Design considerations for two-stage enrichment clinical trials

Design Considerations for Two Stage Enrichment Clinical Trials

Rosamarie Frieri

Department of Statistical Sciences, University of Bologna, Bologna, Italy

email: rosamarie.frieri2@unibo.it

and

William F. Rosenberger

Department of Statistics, George Mason University, Fairfax, VA, US

email: wrosenbe@gmu.edu

and

Nancy Flournoy

Department of Statistics, University of Missouri, Columbia, MO, US

email: flournoyn@missouri.edu

and

Zhantao Lin

Department of Statistics, Data and Analytics, Eli Lilly and Company, Indianapolis, IN, US

email: zlin4@masonlive.gmu.edu

SUMMARY: When there is a predictive biomarker, enrichment can focus the clinical trial on a benefiting subpopulation. We describe a two-stage enrichment design, in which the first stage is designed to efficiently estimate a threshold and the second stage is a "phase III-like" trial on the enriched population. The goal of this paper is to explore design issues: sample size in stages 1 and 2, and re-estimation of the stage 2 sample size following stage 1. By treating these as separate trials, we can gain insight into how the predictive nature of the biomarker specifically impacts the sample size. We also show that failure to adequately estimate the threshold can have disastrous consequences in the second stage. While any bivariate model could be used, we assume a continuous outcome and continuous

biomarker, described by a bivariate normal model. The correlation coefficient between the outcome and biomarker is the key to understanding the behavior of the design, both for predictive and prognostic biomarkers. Through a series of simulations we illustrate the impact of model misspecification, consequences of poor threshold estimation, and requisite sample sizes that depend on the predictive nature of the biomarker. Such insight should be helpful in understanding and designing enrichment trials.

Key words: Bivariate normal model, Predictive biomarker, Sample size and power, Threshold estimation.

1. Introduction

Traditional clinical trials usually assume homogeneous treatment effects for all patients. However, patients can respond differently to the same medication. For example, beta 2-agonists for asthma are ineffective for 40 to 70% of patients (Abrahams and Silver, 2009). In this situation, the homogeneity assumption is violated. Therefore, a design based on this assumption may conclude incorrectly that a new drug is ineffective, while in fact it is effective in a subgroup of patients. The enrichment design is an efficient approach for precision medicine, where a subpopulation of patients is selected to enroll in a clinical trial based on one or more biomarkers, and the safety and/or effectiveness of the drug in selected populations is evaluated.

Since the subpopulation is identified by means of biomarkers that are suspected to impact the treatment effect, the more the knowledge of candidate biomarkers we have, the more likely we are to correctly identify patients belonging to the benefiting subgroup of patients. Biomarkers can be categorized into two types. A prognostic biomarker is correlated with the patient outcome, independent of the treatment received. A predictive biomarker is correlated with both the treatment and the patient outcome. To achieve a significant treatment effect and improve the efficiency of clinical trials, predictive biomarkers can be utilized. In some cases, subgroups of patients are well-defined by biomarkers, so outcomes of different treatments can be compared within each subgroup. However, in many cases, biomarkers are measured on a continuous scale without a known dichotomy [e.g., Spencer et al. (2016); Stallard (2022)]. A threshold that identifies a benefiting population must be defined.

There are several philosophical approaches to the design of enrichment trials in the literature. The first is a planned adaptive design, in which an interim inspection of the data is conducted to determine if there is a subpopulation that benefits. If there is no subpopulation, the trial is continued as planned; otherwise, the randomization is restricted to include only

patients in the subpopulation. Data from both stages are included in the final inference. Consequently, subjects in the first stage are more heterogeneous than the second stage, and the final treatment effect in the subpopulation is dampened (or biased) by those not in the subpopulation in the first stage. Simon and Simon (2013) describe using bootstrapping techniques to correct for the bias induced by combining the two stages. One question of interest is the preservation of type I error rates in the adaptive structure of the design [e.g., Simon and Simon (2013); Stallard (2022)]. Zhao and LeBlanc (2020) efficiently design a clinical trial to yield a larger impact in the full population. In their approach, an optimality criterion is established to increase positive outcomes of patients both in the subpopulation and outside the subpopulation. This expands the scope of the trial in cases where the subpopulation benefits, but also patients outside the subpopulation may receive some, albeit less, benefit.

The second approach is more controversial, because it involves an unplanned decision. An interim inspection yields no treatment effect, and yet a further inspection of the data identifies a subpopulation in which there may be some signs of efficacy. At this point, the trial design is changed to enroll only subjects in the subpopulation. Unplanned changes in many aspects of the trial are commonly done, for example, in eligibility criteria to improve recruitment or diminish safety concerns. Other possibilities are changing the endpoint, the interim monitoring plan, or the randomization probabilities. Consideration of type I error rate preservation when such changes are unplanned is of concern [e.g. Posch and Proschan (2012)].

The third approach is the one we take in this paper. A trial is designed to specifically determine a threshold of a predictive biomarker so that a new trial can be designed to target only that subpopulation. The goal of the first trial is solely to efficiently estimate the threshold, which depends on the correlation between a continuous outcome and a continuous

biomarker. We specifically delineate the correlation and the impact on relevant sample size to estimate the threshold within some tolerance. This threshold will then be used to enroll patients in a full-scale, two-armed treatment trial to determine efficacy. This is the approach of Spencer et al. (2016), who designed a trial to estimate the boundary of the biomarker, which "... allows later phase studies to focus on this subpopulation alone". We describe the design of such a trial. Our goal is to provide a complete template for the design of the two studies, should this approach be chosen.

The idea of characterizing the subgroup in a training sample and then evaluating the subgroup in a test sample is referred to as "sample splitting" by Simon (2015). Simon and Simon (2013, 2017) and Simon (2015) point out that using data from an early stage clinical trial will likely lack the requisite sample size to estimate an appropriate threshold. But Wang et al. (2009) indicate that "... a replication study is usually needed to confirm the treatment effect found for the adaptively enriched subset and, the study is likely to be smaller due to a larger ... effect size compared with the overall effect size". One goal of this paper is to determine how large the preliminary study needs to be to efficiently estimate the threshold, and describe the consequences in the second phase of the trial if the sample size is not adequate to estimate an appropriate threshold within some bound on the mean squared error. Following Lin, Rosenberger and Flournoy (2021) we use a bivariate normal model, which directly shows the relationship of the correlation of the biomarker and outcome. The model allows insight into how the predictive nature of the biomarker impacts sample size and power. Besides providing insights we present a complete template for design considerations for use in practice.

The two stage enrichment design proposed in this paper works as follows. In stage 1, patients meeting the eligibility criteria are randomized to either treatment or control group and the value of a biomarker X is observed for each patient. Data collected at stage 1

are used to select a biomarker threshold that determines a subpopulation of patients who will likely benefit from the treatment. We assume that a larger response is better and that X is a predictive biomarker. In addition, it is expected that X predicts the response in a monotonic relationship, meaning that higher biomarker values are associated with larger responses. Thus, if x^* is the selected biomarker threshold, then the population of interest consists of patients whose biomarker value is above x^*. In stage 2, only patients in that subpopulation are enrolled. Clearly the value of x^* depends on the predictive nature of the biomarker; when x^* is too small the subpopulation tends to coincide with the whole population, whereas when x^* is too large the subpopulation tends to be more exclusive. In Section 2 the two stage adaptive enrichment design is described and considerations on the size of the subpopulation are given. The sample size selection procedures for both stage 1 and 2 are reported in Section 3. Extensive simulation studies are contained in Section 4. The paper ends with a discussion and suggestions for future developments (Section 5).

2. Two stage adaptive enrichment design

The clinical trial consists of two stages: n_1 patients are enrolled in stage 1 and n_2 in stage 2, with $n = n_1 + n_2$. Each patient is assigned to treatment A (experimental) or B (standard treatment) by some randomization procedure that does not depend on outcomes or biomarkers. Let T_i be the treatment indicator such that $T_i = 1$ if patient i has received treatment A and 0 otherwise for $i = 1, \ldots, n$. For each patient we observe a continuous response y_i and a single continuous biomarker value x_i. For ease of notation let us denote $Y_A = Y_i|(T_i = 1)$ and $Y_B = Y_i|(T_i = 0)$ for $i = 1, \ldots, n$. We assume $Y_A \sim N(\mu_{Y|A}, \sigma^2_{Y|A})$ and $Y_B \sim N(\mu_{Y|B}, \sigma^2_{Y|B})$ with $\sigma^2_{Y|A}$ and $\sigma^2_{Y|B}$ known and $X \sim N(\mu_X, \sigma^2_X)$ (the distribution of the biomarker does not depend on the treatment assigned). Moreover, $Cov(Y_A, X) = \rho_A \sigma_{Y|A} \sigma_X$

and $Cov(Y_B, X) = \rho_B \sigma_{Y|B} \sigma_X$ so that

$$\begin{aligned}(Y_A, X)^T &\sim bN(\mu_{Y|A}, \mu_X, \sigma_{Y|A}, \sigma_X, \rho_A) \\ (Y_B, X)^T &\sim bN(\mu_{Y|B}, \mu_X, \sigma_{Y|B}, \sigma_X, \rho_B),\end{aligned} \tag{1}$$

where $bN(\cdot)$ is the bivariate normal density. We work under the assumption that subject observations (Y_i, X_i) for $i = 1, \ldots n$ are independent, conditional on the treatment assignment. We also assume that a higher response is more desirable for patients, a larger biomarker value implies a better outcome and that X is a predictive biomarker (namely $\rho_A > \rho_B$).

Note that, under model (1), if $\rho_A = \rho_B = 0$ the biomarker is neither prognostic nor predictive since it is unrelated to the outcome. If, instead $\rho_A = \rho_B > 0$ the biomarker is prognostic but not predictive; higher values lead to a better responses regardless of the treatment. Clearly, in both these cases no target population can be identified based on the biomarker. On the other hand, when $\rho_A > \rho_B$, the biomarker is predictive and it is reasonable to design a trial to discover whether a benefiting subpopulation could be defined.

2.1 *Stage 1*

The aim of stage 1 is to estimate a subpopulation defined by a biomarker threshold in which a given treatment effect can be detected. In particular, we are interested in the subpopulation in which the treatment effect is greater than a given value, say c with $c > \mu_{Y|A} - \mu_{Y|B}$. Considerations on the definition of c are given later in this subsection. Since we assume that the biomarker is positively correlated with the response, such a target population is identified as the subset of patients whose biomarker value is above a threshold, say x^*. Let $\theta_A = (\mu_{Y|A}, \rho_A, \mu_X, \sigma_X^2)$ and $\theta_B = (\mu_{Y|B}, \rho_B, \mu_X, \sigma_X^2)$; by the computations in Web Appendix 1, the treatment effect in the subpopulation identified by $X > x^*$ is

$$\begin{aligned}\delta(x^*) = \delta(x^*, \theta_A, \theta_B) &= \mathbb{E}[Y_A | X > x^*] - \mathbb{E}[Y_B | X > x^*] \\ &= \mu_{Y|A} - \mu_{Y|B} + \left(\rho_A \sigma_{Y|A} - \rho_B \sigma_{Y|B}\right) \frac{\phi\left(\frac{x^* - \mu_X}{\sigma_X}\right)}{1 - \Phi\left(\frac{x^* - \mu_X}{\sigma_X}\right)},\end{aligned} \tag{2}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of a standard normal distribution, respectively. Here we assume equal variances $\sigma_{Y|A} = \sigma_{Y|B}$ and since X is a predictive biomarker, the term $\rho_A\sigma_{Y|A} - \rho_B\sigma_{Y|B} > 0$. Let us now define $z^* = (x^* - \mu_X)/\sigma_X$, and $\lambda(z^*) = \phi(z^*)/\{1 - \Phi(z^*)\} = \mathbb{E}[Z|Z > z^*]$. Note that $\lambda(z^*)$ it is the mean of a standard normal random variable $Z = (X - \mu_X)/\sigma_X$ truncated at z^* (or the inverse Mill's Ratio) and can be interpreted as the normal hazard function of z^*. Thus $\lambda(z)$ is an increasing function of z and $\lambda(z) \geqslant z$ holds for any $z \in \mathbb{R}$. Analogously, the treatment effect in the subpopulation defined by the threshold z^* is

$$\begin{aligned} \delta(z^*) = \delta(z^*, \theta_A, \theta_B) &= \mathbb{E}[Y_A|Z > z^*] - \mathbb{E}[Y_B|Z > z^*] \\ &= \mu_{Y|A} - \mu_{Y|B} + \left(\rho_A\sigma_{Y|A} - \rho_B\sigma_{Y|B}\right)\lambda(z^*). \end{aligned} \tag{3}$$

Since $\delta(z)$ depends on z only through the function $\lambda(\cdot)$ and $\rho_A > \rho_B$ holds, $\delta(z)$ is an increasing function of z. Then finding $\min\{z \in \mathbb{R} \text{ s.t. } \delta(z) = \delta(z, \theta_A, \theta_B) \geqslant c\}$ is equivalent to finding

$$z^* \in \mathbb{R} \text{ s.t. } \delta(z^*) = \delta(z^*, \theta_A, \theta_B) = c, \tag{4}$$

or

$$z^* \in \mathbb{R} \text{ s.t. } \lambda(z^*) = \frac{c - (\mu_{Y|A} - \mu_{Y|B})}{\rho_A\sigma_{Y|A} - \rho_B\sigma_{Y|B}}. \tag{5}$$

Therefore, the problem can be restated in terms of finding the truncation point z^* of a standard normal distribution such that the mean of the truncated distribution is equal to $[c - (\mu_{Y|A} - \mu_{Y|B})]/(\rho_A\sigma_{Y|A} - \rho_B\sigma_{Y|B})$. As c increases, $\lambda(\cdot)$ increases and so z^* increases.

The minimum treatment effect used to select the biomarker threshold at the end of stage 1, namely c, can be thought of as a sum of two components. We can define $c = d + \Delta$ where d is the minimal clinical significance difference elicited by the investigator and Δ is the "additional" size effect required in order to obtain a conclusive result at the end of stage 2. While d is set according to clinical considerations, Δ is set according to the sample size that can be afforded in stage 2. In practice, there is a trade-off between large Δ and small

n_2. Indeed, a large Δ implies a large value for the biomarker threshold z^* that can make $P(Z > z^*)$ so small that it becomes meaningless to continue the trial to stage 2 on a too small portion of the population.

The behavior of the selected threshold z^* under several sets of model parameters with $\sigma_{Y|A} = \sigma_{Y|B} = 1.5$ is displayed in Figure 1. In particular, the left panel displays z^* for increasing values of the difference between c and $\mu_{Y|A} - \mu_{Y|B}$. For a fixed value of this difference, to achieve a treatment effect in the subpopulation at least equal to c, the threshold z^* will be larger when ρ_A is close to ρ_B or the biomarker is not very predictive. Analogously, the right panel shows the behavior of z^* for increasing $\rho_A - \rho_B$ and several values of $c - (\mu_{Y|A} - \mu_{Y|B})$.

[Figure 1 about here.]

2.2 *Analysis after stage 1 and the choice of the biomarker threshold*

In stage 1 patients are recruited from the full population under model (1). By randomizing n_1 patients with balanced allocation, at the end of stage 1 we estimate the maximum likelihood estimators (MLEs) $\hat{\theta}_A = (\hat{\mu}_{Y|A},\, \hat{\rho}_A, \hat{\mu}_{X|A}, \hat{\sigma}^2_{X|A})$ based on half the data and $\hat{\theta}_B = (\hat{\mu}_{Y|B}, \hat{\rho}_B, \hat{\mu}_{X|B}, \hat{\sigma}^2_{X|B})$ based on half the data, where $\hat{\mu}_{X|A}$ and $\hat{\mu}_{X|B}$ are MLEs of μ_X and $\hat{\sigma}^2_{X|A}$ and $\hat{\sigma}^2_{X|B}$ are the MLEs of σ^2_X computed on group A and B, respectively. Since we work under the assumption that the biomarker distribution does not depend on the treatment assignment, we can combine data from group A and B and compute pooled estimators $\hat{\mu}_X = (\hat{\mu}_{X|A} + \hat{\mu}_{X|B})/2$ and $\hat{\sigma}^2_X = (\hat{\sigma}^2_{X|A} + \hat{\sigma}^2_{X|B})/2$.

For any z, the stage 1 MLEs for the treatment effect in the subpopulation defined by the biomarker threshold can be derived by substituting the model parameter estimates obtained with the data from the first stage. We can estimate $\delta(z) = \delta(z, \theta_A, \theta_B)$ by $\hat{\delta}^{(1)}(z) = \delta(z, \hat{\theta}_A, \hat{\theta}_B) = \hat{\mu}_{Y|A} - \hat{\mu}_{Y|B} + \left(\hat{\rho}_A \sigma_{Y|A} - \hat{\rho}_B \sigma_{Y|B}\right) \lambda(z)$, where $z = (x - \hat{\mu}_X)/\hat{\sigma}_X$. The true biomarker threshold z^* is unknown, but its estimate $\hat{z}^*$ can be obtained by replacing

$\delta(z)$ with $\hat{\delta}^{(1)}(z)$ in (4) or model parameters' estimates in (5), having at the end of stage 1, $\hat{\delta}^{(1)}(\hat{z}^*) = \hat{\delta}(\hat{z}^*, \hat{\theta}_A, \hat{\theta}_B) = c$ and

$$\lambda(\hat{z}^*) = \frac{c - (\hat{\mu}_{Y|A} - \hat{\mu}_{Y|B})}{\hat{\rho}_A \sigma_{Y|A} - \hat{\rho}_B \sigma_{Y|B}}. \tag{6}$$

From a practical viewpoint it may be not reasonable to look for the minimum z s.t. $\delta(z) \geqslant c$ over the whole real line. On the one hand, it could be useless to select a threshold such that the chance of recruiting patients at stage 2 is very low and the subpopulation that can benefit from the treatment is too small. Clearly this is a disease-specific issue. For instance, in a clinical trial for a rare/grave disease, even if a small population benefits, it may be worth enriching and continuing the trial. On the other hand, a very low threshold, meaning that the benefiting group is almost the whole population, may lead to the decision of not enriching. However, if the drug under study has non-negligible side effects, it may be worthwhile enriching to not expose patients who will not benefit. As discussed in Simon (2015), these decisions are driven by ethical and cost considerations. Indeed, the financial cost of evaluating the biomarker needs to be taken into account in both the small and the large benefiting population scenarios and can eventually influence the decision about whether or not to enrich (Stallard et al., 2014). Our modelling of the adaptive enrichment problem allows us to directly assess the size of the population in terms of $P(Z > z^*) = \Phi(-z^*)$ so that, for example, we could stop the trial after stage 1 if $\Phi(-z^*) \leqslant \xi$, where ξ can be set according to the above considerations on the subpopulation size.

The estimated threshold $\hat{z}^*$ is a random function of the data from stage 1 because it depends on the estimates on the model parameters $\hat{\theta}_A, \hat{\theta}_B$ that, in turn, depend on n_1: thus the distribution of $\hat{z}^*$ depends on n_1 (see also Section 4.4).

2.3 *Stage 2*

Once a biomarker threshold $\hat{z}^*$ (or equivalently $\hat{x}^* = \hat{\sigma}_X \hat{z}^* + \hat{\mu}_X$) has been estimated after stage 1, then only patients with $Z > \hat{z}^*$ are recruited in stage 2. The subject observations in stage 2 arise from a truncated bivariate normal distribution $[TbN(\cdot)]$,

$$\begin{aligned}(Y_A, Z)|(Z > \hat{z}^*) &\sim TbN(\mu_{Y_A|Z>\hat{z}^*}, \mu_{Z|Z>\hat{z}^*}, \sigma_{Y_A|Z>\hat{z}^*}, \sigma_{Z|Z>\hat{z}^*}, \rho_{A|Z>\hat{z}^*}),\\ (Y_B, Z)|(Z > \hat{z}^*) &\sim TbN(\mu_{Y_B|Z>\hat{z}^*}, \mu_{Z|Z>\hat{z}^*}, \sigma_{Y_B,Z>\hat{z}^*}, \sigma_{Z|X>\hat{z}^*}, \rho_{B|Z>\hat{z}^*})\end{aligned} \tag{7}$$

where the expression for all the parameters can be obtained from the theory of the truncated bivariate normal distribution. Recalling that $\lambda(\hat{z}^*)$ is estimated at stage 1 [eq. (6)] and then fixed for stage 2, then

$$\mu_{Y_A,Z|Z>\hat{z}^*} = \mathbb{E}[Y_A|Z > \hat{z}^*] = \mu_{Y|A} + \sigma_{Y|A}\rho_A\lambda(\hat{z}^*);$$

$$\mu_{Z|Z>\hat{z}^*} = \mathbb{E}[Z|Z > \hat{z}^*] = \lambda(\hat{z}^*);$$

$$\sigma^2_{Y_A|Z>\hat{z}^*} = \mathbb{V}[Y_A|Z > \hat{z}^*] = \sigma^2_{Y|A}\{1 - \rho^2_A\lambda(\hat{z}^*)[\lambda(\hat{z}^*) - \hat{z}^*]\};$$

$$\sigma^2_{Z|Z>\hat{z}^*} = \mathbb{V}[Z|Z > \hat{z}^*] = 1 - \lambda(\hat{z}^*)[\lambda(\hat{z}^*) - \hat{z}^*];$$

$$\rho_{A|Z>\hat{z}^*} = \rho_A\sqrt{\sigma^2_{Z|Z>\hat{z}^*}/\sigma^2_{Y_A|Z>\hat{z}^*}}.$$

Analogous forms for treatment B are obvious. Since $\lambda(z)$ is an increasing function of z and $\lambda(z) \geqslant z$, $\forall z$, then $\mu_{Y_A|Z>\hat{z}^*}$ and $\mu_{Z|Z>\hat{z}^*}$ are increasing in z^* while $\sigma^2_{Y_A|Z>\hat{z}^*}$ and $\sigma^2_{Z|Z>\hat{z}^*}$ are decreasing in z^*. Notice that $\sigma_{Z|Z>\hat{z}^*} > \sigma_{Y_A|Z>\hat{z}^*}$, so $|\rho_{A|Z>\hat{z}^*}| \leqslant |\rho_A|$. Note also that even if $\sigma_{Y|A} = \sigma_{Y|B}$, then at stage 2 we have $\sigma_{Y_A|Z>\hat{z}^*} \neq \sigma_{Y_B|Z>\hat{z}^*}$.

In treatment group A (analogous conclusions hold also for group B), the joint density of the truncated bivariate normal is $f_{(Y_A,Z>\hat{z}^*)}(y, z) = f_{(Y_A,Z)}(y, z)/\Phi(-\hat{z}^*)$ with support $-\infty < y < +\infty$ and $\hat{z}^* \leqslant z \leqslant +\infty$; where $f_{(Y_A,Z)}(\cdot)$ is the joint normal density of (Y_A, Z). Clearly, the marginal distribution of Z has support $(\hat{z}^*, +\infty)$ but the marginal distribution of Y_A, say $f_{Y_A|Z>\hat{z}^*}(y)$, has still support $(-\infty, +\infty)$ due to the single truncation on the bivariate

normal distribution. In particular,

$$\begin{aligned} f_{Y_A|Z>\hat{z}^*}(y) &= \int_{\hat{z}^*}^{+\infty} f_{(Y_A,Z)}(y,z)dz/\Phi(-\hat{z}^*) \\ &= \frac{f_{Y_A}(y)}{\Phi(-\hat{z}^*)} \int_{\hat{z}^*}^{+\infty} f_{Z|Y_A}(z|y)dz \\ &= \frac{f_{Y_A}(y)}{\Phi(-\hat{z}^*)} \Phi\left(-\frac{\hat{z}^* - \frac{\rho_A}{\sigma_{Y|A}}(y-\mu_{Y|A})}{\sqrt{1-\rho_A^2}} \right), \end{aligned}$$

where $f_{Y_A}(y)$ is the pdf of $Y_A \sim N(\mu_{Y|A}, \sigma^2_{Y|A})$. Note that the marginal distributions of the truncated bivariate normal are, in general, not truncated normal distributions (Horrace, 2005). Now, $f_{Y_A|Z>\hat{z}^*}(y)$ is skewed due to the term $\Phi\left(-\left[\hat{z}^* - \frac{\rho_A}{\sigma_{Y|A}}\right]/[y-\mu_{Y|A}]\sqrt{1-\rho_A^2}\right)$ which depends on y, but all else being equal, the distribution is more skewed for large values of $\rho_A/\sigma_{Y|A}$ since the term $\Phi(\cdot)$ tends to be constant with respect to y.

At stage 2, n_2 patients are recruited and assigned to treatment A and B with balanced allocation. The treatment effect at the end of the trial is $\delta^{(2)}(\hat{z}^*) = \mathbb{E}[Y_A|Z > \hat{z}^*] - \mathbb{E}[Y_B|Z > \hat{z}^*] = \mu_{Y_A|Z>\hat{z}^*} - \mu_{Y_B|Z>\hat{z}^*}$. The magnitude of the treatment effect in the target population increases with z^*. Clearly $\delta^{(2)}(\hat{z}^*)$ strongly depends on $\rho_A - \rho_B$. For the same value of the threshold, the treatment effect is much larger when the response on the experimental treatment is strongly correlated with the biomarker and the differences become more pronounced for increasing $\hat{z}^*$ (see also Web Figure 1).

Let $\hat{\mu}_{Y_A|Z>\hat{z}^*}$ and $\hat{\mu}_{Y_B|Z>\hat{z}^*}$ be the sample means and $\hat{\sigma}^2_{Y_A|Z>\hat{z}^*}$ and $\hat{\sigma}^2_{Y_B|Z>\hat{z}^*}$ the sample variances of patients' responses on treatment group A and B computed on stage 2 data. Then, $\hat{\delta}^{(2)}(\hat{z}^*) = \hat{\mu}_{Y_A|Z>\hat{z}^*} - \hat{\mu}_{Y_B|Z>\hat{z}^*}$ is an unbiased estimator for $\delta^{(2)}(z)$ with asymptotic variance estimated by $\hat{\mathbb{V}}[\hat{\delta}^{(2)}(\hat{z}^*)] = 2\left(\hat{\sigma}^2_{Y_A|Z>\hat{z}^*} + \hat{\sigma}^2_{Y_B|Z>\hat{z}^*}\right)/n_2$.

At the end of the trial, the objective is to test if the treatment effect in the selected subpopulation meets the minimal clinical significance difference d. More specifically, let us consider

$$\begin{aligned} H_0 &: \mu_{Y_A|Z>z^*} - \mu_{Y_B|Z>z^*} \leqslant d; \\ H_1 &: \mu_{Y_A|Z>z^*} - \mu_{Y_B|Z>z^*} > d, \end{aligned} \tag{8}$$

which can be tested with the statistic $T_n = \{\hat{\mathbb{V}}[\hat{\delta}^{(2)}(\hat{z}^*)]\}^{-1/2}(\hat{\mu}_{Y_A|Z>\hat{z}^*} - \hat{\mu}_{Y_B|Z>\hat{z}^*} - d)$.

Under the null hypothesis, T_n is asymptotically distributed as a standard normal random variable. If the null hypothesis is rejected we can conclude that the treatment effect in the subpopulation defined by the biomarker threshold $\hat{z}^*$ is greater than the minimal clinical significance difference d. Moreover, the patients group identified by $Z > \hat{z}^*$ is the subpopulation benefiting from the new treatment.

Simon and Simon (2013) describe the failure of randomization tests in their adaptive enrichment design. However, since our design analyzes only subjects in the subpopulation, this is not an issue. While we have assumed parametric models throughout the paper, a re-randomization analysis (Rosenberger et al., 2019) is appropriate and valid for the second stage data. Alternatively, we could extend the Lin et al. (2021) approach to two treatments and use the data from both stages in the inference by use of their asymptotic techniques.

3. Sample size determination for the two stage enrichment design

The procedure to calculate the sample size n_1 and n_2 requires initial guesses of $\mu_{Y|A} - \mu_{Y|B}$, and $\rho_A - \rho_B$, as well as a clinically significant value d. Note that, in this paper we take $\sigma_{Y|A}$ and $\sigma_{Y|B}$ to be known. The approach could be easily extended when the response variances are unknown and, in this case, initial guesses for $\sigma_{Y|A}$ and $\sigma_{Y|B}$ are also required. The following steps detail the sample size calculations.

(1) The initial values of model parameter given by the investigator are used to compute an initial estimate of the biomarker threshold by (5), denoted by $z^*_{(0)}$, that is such that $\delta(z^*_{(0)}) = c = d + \Delta$. Compute $\lambda(z^*_{(0)})$ and then the consequent values of the parameters of the truncated distributions in (7).

(2) Let $n_{2,(0)}$ be an initial value for the sample size at stage 2. Given the type I error rate

α, the sample size required at stage 2 to achieve a power $1-\beta$ is given by

$$n_{2,(0)} = (z_{1-\alpha} + z_{1-\beta})^2 \frac{2(\sigma^2_{Y_A|Z>z^*_{(0)}} + \sigma^2_{Y_B|Z>\hat{z}^*_{(0)}})}{(\mu_{Y_A|Z>z^*_{(0)}} - \mu_{Y_B|Z>z^*_{(0)}} - d)^2}. \quad (9)$$

Now, $n_{2,(0)}$ depends on the parameters of the truncated distributions that depend on $z^*_{(0)}$ that, in turn, depends on c. Since $c = d + \Delta$, but d is a minimal clinical significance difference given by the clinicians, we have that $n_{2,(0)}$ depends on Δ. The behavior of $n_{2,(0)}$ as $z^*_{(0)}$ varies, namely as the parameters of the truncated distributions and Δ vary is presented in Web Figure 2. In particular, if we consider the parameters of the truncated distributions fixed, an increasing $z^*_{(0)}$ is the result of a greater value of Δ.

Steps (1) and (2) are repeated for many candidate values for Δ and a pair Δ and $n_{2,(0)}$ can be selected according to the number of patients that the investigator thinks can be enrolled in stage 2.

(3) Once $n_{2,(0)}$ and Δ (and hence c) are determined, an approximation of the distribution of the estimated threshold can be obtained by simulating stage 1 as follows. Simulate N trials and compute $\hat{z}^*_h$ with (5) for $h = 1, \ldots, N$. Choose n_1 such that the mean squared error (MSE) of the estimated biomarker threshold does not exceed an upper bound γ, namely such that $\text{MSE}(\hat{z}^*) = \frac{1}{N}\sum_{h=1}^{N}(\hat{z}^*_h - z^*_{(0)})^2 \leqslant \gamma$. Note that considerations on the choice of γ may also come from Web Figure 1. When the experimental treatment outcome is weakly correlated with the biomarker, $\delta^{(2)}(z^*)$ increases slowly with z^*. However, in this case, the effect of a poor estimate of the threshold at the end of stage 1 is lower than in the case of high differences $\rho_A - \rho_B$. As discussed in Section 2.2, when $\Phi(-z^*)$ is smaller than an upper bound ξ, the trial could be stopped because of the size of the subpopulation. Thus, it is possible to compute the simulated average probability of incorrectly stopping the trial after stage 1, denoted p in the following, and use it as a further criterion to determine n_1.

(4) Once n_1 is selected, stage 1 can start and, at the end, $\hat{\theta}_A$ and $\hat{\theta}_B$ are available as well as

$\hat{z}^*$. Now the first estimate of the sample size at stage 2 $n_{2,(0)}$ can be updated with the current information accrued so far. A re-estimation for the sample size at stage 2 can be obtained by using the data from stage 1 and replacing the current parameters' estimate in equation (9), obtaining n_2.

4. Simulation study

In this Section, the proposed procedure is implemented under several scenarios. In particular, we focus on the experimental set up in which no treatment effect exists in the whole population by setting $\mu_{Y|A} = \mu_{Y|B} = 1$. Moreover, $\sigma^2_{Y|A} = \sigma^2_{Y|B} = 1$, $\mu_X = 0$, $\sigma_X = 1$. Several configurations of ρ_A and ρ_B are taken into account and we referred to this as setting I, II, III indicating that $\rho_B = 0.1, 0.2, 0.3$ respectively and the letters A, B, C denote $\rho_A - \rho_B = 0.3, 0.4, 0.5$ respectively.

Regarding the minimal clinical significant difference, we set $d = 0$ so that $c = \Delta$. All the simulations of this Section refer to 10000 Monte Carlo iterations. The remainder of the Section is organized as follows. The pre-experimental choice of n_2 is in Section 4.1, the simulation results for the choice of n_1 are in Section 4.2 while those regarding stage 2 are in Section 4.3. Further insights on the effect of n_1 on the estimated threshold are given in Section 4.4. Finally, the results on the effect of model misspecification are reported in Section 4.6.

4.1 *Initial choice of* n_2

Steps (1) and (2) described in Section 3 are performed for several values of Δ. As Δ and so c increase, the biomarker threshold becomes larger and so the size of the subpopulation smaller. (In Web Table 1, we report $\hat{z}^*_{(0)}$ with $\Phi(-\hat{z}^*_{(0)})$ in brackets for increasing Δ). For each Δ, the corresponding $n_{2,(0)}$, computed by (9) with $\alpha = 0.05$ and $\beta = 0.2$ are reported in Table 1 in several experimental settings, corresponding to different configurations of ρ_A

and ρ_B. A more restrictive threshold in stage 1 results in a smaller sample size at stage 2 to achieve a given power of the test (8), but at the same time the subpopulation cannot be too small. In general, the sample size at stage 2 decreases as Δ increases but its behavior depends on both the difference and the magnitude of the correlation coefficients. Indeed, when $\rho_A - \rho_B = 0.3$ (Setting A), Δ being equal, higher values of ρ_A and ρ_B are associated to slightly smaller sample sizes. Clearly, the same choice of Δ (e.g. comparing IA-IC with IIIA-IIIC), leads to the same $z^*_{(0)}$ when $\rho_A - \rho_B$ is constant. Nonetheless $n_{2,(0)}$ is smaller for larger magnitude of the ρ_A and ρ_B. For the simulations of this Section we set $\Delta = 0.3$ and so the first estimates of the stage 2 sample sizes are given by the corresponding $n_{2,(0)}$ in Table 1.

[Table 1 about here.]

4.2 *Choice of* n_1

After setting $n_{2,(0)}$ and c, the selection of n_1 is performed, that is Step (3). Under the initial configuration of model parameters provided by the investigator, stage 1 has been simulated 10000 times for each value of $n_1 \in [80, 500]$ and, in each hypothetical trial, the threshold to select the target population is computed. At the end, the average of the squared distance between between z^*_h and $z^*_{(0)}$ is the simulated average MSE. According to the decision making at the end of stage 1 we could have different designs as follows.

- Design ($D1$): the trial moves to the second stage whatever the result of stage 1 is. In this case an upper bound on the threshold and on n_2 are required to avoid degenerate simulations. For the simulation study when the threshold is greater than $z^*_{\max} = \Phi^{-1}(0.95) \approx 1.64$, the threshold is set equal to $z^*_{\max}$; when the re-estimated $n_2 > 1000$ it is set equal to 1000.
- Design ($D2$): when $\hat{z}^*$ is greater than a given upper bound the trial is stopped after stage 1 for futility. As discussed in Section 2.2, in our framework the upper bound on the threshold

is directly related to the size of the subpopulation. For the simulations of $(D2)$ when $\hat{z}^*$ is such that $\Phi(-\hat{z}^*) < \xi = 0.1$ the trial fails to proceed to stage 2.

For increasing n_1, the behavior of MSE($\hat{z}^*$) is reported in Figure 2 for the designs $(D1)$ and $(D2)$. The MSE is higher when the difference among the correlations in the two groups is smaller (solid lines) while it takes much smaller values when $\rho_A - \rho_B = 0.5$. Notice that cases I and II are not reported since the line were essentially superimposed. The upper bound on the MSE($\hat{z}^*$), namely γ, drives the selection of n_1. The choice of γ depends on the clinical trial. In this simulation study we set $\gamma = 0.5$ (light gray dashed line in Figure 2). In addition, in the case of $(D2)$, the relative frequency of incorrect early stopping of the trial can be computed. The results for the simulated averages probabilities p of incorrect stopping the study are displayed in Web Figure 3. When $\rho_A - \rho_B = 0.3$ such probability is quite high even for moderate n_1. It drastically decreases for larger $\rho_A - \rho_B$. Moreover, even when this difference is kept constant there exists an effect of the magnitude of ρ. By comparing the black solid line with the gray ones, higher values of ρ are associated with a smaller probability of reaching the wrong conclusion. In general, the choice of n_1 strongly influences such a probability: this can be taken as a combined criterion together with the one based on the MSE for sample size selection. Clearly as the upper bound on the subpopulation size ξ increases (see Section 2.2), the probability of wrongly stopping the trial early, p, increases as well. We investigate more on the effects of several choice of n_1 in both designs $(D1)$ and $(D2)$ in Section 4.4.

[Figure 2 about here.]

4.3 *Re-estimation of n_2 and stage 2*

Once n_1 is computed, stage 1 can start and, at the end, n_2 is re-estimated on the basis of the model parameters estimates and the computed threshold.

The results for the design $(D1)$ are reported in Table 2. When $\rho_A - \rho_B = 0.3$, a much larger

n_1 is needed to achieve the desired MSE on the estimated threshold and a slightly larger n_2 with respect to the initially planned value. In addition, since under design $(D1)$ whatever the results of stage 1, the trial is taken to the second stage, the standard deviation computed on n_2 is large, with this effect being more pronounced when $\rho_A = 0.4$. For larger $\rho_A - \rho_B$ (settings B and C) the values of n_2 are very close to the ones obtained pre-experimentally. We compute the measure of the power of the test at the end of the second stage: $pow_{(MC)}$ is the simulated average power computed as the relative frequency of the simulations in which the observed test statistic is greater than $z_{1-\alpha}$, pow is the average power and $sd(pow)$ its standard deviation. The power achieves the desired level in all the considered scenarios with a decreasing standard deviation from scenario A to C.

For $(D2)$, Table 3 provides a summary of the simulation study. In this case, the MSE$(\hat{z}^*)$ is computed only on the thresholds that will be actually used, namely only when $\hat{z}^*$ is less than the upper bound (with $\xi = 0.1$, $\hat{z}^*$ is close to 1.28); this leads to smaller n_1 with respect to the n_1 obtained from design $(D1)$, especially when $\rho_A - \rho_B = 0.3$. In general, as this difference increases, both n_1 and p decrease (recalling that p is the probability of incorrectly the trial). Clearly in this context it is crucial to check the magnitude of p. In the case of setting A, these are quite large, while they are always smaller than 0.044 in scenarios B and C. Regarding the reported n_2 and the power, these are clearly computed only on those trials that proceed to stage 2. The re-estimated sample size is closer to the planned one with respect to the case of design $(D1)$, with also a smaller variability. Moreover, n_2 decreases when the magnitude of ρ and so the predictive ability of the biomarker increases, with a slight inflation in the standard deviation. Again, the power is at the desired level 0.8 with low variability, confirming the correct re-estimation of n_2.

[Table 2 about here.]

[Table 3 about here.]

4.4 *The effect of* n_1

Now we further investigate the performance of the two stage enrichment design as n_1 changes. Figures 2 and Web Figure 3 highlight how delicate the choice of n_1 is and its effect on the estimation of the biomarker threshold and the consequences of a bad choice of $\hat{z}^*$. We consider both the designs $(D1)$ and $(D2)$ by taking a value of n_1 around half of those selected in Tables 2 and 3: the results are reported in Web Table 3 and Web Table 4, respectively. For the design $(D1)$, besides a much higher MSE$(\hat{z}^*)$, the re-estimated n_2 is generally larger than the one in Tables 1 and 2 (especially in setting A), and the effect of the small n_1 strongly affects $sd(n_2)$. It is worth recalling that $n_2 = 1000$ has been chosen as an upper bound for the stage 2 sample size in this design. In addition, while the average power (both $pow_{(MC)}$ and pow) presents a small deflation, the corresponding standard deviations are around twice the ones obtained with a larger n_1. For design $(D2)$, a too small n_1 strongly increases the probability of incorrectly stopping the study after stage 1, reaching around 20% in setting A, and it is more than doubled in the remaining scenarios. For the design $(D2)$ the values of n_2 fluctuate around $n_{2,(0)}$ but the standard deviations vary from 19 to around 29. Finally, Figure 3 further highlights the crucial role of n_1 in the estimation of the biomarker threshold, showing a distribution more concentrated around the true value as more patients are enrolled in stage 1. In conclusion, a too small stage 1 may lead to a bad estimate of the threshold, which can strongly affect the remainder of the trial, whether it results in a incorrect early stopping or in a less reliable stage 2.

[Figure 3 about here.]

4.5 *Type I error rate control*

The simulation study of this section has the objective of assessing the performance of the proposed two stage adaptive design in terms of the type I error rate for the test in (8). We consider stage 1 in which we set $d = 0.3$ and $\Delta = 0$ (instead of $d = 0$ as in the previous

Section). Thus, the threshold $\hat{z}^*$ is chosen in such a way that $\hat{\delta}^{(1)}(z) = d$. In both cases though, c has the same value and so the selected biomarker thresholds coincide with those calculated previously but in test at the end of stage 2, we are under H_0. Then, we calculate the simulated average type I error ($\alpha^{*}{}_{(MC)}$) for both designs ($D1$) and ($D2$), computed as the relative frequency of the occurrence in which the test statistic is greater than $z_{1-\alpha}$. The results are reported in Web Table 2 and confirm that under the proposed adaptive enrichment design the type I error rate is controlled by appropriately choosing the sample sizes of both stages. The inherent variability in estimating the threshold in stage 1 is incorporated into the null distribution of the final test statistics in our simulations and we don't observe inflation of the type I error rate in Web Table 2. However, since n_2 is clearly based on the distribution of the estimated threshold we were also interested in exploring the behavior type I error rate when stage 1 has a smaller size so, in the setting of Web Table 3 and 4, we report in the last column of these tables $\alpha^*_{(MC)}$ (which was obtained by setting $d = 0.3$ and $\Delta = 0$). For both designs ($D1$) and ($D2$) with halved n_1, the larger variability in estimating the threshold, results in wider fluctuations of the type I error rate around the nominal level.

4.6 *Model misspecification*

The aim of this Section is to check the robustness of the proposed procedure when the true model is not bivariate normal. In particular, the effect of the departure from symmetry has been taken into account assuming a bivariate skew normal distribution as a data generating model. Thus, instead of the model in (1), assume for treatment group A,

$$(Y_A, X)^T = (\mu_{Y|A}, \mu_X)^T + (U_{1A}, U_{2A})^T$$

with $(U_{1A}, U_{2A})^T$ has a bivariate-skew normal distribution,

$$f_{(U_{1A},U_{2A})}(u, v) = 2f(u, v)\Phi(a_1u + a_2v),$$

where $f(u, v)$ is bivariate normal density with zero mean and variance-covariance matrix equal to the one of $(Y_A, Z)^T$ in model (1) (Azzalini and Dalla Valle, 1996). The parameters a_1 and a_2 allow different shapes for the bivariate density (when they are both zero, there is no skewness). Clearly analogous reasoning leads to the model for treatment B.

As in the other simulations of this Section we set $\mu_{Y|A} = \mu_{Y|B} = 1$, $\sigma^2_{Y|A} = \sigma^2_{Y|B} = 1$, $\mu_X = 0$, $\sigma_X = 1$. We adopt the design $(D2)$ under scenarios IIIA, IIIB and IIIC ($\rho_A = 0.6, 0.7, 0.8$ respectively and $\rho_B = 0.3$) and we study the procedure's performance when the model is misspecified. The pre-experimental calculations, namely when $n_{2,(0)}, \Delta$ and n_1 are set, are not affected by the model misspecification so that the results in Table 1 still hold. For the same reason, the choice of n_1 is the same reported in Table 3. The power and the re-estimated n_2, when the model is misspecified for several configurations of a_1 and a_2 is shown in Web Table 5 where n_2 is the sample size re-estimated at the end of stage 1 using the formula in (9). The model mispecification effects are seen here both in the power and type I error in the case of skewness in X only ($a_1 = 0$), in the case of skewness in Y_A and Y_B only ($a_2 = 0$) and in the case of skewness in both the responses and biomarker distribution. However, it is worth noticing that both the power and the type I error rate depend on n_2 that, according to the two stage adaptive design presented in this paper, is re-estimated at the end of stage 1. Thus, if the experimenter recognizes that the data in stage 1 are not bivariate normal, a further simulation study can be performed to discover how n_2 should be increased in order to obtain the desired power/type I error.

5. Discussion

In this paper, we have attempted to give some insight into the subtleties of enriching a clinical trials population in two stages. While the design we have described is somewhat limited by the bivariate normal model (although we discuss model misspecification), we have specifically shown several important aspects that investigators should consider before

enriching. First, we have shown the direct relationship between the predictive nature of the biomarker and sample size considerations. This was only possible by explicitly characterizing the correlations between the biomarker and outcome for each treatment under a bivariate normal model. Second, we have shown that the validity of the second stage depends largely on the success of the first stage in estimating the threshold. We have chosen a bound on the mean squared error of the threshold estimator to determine the first stage sample size. Under this criterion, we demonstrate the consequence of not achieving the requisite sample size in stage 1. While investigators may balk at expending such resources on a preliminary trial just to estimate a threshold, the cost of not doing so is a much higher likelihood of having a considerably larger second stage with less power. Third, we have provided a complete template for sample size considerations for both stages, including the re-estimation of the sample size in stage 2 at the conclusion of stage 1. We also describe some of the dilemmas in making a decision after stage 1 as to whether to enrich based on the predictive level of the biomarker (measured by the correlation coefficient). In our framework, after stage 1, the investigator has an estimate of the size of the benefiting population. This is a crucial aspect of adaptive enrichment designs from both the ethical and the financial cost viewpoints, as pointed out also by Stallard et al. (2014) and Simon (2015). If the subpopulation is too small or too large to be cost effective, the study might not be taken to the second stage. However, in the case of rare/grave diseases, it may be worth enriching and continuing the trial on a tiny subpopulation or in the case of harmful side effects even if the subpopulation is almost as large as the whole population it may be meaningful identify the subset of benefiting patients.

Given these insights, one might reasonably question whether enrichment is worth the cost. It is clear that unless an investment is made in threshold estimation accuracy, the trial will likely lead to wrong conclusions. (This is analogous to the problem in phase I oncology trials, where a too small sample size may lead to the wrong maximum tolerated dose, thus

jeopardizing future clinical phases.) Clearly the considerations for enrichment will depend on the gravity of the disease, the potential benefit to an appropriate subpopulation and the risk of a negative result if the true subpopulation is not correctly selected. Other practical considerations are discussed in Stallard (2022). Because our design is performed in two stages, the total amount of time it takes to run the trial will generally be increased with respect to standard clinical studies. However, the treatment effect may be attenuated by including non-benefiting patients in the trial leading to a reduced statistical power of such a study. More considerations on the total study duration of adaptive enrichment trial can be found in Simon and Simon (2013) and Simon (2015).

While every attempt should be made to collect complete data, there is always some missing data in any clinical trial. A missing biomarker value or outcome in stage 1 clearly has a different impact depending on the mechanism that generates the missingness, and the accuracy of threshold estimation can be impacted, and consequently the definition of the target population. When a biomarker value is not observed in stage 2, the corresponding patient cannot be enrolled in the study, as he or she fails to meet the eligibility criterion. If the outcome is missing in stage 2, one could use imputation techniques as appropriate.

It is worth noticing that when the covariates are treated as random variables, the bivariate normal model coincides with a regression approach with random regressors when the sample sizes in the two groups are equal, so we could have approached the theoretical problem in that way. In this case, the correlations among the response and the biomarker can be depicted as a function of slopes of the regression line of the response conditional on the biomarker value.

Notice also that there is no reason that the data from both stages cannot be used, if desired, in the final treatment comparison. Several authors have described techniques to do this, including bootstrap methods (Simon and Simon, 2017) or asymptotic inference approaches [e.g., Lin et al. (2021) and Tarima and Flournoy (2021)]. These methods still

rely on the accuracy of the threshold estimator in stage 1, but sample size considerations in stage 2 would require different methods.

Many potential directions for future research have been identified. For instance, we considered a normal bivariate model but the procedure could be extended for trials with binary and survival outcomes. However, in the case of non-normal outcomes, the correlation coefficient is no longer a valid metric of the value of the biomarker as a surrogate for predicting the treatment effect in the subpopulation, although maximum likelihood could still be used for estimation of the threshold if we can characterize the bivariate model. In the case of a binary outcome and continuous biomarker, the threshold estimation problem reduces to a two-group discriminant analysis, and standard methodology can be used.

In addition, the inclusion of multiple biomarkers instead of a single one could make the identification of the target population more accurate. In the presence of multiple correlated normal biomarkers, an extension is possible by using the multivariate normal model, but it becomes difficult to determine how to design stage 2 when there are multiple thresholds that must be satisfied. The most reasonable solution solution is to develop a composite biomarker using dimension reduction techniques for creating propensity score (for instance via principal components or independent component analysis) by analyzing first stage data.

We do not extend our procedure to the case of multiple endpoints, although in principle our techniques should be applicable. Usually when we consider co-primary endpoints, we believe that they are independent or the correlation is low. Therefore, we still can model them separately. However, they may have different predictive biomarkers, so the final target population needs to be defined by multiple thresholds of the predictive biomarkers. The benefiting population is therefore more complicated to determine.

We have only investigated equal allocation design. The optimal allocation problem for two

stage enrichment design could be a very interesting new line of research that, to the best of our knowledge, has not been explored.

Acknowledgements

This work has begun when R. Frieri was a postdoctoral fellow at George Mason University and continued when W.F. Rosenberger was a visiting scholar at University of Bologna. They thank the respective departments for the hospitality. The authors of this paper wish to thank the Co-Editor, Associate Editor, and two referees who made substantial comments that improved the paper.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this paper.

Received Month year. *Revised February* year. *Accepted month* yer.

References

Abrahams, E. and Silver, M. (2009). The case for personalized medicine. *Journal of Diabetes Science and Technology* **3,** 680–684.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83,** 715–726.

Horrace, W. (2005). Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis* **94,** 209–221.

Lin, Z., Flournoy, N., and Rosenberger, W. (2021). Inference for a two-stage enrichment design. *The Annals of Statistics* **49,** 2697–2720.

Posch, M. and Proschan, M. A. (2012). Unplanned adaptations before breaking the blind. *Statistics in Medicine* **31,** 4146–4153.

Rosenberger, W., Uschner, D., and Yanying, W. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine* **38,** 1–12.

Simon, N. (2015). Adaptive enrichment designs: applications and challenges. *Clinical Investigation* **5,** 383–391.

Simon, N. and Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics* **14,** 613–625.

Simon, R. and Simon, N. (2017). Inference for multimarker adaptive enrichment trials. *Statistics in medicine* **36,** 4083–4093.

Spencer, A., Harbron, C., Mander, A., Wason, J., and Peersa, I. (2016). An adaptive design for updating the threshold value of a continuous biomarker. *Statistics in Medicine* **35,** 4909–4923.

Stallard, N. (2022). Adaptive enrichment designs with a continuous biomarker (with discussion). *Biometrics* pages 1–11.

Stallard, N., Hamborg, T., Parsons, N., and Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics* **24,** 168–187.

Tarima, S. and Flournoy, N. (2021). Choosing interim sample sizes in group sequential designs. *Studies in Health Technology and Informatics* **24,** 11–16.

Wang, S. J., Hung, H. M. J., and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* **51,** 358–374.

Zhao, Y. Q. and LeBlanc, M. L. (2020). Designing precision medicine trials to yield a greater population impact. *Biometrics* **76,** 643–653.

Supporting Information

The derivation of equation (2), Web Tables and Figures referenced in Sections 2.3, 3 and 4 are available at the Biometrics website on Wiley Online Library.

In addition, the `R` code to implement the design can be found online as supporting information.

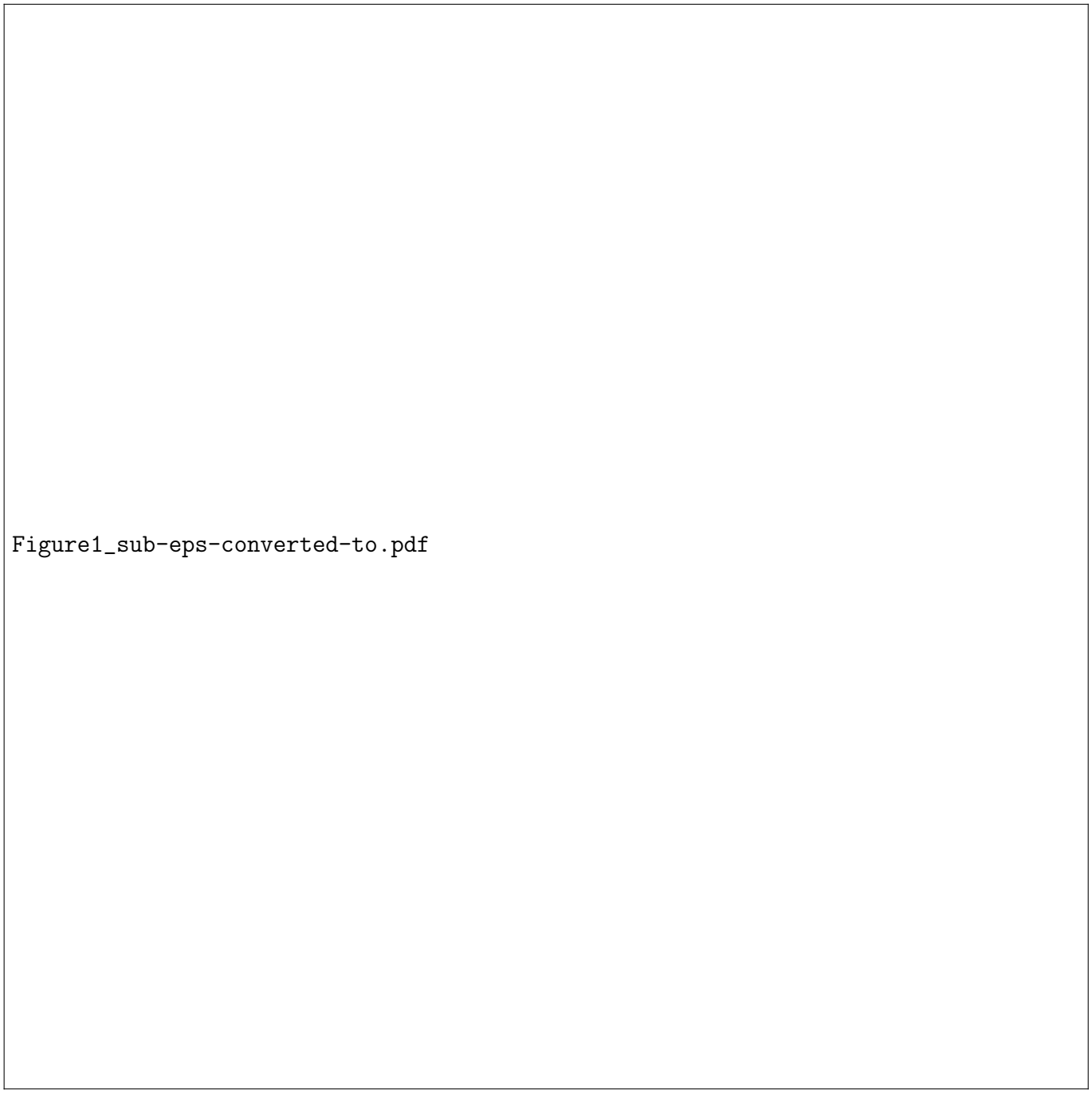

Figure 1. Selected threshold for $\sigma_{Y|A} = \sigma_{Y|B} = 1.5$.

Figure 2. $MSE(\hat{z}^*)$ for increasing n_1 in scenarios IIIA-IIIC, namely for $\rho_B = 0.3$ and $\rho_A = 0.6, 0.7, 0.8$, respectively. The horizontal line is in correspondence of $\gamma = 0.5$. In Design $(D1)$: stage 2 is performed regardless of the result of stage 1; in Design $(D2)$ the trial is stopped for futility after stage 1 when $\Phi(\hat{z}^*) < 0.1$.

Figure 3. Density of the estimated threshold for Design ($D2$): the trial is stopped for futility after stage 1 when $\Phi(\hat{z}^*) < 0.1$ and in Scenario IIIC: $\rho_A = 0.8, \rho_B = 0.3$. The true threshold is -0.33 (light gray dashed line).

Table 1

Values of initial guesses of stage 2 sample size $n_{2,(0)}$ according to equation (9) as Δ varies with $\mu_{Y|A} = \mu_{Y|B} = 1$, $\sigma^2_{Y|A} = \sigma^2_{Y|B} = 1$, $\mu_X = 0$, $\sigma_X = 1$, $d = 0$, $\alpha = 0.05$, $\beta = 0.2$.

Setting			Δ						
			0.20	0.25	0.30	0.35	0.40	0.45	0.50
	ρ_A	ρ_B				$n_{2,(0)}$			
IA	0.4	0.1	587	374	258	189	144	114	92
IIA	0.5	0.2	566	358	247	180	137	108	87
IIIA	0.6	0.3	537	338	232	168	128	100	81
	ρ_A	ρ_B				$n_{2,(0)}$			
IB	0.5	0.1	577	366	253	185	141	111	89
IIB	0.6	0.2	555	351	241	175	133	104	84
IIIB	0.7	0.3	527	330	225	163	123	96	77
	ρ_A	ρ_B				$n_{2,(0)}$			
IC	0.6	0.1	567	358	246	179	136	107	86
IIC	0.7	0.2	544	342	234	170	128	100	81
IIIC	0.8	0.3	517	322	219	158	119	92	74

Table 2

Summary table of Design (D1): n_1 obtained by setting $\gamma = 0.5$, n_2 computed by using stage 1 data, $pow_{(MC)}$ is the simulated average power computed as the relative frequency of the occurrence in which the test statistic is greater than $z_{1-\alpha}$, pow is the average power at the end of the study and $sd(\cdot)$ the corresponding standard deviations.

Setting	ρ_A	ρ_B	n_1	n_2	$sd(n_2)$	$pow_{(MC)}$	pow	$sd(pow)$
IA	0.4	0.1	375	281	113.0	0.775	0.774	0.065
IIA	0.5	0.2	360	266	100.9	0.795	0.796	0.060
IIIA	0.6	0.3	345	245	83.0	0.792	0.797	0.052
IB	0.5	0.1	270	257	53.2	0.798	0.799	0.040
IIB	0.6	0.2	260	244	43.1	0.801	0.799	0.040
IIIB	0.7	0.3	255	227	32.5	0.800	0.800	0.037
IC	0.6	0.1	210	246	24.5	0.809	0.800	0.034
IIC	0.7	0.2	205	234	23.6	0.800	0.800	0.035
IIIC	0.8	0.3	200	218	19.3	0.807	0.800	0.035

Table 3

Summary table of Design (D2) : n_1 obtained by setting $\gamma = 0.5$, n_2 computed by using stage 1 data, p simulated average probability of stopping after stage 1, $pow_{(MC)}$ is the simulated average power computed as the relative frequency of the occurrence in which the test statistic is greater than $z_{1-\alpha}$, pow is the average power at the end of the study and $sd(\cdot)$ the corresponding standard deviations.

Setting	ρ_A	ρ_B	n_1	p	n_2	$sd(n_2)$	$pow_{(MC)}$	pow	$sd(pow)$
IA	0.4	0.1	290	0.148	258	7.1	0.794	0.800	0.031
IIA	0.5	0.2	280	0.149	248	9.5	0.808	0.801	0.032
IIIA	0.6	0.3	275	0.127	233	10.6	0.802	0.801	0.033
IB	0.5	0.1	240	0.044	252	9.7	0.792	0.800	0.031
IIB	0.6	0.2	230	0.035	241	11.3	0.798	0.801	0.032
IIIB	0.7	0.3	230	0.026	226	15.6	0.803	0.800	0.034
IC	0.6	0.1	205	0.011	245	11.7	0.806	0.800	0.032
IIC	0.7	0.2	205	0.006	234	15.0	0.797	0.800	0.033
IIIC	0.8	0.3	200	0.003	218	19.2	0.796	0.800	0.036