnbull 2013). However, again, while it is obviously desirable to choose the best treatment it is often not appropriate to select only the “best" subpopulation.

5.2 Multiple biomarkers

As more predictive biomarkers are involved in subgroup identification, likely, the target population could be more precisely identified, but, at the same time, more complicated eligibility criteria have to be involved. Sometimes variable/model selection methodologies or methods to identify a reliable binary classifier can be helpful. Joshi et al. (2020) considered a three-stage design with multiple continuous biomarkers and continuous responses. They use equal allocation to T and C and, at the first interim analysis, the best subgroup is chosen by adopting ID5 with $w = 0.75$ and the recruitment of the second stage is restricted to the ensuing subpopulation. In the second interim analysis, stage 1–2 data are used to refine the subpopulation, now identified by ID5 with $w = 0.5$: this is the eligible population for stage 3. At the end of the trial, a weighted average of the stage-wise test statistic is used to test the treatment effect. Their strategy is similar to the one in Simon and Simon (2013) and thus suffers from the same drawbacks.

5.3 Multiple treatments

In the presence of more than two arms, the enrichment problem becomes something different. The main objective becomes to learn which patient subgroup benefits from which of the treatments under comparison. In this case, there is some overlap between enrichment design and master protocols. These procedures are commonly classified in the following categories (e.g. Woodcock and LaVange 2017).

- Platform trials: multiple treatments and possibly multiple subpopulations. Usually, a common control group is taken for comparison. They are very flexible, allowing for dropping ineffective arms or adding new ones during the trial.
- Umbrella trials: multiple treatments for a single disease. Based on predictive biomarker values, patients are enrolled in different cohorts.
- Basket trials: a single targeted therapy is evaluated on multiple diseases. For instance, a single drug targeting a particular molecular pathway or mutation is tested on different kinds of tumors to find out those for whom the drug is more effective.

For instance, Steingrimsson et al. (2021) introduced an adaptive enrichment design comparing two treatments to a common control in $\mathcal{B}^+$ and $\mathcal{B}^-$ a priori known. This is an example of a platform trial where six treatment-by-groups are studied to find which one leads to the better outcome. The authors pointed out that it would be interesting to extend their proposal to methods that can accommodate a continuous biomarker and adaptively find the cutoff based on the accumulating data during the trial. For a Bayesian adaptive enrichment umbrella trial see the paper by Ballarini et al. (2021).

5.4 Design considerations

Sample size. In many enrichment papers sample size is fixed in advance and planning considerations are restricted to sample size recalculations for the second stage. In

other papers, the total sample size is computed to detect a certain treatment effect in the overall population. Important parameters that affect sample size are the $\mathcal{B}^+$ prevalence and the variance of the outcome in this group (see e.g. Frieri et al. 2022; Placzek and Friede 2022). Also the time at which interim analysis has to be planned (i.e., $n^{(1)}$) is a critical issue (Rosenblum and Van Der Laan 2011; Renfro et al 2014; Lin et al. 2021). A later interim analysis would take advantage of more information for interim decisions and would share similar characteristics to a potential non-adaptive clinical trial. Instead, if an earlier interim analysis enables more flexibility due to more space for adaptation in the rest of the trial, the risk is to jeopardize the whole study by basing the interim decision on inadequate data (see e.g. Wu et al. 2022). The question that bounded many authors of enrichment designs is how to make a reliable interim decision. Unfortunately, very little work has been done to specifically determine the sample sizes ($n^{(1)}$ and $n^{(2)}$) according to optimal considerations in adaptive enrichment designs.

Optimal allocation. The majority of the papers on adaptive enrichment designs we found, adopt balanced allocation to T and C. Especially when patient heterogeneity is taken into account and covariate information is available at the design stage, comparable groups in terms of important prognostic covariates would be desirable. In addition, in the presence of predictive biomarkers, other objectives related to the enrichment problem may be of interest: for instance, maximizing the inferential precision in the estimation of the threshold of a continuous biomarker or maximizing the power of the test at the final analysis. Then some work should be done to derive allocation proportions that fulfill these objectives. Few exceptions are mentioned in the following. With a single dichotomous predictive biomarker, Zhu et al. (2013) proposed adaptive allocation to maximize the power of the test on the treatment-biomarker interaction in a linear model. Zhao et al. (2022) recently proposed a design that balances over prognostic covariates and assigns more patients to the best treatment based on the predictive covariates. They have some continuous covariates that are discretized, but the direct implementation on a continuous scale can be extended, for instance, with the recent results by Baldi Antognini et al. (2022). It would be desirable to adjust covariate adaptive randomization procedures to fit the enrichment design framework, with the objective of balancing the assignments also across important covariates. However, whether this is the optimal strategy or not is currently unknown and likely dependent on the specific trial objectives and final analysis (e.g., balanced allocation may not be optimal under heteroscedasticity). We feel this topic is worth investigation in future research. In addition, it would be interesting to adapt Covariate-Adjusted Response Adaptive procedures to study adaptive sequential enrichment designs in which multiple treatments and multiple patients subgroups are involved and, based on the responses, the allocation probabilities to each treatment change by assigning more patients to the best treatment for each subgroup (including the possibility of dropping ineffective arms within each subgroup as in platform trials).

5.5 Is it worth enriching?

There are still many interesting open questions regarding how to optimally use the adaptive enrichment framework in clinical development (see e.g. Simon and Simon 2017; Stallard 2023; Frieri et al. 2022). We summarize the typical issues and concerns of the use of enrichment designs, that could contribute to possible wrong study conclusions.

- **Design issues**. See Sect. 5.4.
- **Analysis issues**. See Sect. 3.3. The main issue arises when the same data are used to select patients' subgroups and to make the final inference.
- **Accuracy in the identification of the target population**. Most of the experimental scenarios of enrichment designs assume that $\mathcal{B}^+$ or $\mathcal{B}^-$ are prefixed or correctly determined by the biomarker level/value. This is clearly related to the knowledge and understanding of the biomarker-response relationship. Moreover, in practice, the biomarker assay may have a non-negligible rate of false negative and/or false positive (Wang et al. 2007; Freidlin and Korn 2014). For instance, Maitournam and Simon (2005) discussed the impact of biomarker misclassification on trial efficiency. Another critical issue arises when the biomarker is not binary. In particular, how should we select the cutoff? Or when there are several candidate biomarkers, how do we select one or how do we combine them? When this decision has to be made? Note that when a continuous biomarker is dichotomized, the treatment effect and the size of the target population are functions of the biomarker's cutoff defining the subpopulation. Thus, the uncertainty on the appropriate biomarker threshold may induce increasing complexity/variability in the decision-making (see e.g. Simon 2015; Frieri et al. 2022).
- **Uncertainty in the recruitment rate**. There is often a high uncertainty in the prevalence of biomarkers at the design stage, which makes it difficult to plan the recruitment, and that might increase the study length. Biomarker prevalence is often empirically estimated from a screening stage, but this estimate may be problematic when the study size is small or if the true prevalence is small. Indeed, many enrichment designs claim a gain in terms of sample size; however, it has to be considered that by restricting the eligible population, especially if the biomarker-positive prevalence is low, the amount of time to enroll the required number of patients may drastically increase. Then a large number of patients might have to be screened in order to find enough biomarker-positive patients and the study duration could be even longer than a standard clinical trial design (see e.g. Stallard et al. 2014; Freidlin and Korn 2014; Simon and Simon 2017). For instance, Simon and Maitournam (2004) compared an enrichment and a traditional trial in an experimental scenario in which preliminary data support the efficacy of the treatment only in a subpopulation. Their comparison is based on the number of patients screened for the enrichment trial and the number of patients randomized for the traditional one.
- **Length of the study**. Another time-related aspect is related to the time to observe the outcome (Wang et al. 2007; Steingrimsson et al. 2021; Wu et al. 2022), especially for those therapeutic areas in which the primary endpoint could take longer

to be observed (e.g. overall survival, progression-free survival). However, this is a common problem of adaptive designs. In these cases, it could be convenient to identify (if exists) a short-term surrogate endpoint which is a good proxy of the primary endpoint to be used in place of this for the interim analysis. Note that, despite surrogate endpoints may be observed quicker, due to possible misclassification, the design efficiency may be affected. Other related issues can be found in Uozumi et al. (2019). They discussed patients' recruitment methods in which the enrollment is not stopped at interim, which may help in shortening the total trial period.

- **Size of the subpopulation for intended use**. If the treatment is effective only in a very small subpopulation, economic reasons could lead to not finding it worthwhile to continue the clinical development within this group of patients. On the other hand, if the target population is almost as large as the whole population (and the treatment is not expected to be harmful to the excluded patients), it may not be worth bearing the costs of screening to identify the subpopulation from an economic viewpoint. These considerations might also depend on the gravity/rarity of the disease.
- **Ethical issues**. From the patient's perspective it should be considered that the biomarker screening may be invasive (it often happens in oncology), or may have a high failure rate, and it may increase the time to receive the therapy (Antoniou et al. 2019).
- **Cost and funding issues**. The scientific and logistic complexity of enrichment designs could increase the cost. First, the financial cost of evaluating the biomarker needs to be taken into account, especially when the number of patients that can be enrolled into the trial is small versus the number of those that have to be screened (Stallard et al. 2014; Thall 2021). Moreover, it has to be considered that the screening cost increases as the definition of eligibility criteria becomes more specific (e.g., because multiple biomarkers are used for interim decisions). In practice, on the one hand, founders could be more enthusiastic to support a clinical study with higher flexibility and chance of success, on the other hand, it has to be acknowledged that there is high uncertainty when trying to predict the total cost of a trial (Antoniou et al. 2019). Other sources of the increasing costs may be due to a higher administrative burden and support coming from the collaboration of different expertise.
- **Communication issues**. It is essential to provide accurate and effective information about the enrichment trial to all the relevant stakeholders to make them aware of the characteristics, potential advantages and disadvantages of these studies (Antoniou et al. 2019).
- **Implementation issues**. To facilitate the application of these innovative designs, a user-friendly computer program/commercial software should be made available to practitioners for the implementation and evaluation of adaptive enrichment designs, although some `R` packages have been developed recently (Friede et al. 2020).

5.6 Concluding remarks

Enrichment designs are typically used in phase II and/or phase III clinical trials. For example, the first stage may be phase II, in which the biomarkers are selected based on the responses to the treatment, and the accrual is then restricted in the second stage (i.e., phase III). Other interesting circumstances may be when seamless phase II/III trials can be used or when patient groups are subsequently excluded when there is greater evidence that they do not benefit from the treatment.

As is evident from this review, most of the authors have worked on developing methods for combining the data from two or more stages. Such an approach is obviously more attractive as it can be claimed at being more efficient. However, others still support the idea that a separate confirmatory trial to specifically target the biomarker-positive subpopulation is required. Indeed, phase III studies are the conclusive stage of clinical development, so they should be designed to assess definitive evidence on the treatment effectiveness (see e.g. Wang et al. 2009; Freidlin and Korn 2014). In addition, Mehta and Gao (2011) pointed out that two separate clinical studies for stages 1 and 2 could be more attractive from the trial sponsor's viewpoint.

More work needs to be done to develop an optimal strategy to update the enrollment criteria of adaptive enrichment trials. This is still a "very open question" (Simon 2015) and *"there is a need for more rigorous methodology and improved approaches for biomarker threshold selection...when naturally continuous or combined biomarkers or signatures are utilized"* as it is pointed out in the review by Renfro et al. (2016). Often, due to the limited amount of theoretical properties and closed form solutions, simulation studies have played a relevant role in assessing the operating characteristic of adaptive enrichment designs (Friede et al. 2020), making more complicated the planning of the study. Adaptive enrichment designs in which no predefined subgroups have to be settled are more flexible: the biomarker can be both selected and validated in the same trial. However, these studies should be carefully planned to maximize their efficiency by including sample size considerations and an appropriate randomization procedure.

Acknowledgements The authors of this paper wish to thank the Guest Editors, and the referee who made substantial comments that improved the paper. This research was supported by EU funding within the NextGenerationEU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (Project no. PE00000007, INF-ACT).

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

References

Amery W, Dony J (1975) A clinical trial design avoiding undue placebo treatment. J Clin Pharmacol 15(10):674–679

Antoniou M, Jorgensen AL, Kolamunnage-Dona R (2016) Biomarker-guided adaptive trial designs in phase II and phase III: A methodological review. PLoS ONE 11(2):1–30

Antoniou M, Kolamunnage-Dona R, Jorgensen AL (2017) Biomarker-guided non-adaptive trial designs in phase II and phase III: a methodological review. J Pers Med 7(1):1

Antoniou M, Kolamunnage-Dona R, Wason J et al (2019) Biomarker-guided trials: challenges in practice. Contemp Clin Trials Commun 16(100493):1–10

Atkinson A, Colburn W, Degruttola V et al (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95

Baldi Antognini A, Frieri R, Zagoraiou M et al (2022) The efficient covariate-adaptive design for high-order balancing of quantitative and qualitative covariates. Stat Pap. https://doi.org/10.1007/s00362-022-01381-1

Ballarini NM, Burnett T, Jaki T et al (2021) Optimizing subgroup selection in two-stage adaptive enrichment and umbrella designs. Stat Med 40(12):2939–2956

Bauer P, Kohne K (1994) Evaluation of experiments with adaptive interim analyses. Biometrics 50(4):1029–1041

Cai H, Lu W, Marceau West R et al (2022) Capital: optimal subgroup identification via constrained policy tree search. Stat Med 41(21):4227–4244

Califf R (2018) Biomarker definitions and their applications. Exp Biol Med 243(3):213–221

Davis CE, Applegate WB, Gordon DJ et al (1995) An empirical evaluation of the placebo run-in. Controlled Clin Trials 16(1):41–50

Diao G, Dong J, Zeng D et al (2018) Biomarker threshold adaptive designs for survival endpoints. J Biopharm Stat 28(6):1038–1054

FDA (2019) Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products guidance for industry

FDA (2020) Enhancing the diversity of clinical trial populations–eligibility criteria, enrollment practices, and trial designs

Fedorov VV, Liu T (2007) Enrichment design. Wiley encyclopedia of clinical trials, New York, pp 1–8

Flournoy N, Tarima S (2023) Discussion on "adaptive enrichment designs with a continuous biomarker" by Nigel Stallard. Biometrics 79(1):31–35

Follmann D (1997) Adaptively changing subgroups proportions in clinical trials. Stat Sin 7:1085–1102

Foster J, Taylor J, Ruberg S (2011) Subgroup identification from randomized clinical trial data. Stat Med 30:2867–2880

Freidlin B, Korn EL (2014) Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol 11(2):81–90

Freidlin B, Simon R (2005) Adaptive signature design: an adaptive clinicaltrial design for generating and prospectivelytesting agene expression signature for sensitive patients. Clin Cancer Res 11(21):7872–7878

Freidlin B, McShane L, Korn E (2010) Randomized clinical trials with biomarkers: design issues. J Natl Cancer Inst 102(3):152–160

Friede T, Stallard N, Parsons N (2020) Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: methods, simulation model and their implementation in R. Biom J 62(5):1264–1283

Frieri R, Rosenberger W, Flournoy N et al (2022) Design considerartions for two stage enrichment trials. Biometrics. https://doi.org/10.1111/biom.13805

Graf A, Wassmer G, Friede T et al (2019) Robustness of testing procedures for confirmatory subpopulation analyses based on a continuous biomarker. Stat Methods Med Res 28(6):1879–1892

Hallstrom AP, Verter J, Friedman L (1991) Randomizing responders. Controlled Clin Trials 12(4):486–503

Hochberg Y, Tamhane A (1987) Multiple comparison procedures. Wiley, New York

Jennison C, Turnbull B (2007) Adaptive seamless designs: selection and prospective testing of hypotheses. J Biopharm Stat 17:1135–1161

Jiang W, Freidlin B, Simon R (2007) Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst 99(13):1036–1043

Johnston SE, Lipkovich I, Dmitrienko A et al (2022) A two-stage adaptive clinical trial design with data-driven subgroup identification at interim analysis. Pharm Stat 21(5):1090–1108
Joshi N, Nguyen C, Ivanova A (2020) Multi-stage adaptive enrichment trial design with subgroup estimation. J Biopharm Stat 30(6):1038–1049
Kelly PJ, Roshini Sooriyarachchi M, Stallard N et al (2005) A practical comparison of group-sequential and adaptive designs. J Biopharm Stat 15(4):719–738
Kimani P, Todd S, Renfro L et al (2020) Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection. Stat Med 39(19):2568–2586
Lai T, Lavori P, Liao O (2014) Adaptive choice of patient subgroup for comparing two treatments. Contemp Clin Trials 39(2):191–200
Lai T, Lavori P, Tsang K (2019) Adaptive enrichment designs for confirmatory trials. Stat Med 38(4):613–624
Lin Z, Flournoy N, Rosenberger W (2021) Inference for a two-stage enrichment design. Ann Stat 49(5):2697–2720
Lipkovich I, Dmitrienko A, D'Agostino R Sr (2017) Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 36(1):136–196
Liu A, Liu C, Li Q et al (2010) A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. Clin Trials 7(5):537–545
Magnusson BP, Turnbull BW (2013) Group sequential enrichment design incorporating subgroup selection. Stat Med 32(16):2695–2714
Maitournam A, Simon R (2005) On the efficiency of targeted clinical trials. Stat Med 24:329–339
Mandrekar S, Sargent D (2009) Clinical trial designs for predictive biomarker validation: one size does not fit all. J Biopharm Stat 19(3):530–542
Mandrekar S, Sargent D (2009) Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol 27(24):4027–4034
Marcus R, Peritz E, Gabriel K (1976) On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63(3):655–660
Mehta CR, Gao P (2011) Population enrichment designs: case study of a large multinational trial. J Biopharm Stat 21(4):831–845
Ondra T, Dmitrienko A, Friede T et al (2016) Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. J Biopharm Stat 26(1):99–119
Ondra T, Jobjörnsson S, Beckman RA et al (2019) Optimized adaptive enrichment designs. Stat Methods Med Res 28(7):2096–2111
Placzek M, Friede T (2019) A conditional error function approach for adaptive enrichment designs with continuous endpoints. Stat Med 38(17):3105–3122
Placzek M, Friede T (2022) Blinded sample size recalculation in adaptive enrichment designs. Biom J. https://doi.org/10.1002/bimj.202000345
Renfro LA, Coughlin CM, Grothey AM et al (2014) Adaptive randomized phase ii design for biomarker threshold selection and independent evaluation. Chin Clin Oncol 3(1):3489
Renfro LA, Mallick H, An MW et al (2016) Clinical trial designs incorporating predictive biomarkers. Cancer Treat Rev 43:74–82
Rosenblum M, Van Der Laan M (2011) Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. Biometrika 98(4):845–860
Rosenblum M, Fang E, Liu H (2020) Optimal, two-stage, adaptive enrichment designs for randomized trials, using sparse linear programming. J R Stat Soc Ser B 82:749–772
Russek-Coen E, Simon R (1997) Evaluating treatments when a gender by treatment interaction may exists. Stat Med 16:455–464
Simon N (2015) Adaptive enrichment designs: applications and challenges. Clin Invest (Lond) 5(4):383–391
Simon N, Simon R (2013) Adaptive enrichment designs for clinical trials. Biostatistics 14(4):613–625
Simon R, Maitournam A (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10:6759–6763
Simon R, Simon N (2017) Inference for multimarker adaptive enrichment trials. Stat Med 36:4083–4093
Spencer A, Harbron C, Mander A et al (2016) An adaptive design for updating the threshold value of a continuous biomarker. Stat Med 35:4909–4923
Stallard N (2023) Adaptive enrichment designs with a continuous biomarker (with discussion). Biometrics 79(1):9–19. https://doi.org/10.1111/biom.13644

Stallard N, Hamborg T, Parsons N et al (2014) Adaptive designs for confirmatory clinical trials with subgroup selection. J Biopharm Stat 24(1):168–187
Steingrimsson JA, Betz J, Qian T et al (2021) Optimized adaptive enrichment designs for three-arm trials: learning which subpopulations benefit from different treatments. Biostatistics 22(2):283–297
Tarima S, Flournoy N (2022) Most powerful test sequences with early stopping options. Metrika 85(4):491–513
Temple R (1994) Special study designs: early escape, enrichment, studies in non-responders. Commun Stat 2(23):81–90
Temple R (2010) Enrichment of clinical study populations. Clin Pharmacol Ther 88(6):774–778
Thall P (2021) Adaptive enrichment designs in clinical trials. Annu Rev Stat Appl 8(4):393–411
Uozumi R, Yada S, Kawaguchi A (2019) Patient recruitment strategies for adaptive enrichment designs with time-to-event endpoints. BMC Med Res Methodol 19(1):159
Wang S, O'Neil R, Hung H (2007) Approaches to ecaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat 6:227–244
Wang S, Hung H, O'Neil R (2009) Adaptive patient enrichment designs in therapeutic trials. Biometr J 51(2):357–374
Wassmer G, Dragalin V (2015) Designing issues in confirmatory adaptive population enrichment trials. J Biopharm Stat 25(4):651–669
Woodcock J, LaVange LM (2017) Master protocols to study multiple therapies, multiple diseases, or both. N Engl J Med 377(1):62–70
Wu L, Li Q, Liu M et al (2022) Incorporating surrogate information for adaptive subgroup enrichment design with sample size re-estimation. Stat Biopharm Res 14(4):493–504
Yang B, Zhou Y, Zhang L et al (2015) Enrichment design with patient population augmentation. Contemp Clin Trials 42:60–67
Zhang Z, Chen R, Soon G et al (2017) Treatment evaluation for a data-driven subgroup in adaptive enrichment designs of clinical trials. Stat Med 37:1–11
Zhao W, Ma W, Wang F et al (2022) Incorporating covariates information in adaptive clinical trials for precision medicine. Pharm Stat 21(1):176–195
Zhu H, Hu F, Zhao H (2013) Adaptive clinical trial designs to detect interaction between treatment and a dichotomous biomarker. Can J Stat 41(3):525–539