Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Design and estimation in clinical trials with subpopulation selection

Yi-Da Chiu[a], Franz Koenig[b], Martin Posch[b] and Thomas Jaki[a]*†

Population heterogeneity is frequently observed among patients' treatment responses in clinical trials because of various factors such as clinical background, environmental and genetic factors. Different subpopulations defined by those baseline factors can lead to differences in the benefit or safety profile of a therapeutic intervention. Ignoring heterogeneity between subpopulations can substantially impact on medical practice. One approach to address heterogeneity necessitates designs and analysis of clinical trials with subpopulation selection. Several types of designs have been proposed for different circumstances. In this work we discuss a class of designs that allow selection of a predefined sub-group. Using selection based on the maximum test statistics as the worst case scenario, we then investigate the precision and accuracy of the maximum likelihood estimator (MLE) at the end of the study via simulations. We find that the required sample size is chiefly determined by the subgroup prevalence and show in simulations that the MLE for these designs can be substantially biased. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: bias, enrichment design, maximum likelihood estimator, prevalence, subgroup analysis, subpopulation selection

1. Introduction

Heterogeneity is frequently observed among patients' treatment response in clinical trials. This is due to various factors such as age, race, disease severity or genetic differences. The topic of heterogeneity in treatment effects has received some attention in the literature (e.g. [1, 2, 3]) and graphical methods such as forest plots are routinely use for the purpose of examining heterogeneity in effects (e.g. [4]). Ignoring heterogeneity can substantially impact on medical practice. For example, a treatment might work well in some patients but not in others. Naively estimating the treatment effect across all patients will result in a diluted effect for the group that truly benefits from the treatment. At the same time an ethical issue

[a] *Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, LA1 4YF, U.K.*
[b] *Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.*
** Correspondence to: Thomas Jaki, Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, LA1 4YF, U.K.*

Contract/grant sponsor: This work is independent research arising in part from Dr Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. Funding for this work was also provided by the Medical Research Council (MR/M005755/1). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. All authors have made equal contributions to this manuscript.
† E-mail: t.jaki@lancaster.ac.uk

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

arises due to delivering a treatment to all patients while some might not expect an effect and will potentially be exposed to harmful side-effects. To address these issues, trials that consider (potential) subgroups defined by one or more biomarkers are becoming more popular. In general, a biomarker is some measurable variable that might help to identify distinct groups of patients and some examples include cholesterol levels, genetic variations or age. A biomarker is considered prognostic if it provides information about the value of some other variable of interest (e.g. the primary endpoint of a study) while it is called predictive if it's value yields information about the treatment effect. In this paper we will only consider the latter type of biomarkers.

A number of different designs concerning treatment selection and subgroups within the study populations have been proposed. These designs can be categorized by factors such as design setting (confirmatory or exploratory) or methodology (frequestist, Bayesian or utility/decision function) - see [5, 6, 7]. Additionally, the designs can be categorized into single-stage (fixed sample) designs and multi-stage (adaptive) designs. Both conventionally utilize multiple testing procedures to test for effects in each of the populations of interest. An overview of different multiple testing approaches for this purpose is given in [6] and the references therein. A single-stage design with one biomarker tests, for example, the null hypotheses: the treatment effect of the full population is zero, H_{0F}; and the treatment effect in the subgroup of interest is zero, H_{0S} [5, 8, 9, 10, 11, 12]. These designs are usually employed for exploratory subgroup analysis in phase II (i.e. to identify an interesting subgroup), or for confirmatory subgroup analysis in phase III, examining the treatment benefit of pre-specified subgroups. Corresponding multi-stage designs are constructed either as extensions of group sequential approaches [13] or using combination tests [14]. They can refine the population to either the whole or one or more subgroups at the interim analysis and can allow for early stopping for benefit and lack of benefit (see e.g. [5, 15, 16, 17, 18]).

The accuracy and precision of the treatment effect estimators in subgroup analysis is also crucial to the development of novel treatments and decisions about treatment implementation. Especially, bias is ubiquitous in designs that select (see [19]) and in the designs considered here the bias can come from selecting which (sub)population should be studied further or from selective reporting promising results even in a simple fixed sample design. A variety of papers on treatment effect estimation in the related problem of trials with treatment selection have been published. Approximate bias-correction estimators for single stage designs for normal endpoints are discussed in [20, 21], uniformly minimum variance conditional unbiased estimators (UMVCUE) for two stage designs have been proposed by Cohen and Sackrowitz [22] and further extensions published in [23, 24]. Shrinkage estimators have been discussed in [25] while approaches to construct confidence intervals are described in [26, 27, 28]. Time-to-event endpoints are considered in Brückner et al [29].

In contrast, rather limited literature addresses estimation issues in clinical trials with subpopulation selection. For single-stage designs, Rosenkranz [30] proposed a bias-adjustment method employing bootstrap techniques to calibrate the estimates upon general distributional assumption on outcomes. For multi-stage designs, Kimani et al. [31] proposed two estimators: one is a naive estimator using a weighted average of per-stage means and prevalences for each subgroup; the other is a uniformly minimum variance conditional unbiased estimator (UMVCUE) derived by the Rao-Blackwell theorem. They assessed the performance under several situations, such as different values of prevalence and treatment effect of one subpopulation, and also suggested which estimator should be used according to what population is selected at stage 1. In addition, Magnusson and Turnbull [16] focused on the designs rather than estimation though, they outlined an extended bias-reduction algorithm proposed by Wang and Leung [32] in which uses double bootstrap methods [33] to adjust ML-estimates and build bootstrap confidence interval.

Despite some contributions on estimation, the aforementioned papers do not provide a complete overview of the maximum likelihood estimator (MLE) under various designs and lack exploring the estimator performance in further

conditions. Rosenkranz's [30] simulation work on single-stage designs implicitly regarded the MLE only in circumstances with few different treatment effects for subgroups and thresholds used in the selection rule. Kimani et al. [31] considered two-stage adaptive seamless designs selecting subpopulation based on the stage 1 data but not allowing early stopping, and they only assessed estimators with selection but without reporting promising results. The multi-stage designs of Magusson and Turnbull allow to select multiple subpopulations if the estimates of treatment effects are above certain thresholds at stage 1.

In this paper we discuss a framework to design single and multi-stage design which select subgroups. We illustrate the design properties when selection is based on the maximum statistic and comprehensively evaluate the properties of the MLE for these designs. Note that selecting on the basis of the maximum statistic is the worst case for both type I error (provided that the number of hypothesis remains the same) and bias and hence of particular interest. In Section 2 we derive a subgroup selection design that selects groups based on the maximum test statistic. Section 3 describes a simulation study in which different general design scenarios are evaluated and the bias and MSE of the corresponding maximum likelihood estimators is derived. In Section 4 we remark on the designs with different selection rules, then summarise the results of the simulation study and discuss its implications for future work.

2. Designs

In this section, we first define the basic setting and notation and then provide general ideas for designs with subpopulation selection based on the maximum test statistic.

2.1. Basic Setting and Notation

Assume J mutually disjoint subpopulations are in the full study population (F) and denote the prevalence of the j-th subpopulation (S_j) by λ_j , where $j = 1, \ldots, J$ and $\sum \lambda_j = 1$. The sample size of each subgroup is fixed as a proportion of the total sample size depending on the respective prevalence. We use n_j to denote the sample size in subgroup S_j and more generally use subscripts to denote groups and treatments and superscripts for stages. We consider a normally distributed endpoint with mean $\mu_{j,l}$ with $j = 1, \ldots J$ and $l = T, C$ where subscript T corresponds to the treatment group and C to the control group. Additionally we assume a common variance, σ^2, across subpopulations.

2.1.1. Single stage design For a single-stage design, the test statistics used for selection and decision are distributed as

$$Z_j^{(1)} = I_j^{(1)}\big(\bar{Y}_{j,T}^{(1)} - \bar{Y}_{j,C}^{(1)}\big) \sim \mathcal{N}\big(I_j^{(1)}\theta_j, 1\big).$$

Note that we use the (unnecessary) superscript (1) for consistency with the multi-stage notation used later. $\bar{Y}_{j,T}^{(1)}$ and $\bar{Y}_{j,C}^{(1)}$ are the sample means of the treatment group and of the control group within S_j, respectively. The true treatment difference in S_j is denoted $\theta_j = \mu_{j,T} - \mu_{j,C}$ and $I_j^{(1)} = 1/\big(\sigma\sqrt{1/n_{j,T}^{(1)} + 1/n_{j,C}^{(1)}}\big)$ is the information level for S_j. This further simplifies to $1/\big(2\sigma\sqrt{1/n_j^{(1)}}\big)$ when the assumed treatment allocation ratio is 1:1, where $n_j^{(1)}$ is the total sample size of S_j until the end of stage 1.

Considering a composite population $S_{\mathcal{U}+\mathcal{V}}$ combining two subpopulations $S_{\mathcal{U}}$ and $S_{\mathcal{V}}$ (where $\mathcal{U},\mathcal{V} \subseteq \{1,2,\ldots,J\}$, $\mathcal{U}\cap\mathcal{V}=\varnothing$), the test statistics are distributed as

$$Z^{(1)}_{\mathcal{U}+\mathcal{V}} = \sqrt{\frac{n^{(1)}_{\mathcal{U}}}{n^{(1)}_{\mathcal{U}+\mathcal{V}}}}Z^{(1)}_{\mathcal{U}} + \sqrt{\frac{n^{(1)}_{\mathcal{V}}}{n^{(1)}_{\mathcal{U}+\mathcal{V}}}}Z^{(1)}_{\mathcal{V}} \quad = \quad I^{(1)}_{\mathcal{U}+\mathcal{V}}\big(\bar{Y}^{(1)}_{\mathcal{U}+\mathcal{V},T} - \bar{Y}^{(1)}_{\mathcal{U}+\mathcal{V},C}\big) \sim \mathcal{N}\big(I^{(1)}_{\mathcal{U}+\mathcal{V}}(\mu_{\mathcal{U}+\mathcal{V},T} - \mu_{\mathcal{U}+\mathcal{V},C}),1\big),$$

where $\bar{Y}^{(1)}_{\mathcal{U}+\mathcal{V},T}$ and $\bar{Y}^{(1)}_{\mathcal{U}+\mathcal{V},C}$ are defined as before but the observations are from the combined treatment group and the combined control group of the united subpopulation $S_{\mathcal{U}+\mathcal{V}}$. The true treatment effect size and the information level of $S_{\mathcal{U}+\mathcal{V}}$ are $\theta_{\mathcal{U}+\mathcal{V}} = \mu_{\mathcal{U}+\mathcal{V},T} - \mu_{\mathcal{U}+\mathcal{V},C}$ and $I^{(1)}_{\mathcal{U}+\mathcal{V}} = 1/\big(\sigma\sqrt{1/n^{(1)}_{\mathcal{U}+\mathcal{V},T} + 1/n^{(1)}_{\mathcal{U}+\mathcal{V},C}}\big)$, respectively. $I^{(1)}_{\mathcal{U}+\mathcal{V}}$ is also equal to $1/\big(2\sigma\sqrt{1/(n^{(1)}_{\mathcal{U}} + n^{(1)}_{\mathcal{V}})}\big)$ for equal allocation. Additionally, $\theta_{\mathcal{U}+\mathcal{V}} = (\lambda_{\mathcal{U}}\theta_{\mathcal{U}} + \lambda_{\mathcal{V}}\theta_{\mathcal{V}})/(\lambda_{\mathcal{U}} + \lambda_{\mathcal{V}})$. Note that if $\mathcal{U}$ and $\mathcal{V}$ are complementary, their composite population $S_{\mathcal{U}+\mathcal{V}}$ is the full population F and then the subscript of the above notations are replaced with f. If $\mathcal{U}$ and $\mathcal{V}$ have an individual element for each, such as $\{1\}$ and $\{2\}$, we simplify the notation of $\mathcal{U}+\mathcal{V}$ as $1+2$. This notation simply denotes the union of $S_{\mathcal{U}}$ and $S_{\mathcal{V}}$, and it does not necessarily imply one is nested in the other.

2.1.2. Multi-stage design For multi-stage designs, the test statistic based on the accumulated data at the end of stage k ($k \le K$, the total stage number) for $S_{\mathcal{U}}$ is denoted by

$$Z^{1:k}_{\mathcal{U}} = \sum_{i=1}^{k}\sqrt{\frac{I^{(i)}_{\mathcal{U}}}{I^{1:k}_{\mathcal{U}}}}Z^{(i)}_{\mathcal{U}} = I^{1:k}_{\mathcal{U}}\big(\bar{Y}^{1:k}_{\mathcal{U},T} - \bar{Y}^{1:k}_{\mathcal{U},C}\big) \sim \mathcal{N}\big(I^{1:k}_{\mathcal{U}}\theta_{\mathcal{U}},1\big),$$

where the superscript $1{:}k$ refers to a quantity calculated based on the accumulated data at the end of stage k; therefore, $I^{1:k}_{\mathcal{U}}$ is the accumulated information level defined accordingly as $1/\big(\sigma\sqrt{1/n^{1:k}_{\mathcal{U},T} + 1/n^{1:k}_{\mathcal{U},C}}\big)$.

2.2. Designs considered

We consider designs that control the family-wise error rate (FWER) at level α in the strong sense [34] and the set of hypotheses to be tested

$$H_{0s}: \theta_s \le 0 \text{ versus } H_{as}: \theta_s > 0, \quad s \in \mathcal{S},$$

where $\mathcal{S}$ is the index set corresponding to the subpopulations considered and can index nested groups. For instance if we consider subgroup 1, subgroup 1 and 2 or the full population being of interest, $\mathcal{S} = \{1, 1+2, f\}$.

2.2.1. Single-Stage Designs To select, we use the maximum of the test statistics among $Z^{(1)}_s$, $s \in \mathcal{S}$ for population selection. Its implication and other selection rules will be discussed later. In the evaluation of the operating characteristics we consider the case where population selection is undertaken first and only subsequently the corresponding hypothesis being tested. The testing procedure is making a decision about rejecting H_{0w} if $Z^{(1)}_w \ge C_\alpha$, where w is a realized value of the random variable W and refers to the event that subpopulation S_w is chosen. $Z^{(1)}_w$ is the selected test statistic for S_w, and C_α is the corresponding critical value found to ensure the FWER in the strong sense.

The crucial element to finding the appropriate critical value and sample size is the density of the joint distribution of the selected test statistic $Z^{(1)}_W$ and the selected population index W. While the subsequent results are derived on the basis of selecting based on the maximum statistic, other selection rules can equally be implemented. Using a different rule results in a different density and for illustration purposes we also provide the resulting distribution for selecting any populations whose estimated effect exceeds a pre-specified value, δ, in the Supplementary Materials S.6. The joint

densities $p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta})$, $w \in \mathcal{S}$ govern the probability whether to select S_w and to reject the null hypothesis H_{0w} (where $\boldsymbol{\Theta}$ is a configuration of all mutually disjoint subgroup treatment effects $\theta_1, \theta_2, \ldots, \theta_J$). It can further be decomposed as $p_{Z_w^{(1)}}(z_w^{(1)}; \boldsymbol{\Theta}) \cdot Pr(W = w | Z_w^{(1)} = z_w^{(1)}; \boldsymbol{\Theta})$. Consequently, the joint densities of $Z_W^{(1)}$ and W can be represented as

$$p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}) = \phi(z_w^{(1)} - \theta_w I_w^{(1)}) \Psi_{\mathcal{S}\setminus w}\left(z_w^{(1)}, \ldots, z_w^{(1)}; \boldsymbol{\Theta}\right), \tag{1}$$

where ϕ denotes the standard normal density; $\Psi_{\mathcal{S}\setminus w}(\cdot, \ldots, \cdot; \boldsymbol{\Theta})$ is the cumulative distribution function (CDF) of the $|\mathcal{S}| - 1$-dimensional normal distribution conditional on $Z_w^{(1)}$ under a specified configuration of treatment effects $\boldsymbol{\Theta}$, where $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$. The covariance matrix depends on whether subgroups are nested or not (see examples in Supplementary Materials S.2 and S.3). The CDF specifies $Pr(W = w | Z_w^{(1)} = z_w; \boldsymbol{\Theta})$. It is noted that (1) is similar to the integrand of equation (4) in [9] where two co-primary analyses are performed on the full population and a subgroup, and the significance level for F is pre-specified.

Using an iterative search, C_α can then be found using the following inequality

$$\alpha \geq \sum_{w \in \mathcal{S}} \int_{C_\alpha}^{\infty} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_0) dz_w^{(1)}, \tag{2}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$ denotes the global null hypothesis H_0, $\theta_1 = \theta_2 = \ldots = \theta_J = 0$. Note that finding the critical value under this setting implies weak control of the FWER. Following [35] it can be shown, however, that weak control implies strong control since $\theta_1 = \theta_2 = \ldots = \theta_J = 0$ maximises the type I error when selection is based on the maximum. Similarly, assume an alternative hypothesis that exactly one subgroup (say S_w, w in $\mathcal{S}$) has nonzero positive effect size, δ, but others have none is true, the required total sample size for the full population $n_f^{(1)}$ can be found using the above critical values, a desired effect and a specified power level, $1 - \beta$. The related equation is

$$1 - \beta \leq \int_{C_\alpha}^{\infty} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_a) dz_w^{(1)}, \tag{3}$$

where $\boldsymbol{\Theta}_a$ denotes the alternative hypothesis, a vector of size J whose elements are all 0 except for the w^{th} element which is δ. The desired $n_f^{(1)}$ is obtained by iteratively increasing the sample size until equation (3) holds.

Note that only rejection of the hypothesis with the truly largest effect is considered in this power requirement. Similar considerations can be used to find the power to reject any false null hypothesis (see Figure 1 for an example).

We have derived the above formula here for consistency as for the multi-stage designs considered below only the selected subgroup continues to subsequent stages.

The derivations of (2) and (3) are provided in the Supplementary Materials S.1 and more specific example solutions for the single-stage design with two and three subgroups are given in Supplementary Materials S.2 and S.3 when the index set of selection population is $\mathcal{S} = \{1, f\}$ and $\mathcal{S} = \{1, 1+2, f\}$.

2.2.2. Multi-Stage Designs The multi-stage designs we consider follow similar procedures as the aforementioned single-stage designs. Population selection is performed at the first interim analysis, but any population in $\mathcal{S}$ can be selected. We consider the case where data after stage 1 are enriched so that the total sample size in the trial remains fixed but the sample size of subgroups that have not been selected is reallocated to the remaining populations. Suppose the selected population is S_w, the difference is that at stage k the testing procedure stops by rejecting H_{0w} if $Z_w^{1:k} \geq C_{u_k,\alpha}$, or stops with retaining H_{0w} if $Z_w^{1:k} \leq C_{l_k}$, or the procedure continues to stage $k+1$ if $C_{l_k} \leq Z_w^{1:k} \leq C_{u_k,\alpha}$, where $C_{u_k,\alpha}$ and C_{l_k} are the corresponding upper and lower stopping boundaries at stage k.

Two elements are required for appropriate stopping boundaries and stagewise sample sizes. The first is the joint density of $(Z_W^{(1)}, W)$, shown in (1). The second element is the density of the conditional distribution of the test statistics $Z_w^{1:k}$ (with accumulated data until stage k) given its precursor $Z_w^{1:(k-1)}$ at stage $k-1$. We denote this conditional density by $p_{w,k|k-1}(z_w^{1:k}|z_w^{1:(k-1)}; \boldsymbol{\Theta})$ and its general mathematical form is given in Supplementary materials S.4.

The stagewise density comprising of the two elements can then be used to determine the probability of stopping for efficacy or for futility at stage k. For example, the stagewise densities at stage 2 with different values of W are specified as

$$p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}) \cdot p_{w,2|1}(z_w^{1:2}|z_w^1; \boldsymbol{\Theta}), \quad w \in \mathcal{S}. \tag{4}$$

Then given $\boldsymbol{\Theta} = \Theta_0$ (i.e. under the global null hypothesis), the probability of early stopping at stage 2 (either for lack of effect or early rejection) for the subgroup S_w can be calculated as

$$\int_{C_{l_1}}^{C_{u_1,\alpha}} \int_{C_{u_2,\alpha}}^{\infty} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_0) \cdot p_{w,2|1}(z_w^{1:2}|z_w^{(1)}; \boldsymbol{\Theta}_0) dz_w^{1:2} dz_w^{(1)}, \quad w \in \mathcal{S},$$

where the integral bounds signify that the design continues after stage 1 but stops at stage 2 for efficacy. The conditional function $p_{w,2|1}(z_w^{1:2}|z_w^{(1)}; \boldsymbol{\Theta})$ is used to calculate stopping probability at stage 2 given that the design does not stop at the preceding stage. Similarly, the stagewise densities at stage k are the product of the expression in (1) multiplying the factor $\prod_{m=k}^{1} p_{w,m|m-1}(z_w^{1:m}|z_w^{1:(m-1)}; \boldsymbol{\Theta})$. The value of the k-fold multiple integral within the integrand region defined by stopping boundaries before stage $k+1$ is the early stopping probability at stage k. Each conditional density $p_{w,m|m-1}(z_w^{1:m}|z_w^{1:(m-1)}; \boldsymbol{\Theta})$ with its respective integral bound controls the probability of whether the design stops or continues, given that the design has proceeded at the previous stage.

To find boundaries that ensure FWER control an iterative search over the stopping boundaries is conducted based on the following inequality

$$\alpha \geq \sum_{w \in \mathcal{S}} \left\{ \sum_{k=1}^{K} \Big[\int \dots \int_{A_k} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_0) \cdot \Big(\prod_{m=k}^{1} p_{w,m|m-1}(z_w^{1:m}|z_w^{1:(m-1)}; \boldsymbol{\Theta}_0) \Big) dz_w^{1:k} \dots dz_w^{(1)} \Big] \right\}, \tag{5}$$

where the integration region A_k

$$A_k \quad = \quad [C_{l_1}, C_{u_1,\alpha}] \times [C_{l_2}, C_{u_2,\alpha}] \times \dots \times [C_{u_k,\alpha}, \infty) \text{ in } z_w^{(1)} \times z_w^{1:2} \dots \times z_w^{1:k},$$

where $\boldsymbol{\Theta}_0$ denotes the globe null hypothesis. We define $z_w^{1:0} = z_w^{1:1}$ and therefore $p_{w,1|1}(z_w^{1:1}|z_w^{1:1}; \boldsymbol{\Theta}) = 1$. Note that this yields only one inequality while $C_{l_1}, \dots, C_{l_k}$ and $C_{u_1,\alpha}, \dots, C_{u_K,\alpha}$ are all unknown. To overcome this, we set them to follow a specific functional form, where $C_{l_k} = C_{u_K,\alpha}$ for the K stage design. For example, when using the O'Brien Fleming (OBF) [13, 36] type stopping boundaries, $C_{u_k,\alpha} = C_{\text{OBF}}(K, \alpha)\sqrt{K/k}$ and C_{l_k} is a certain function of k. In addition, the calculations in (5) assumes that the futility bounds are binding. For non-binding bounds, one can simply set the lower bounds to $-\infty$.

As before, (5) implies weak control of the FWER but also guarantees strong control following the arguments in Magirr et al. [35].

Suppose an alternative hypothesis of the from $\theta_w = \delta > 0$ for exactly one element (say w) in $\mathcal{S}$ and $\theta_{w^*} = 0 \, \forall w^* \neq w \in \mathcal{S}$ is true. Then under this alternative hypothesis, the above critical values and specified power, the stagewise total sample

size for the full population $n_f^{(k)}$ can be found to satisfy the following inequality:

$$1-\beta \leq \sum_{k=1}^{K}\Big[\int\dots\int_{A_k} p_{Z_W^{(1)},W}(z_w^{(1)},w;\boldsymbol{\Theta}_a)\cdot\Big(\prod_{m=k}^{1} p_{w,m|m-1}(z_w^{1:m}|z_w^{1:(m-1)};\boldsymbol{\Theta}_a)\Big)dz_w^{1:k}\dots dz_w^{(1)}\Big], \tag{6}$$

where the configuration $\boldsymbol{\Theta}_a$ has an non-zero positive effect δ on the w^{th} element but the other $J-1$ elements are zero. Detailed derivations of (5) and (6) are provided in Supplementary Materials S.1 and the design details of two-stage designs with two subgroups (considering selection of S_1 or F) in Supplementary Materials S.5.

2.2.3. An illustrative example The Dose Ranging Efficacy And safety with Mepolizumab in severe asthma (DREAM) trial [37] investigates, amongst other endpoints, the effect of mepolizumab on exacerbations and forced expiratory volume in 1 second (FEV_1). Subsequent secondary analyses of the trial data [38, 39] finds that the treatment effect of mepolizumab depends on the baseline levels of eosinophil and suggests that only patients with blood eosinophil levels of more than 150 cells per μL receive benefit from the treatment.

Suppose that, on the basis of these exploratory findings, we wish to embark on a prospective evaluation of the claim that mepolizumab results in meaningful improvements only for patients with baseline levels of eosinophil of 150 or more cells per μL in the blood. We will use change in FEV_1 from baseline to 90 days modelled as normally distributed as the primary endpoints although the same arguments hold for other endpoints such as exacerbations. Additionally we suppose that the prevalence of each group (below and above 150 cells per μL blood) is 50%. Following [40] we assume that the standard deviation is 0.72L and consider a reduction of FEV_1 of 0.23L as the minimum clinically relevant treatment difference and consequently seek to power our evaluations for this effect.

Three different evaluation strategies are considered: 1) running two separate studies in each of the two subgroups, 2) a single stage study with one subgroup versus the full population (see section 2.2.1) and 3) a two-stage enrichment design where the best performing group is selected at the halfway point and early stopping using O'Brien and Flemming bounds [36] are used (see section 2.2.2). For each of the three designs we consider a type I error per study of 2.5%, require a power of 80% to reject any false null hypothesis. Further we assume that 25 patients are recruited per month and that it takes two month to conduct the interim analysis for strategy 3.

Strategy	max FWER	N	% superior	duration (months)
Separate studies	0.0494	616	50%	27.64
Single stage study	0.0250	684	50%	30.36
2-stage design	0.0250	552	75%	25.08 or 36.12

Table 1. Comparison of different evaluation strategies. max FWER is the maximum family-wise error of the strategy, N is the total sample size, % superior is the percentage of patient studied in the better performing subgroup and duration is the time from recruiting the first patient until the primary endpoint is available for all patients.

A summary of the characteristics of the different strategies is given in Table 1. The strategy using two separate studies requires just over 600 patients to be recruited while the single stage design with two groups does need almost 70 patients more. The reason for this is that no attempt has been made in the first approach to control the family-wise error rate. If we were to correct for multiplicity for the separate studies using a Bonferroni correction, the required sample size would increase to 748 patients. Using a 2-stage selection design allows us to reduce the required sample size even further to around 550 patients, a reduction of 10% and 30% compared to the uncorrected and multiplicity corrected separate study strategy, respectively. Additionally, the 2-stage design does investigate more patients in the group that is truly benefitting from treatment which is one of the reasons for the reduction in required sample size. Besides the reduction in sample size, running a single study rather than two separate ones does also yield organisational advantages. The main drawback of this

approach is that the duration of the study is increased by almost 9 month should the subgroup be selected (although a small reduction in the duration is expected if the full population is selected at interim.

Note that in addition to the advantages illustrated above the family-wise error rate in the 2-stage enrichment design is controlled for the worst case situation in terms of selection and hence other selection rules can be used without error rate inflation.

2.2.4. Alternative Designs We have illustrated how to obtain critical bounds and sample size for general enrichment designs above. Here, we discuss alternative designs considering different type-I-error and power configurations.

Significance levels and stopping boundaries:

An alternative to specifying the design and corresponding stagewise α levels via the boundaries is to specify marginal significance level α_k to each stage k (where $\sum_k \alpha_k = \alpha$) and use an error spending approach as used in classic group sequential designs [13]. Such considerations affect the way we find stopping boundaries where the same boundaries are shared by all the populations considered. More specifically, based on the following inequality (7) it is required to search the critical value used in A_{k-1} first under the upper limit of α_{k-1} (where the subscript of the upper bounds is changed accordingly). Then substitute those critical values for the associated bounds used in A_k under the upper limit of α_k for finding the remaining critical values and so on.

$$\sum_{i=1}^{k} \alpha_i \geq \sum_{w \in \mathcal{S}} \left\{ \Big[\int \dots \int_{A_k} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_0) \cdot \Big(\prod_{m=k}^{1} p_{w,m|m-1}(z_w^{1:m} | z_w^{1:(m-1)}; \boldsymbol{\Theta}_0) \Big) dz_w^{1:k} \dots dz_w^{(1)} \Big] \right\}, \tag{7}$$

Note that there are several ways to determine the lower stopping boundaries; for example, one could set symmetric values with respect to the upper critical values, or simply set 0.

One can further pre-specify the marginal significance levels for $|\mathcal{S}| - 1$ specific populations at each stage. One example of taking this consideration can be found in [9] although they only consider single-stage designs. Such design features may lead to different stopping boundaries for all the populations included in $\mathcal{S}$.

Incidentally, for two-stage designs if early stopping is not considered at stage 1 (that is, the stage-1 data is only used for population selection), then the first bound of integration in equation (5) and (6), A_k, is $(-\infty, \infty)$, where $k > 1$. Meanwhile, the upper bound C_{u_1,α_1} of A_1 is defined as ∞ and therefore the integral $\int_{A_1} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_0) dz_w^{(1)}$ is 0. Such designs are the same as the two-stage adaptive seamless designs used in [31].

Power:

The power of the designs in Section 2.2 is defined as the probability to detect the treatment effect of the population of interest under H_a. Alternatively we can define power to detect any treatment effects wherever they are from a set of specific subpopulations. Such change leads the total sample size for F to be different because of its influence on equation (6), which is the basis of searching $n_f^{(k)}$. Moreover, the equation becomes

$$1 - \beta \leq \sum_{w \in \mathcal{S}^*} \left\{ \sum_{k=1}^{K} \Big[\int \dots \int_{A_k} p_{Z_W^{(1)},W}(z_w^{(1)}, w; \boldsymbol{\Theta}_a) \cdot \Big(\prod_{m=k}^{1} p_{w,m|m-1}(z_w^{1:m} | z_w^{1:(m-1)}; \boldsymbol{\Theta}_a) \Big) dz_w^{1:k} \dots dz_w^{(1)} \Big] \right\}, \tag{8}$$

where $\mathcal{S}^*$ is the subset of $\mathcal{S}$ and contains the specified subpopulations of interest. Take an example that if $\mathcal{S} = \{1, f\}$ and $\mathcal{S}^* = \mathcal{S}$, Figure 1 shows the resulting total sample sizes $n_f^{(1)}$ in a single-stage design, corresponding to different

prevalence values of S_1, under different definitions of power. The left panel is computed to have power $1-\beta$ for selecting the subpopulation with the largest true effect and rejecting the corresponding null hypothesis, while the right panel considers any correct rejection. Under the left power definition the required sample size is large when the prevalence of the subgroup with a positive treatment effect is small as the number of patients having said effect is (relatively) small. As the prevalence λ_1 approaches 1, $n_f^{(1)}$ increases again as the effect of the subgroup dominates the effect in the full population and differentiating between the two populations becomes more difficult. In contrast $n_f^{(1)}$ always decreases under the definition of power to detect $\theta_1 > 0$ or $\theta_f > 0$. Since the effect sizes for S_1 and F are close, it is difficult to select the correct subgroup and thus large sample sizes are needed. The reason that the behaviour of $n_f^{(1)}$ is always decreasing for larger prevalences in the right panel is that there is no restriction on selecting a pre-specified population and reporting the efficacy. The decreasing pattern can be similar to that using the closed testing procedure [41] in a single-stage design, where the total sample is available for investigating any subpopulation without considering selection. Note that all the patterns observed in Figure 1 emerge in a case of multi-stage designs as well (not shown in this paper).

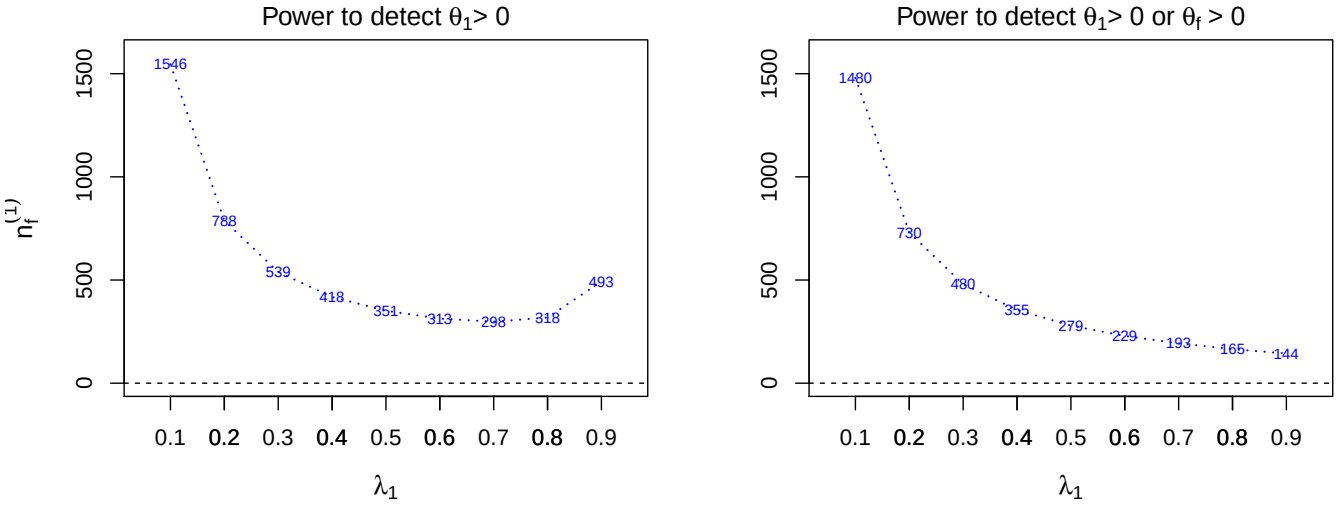

Figure 1. The total sample sizes of the full population F ($n_f^{(1)}$) across prevalence rates of S_1 (λ_1) for two different definitions of power. The design is a single-stage design with two subpopulations where the treatment effects θ_1 and θ_2 for S_1 and S_2 are 0.5 and 0, respectively. The type-I error and power are specified at 0.025 and 80%.

3. Estimation Assessment

In this section we report a simulation study assessing the properties of MLEs. Note that in the reported figures different scales for the y-axes are used in order to highlight patterns.

3.1. Simulation Set-up

In our evaluations, we specify the family-wise error rate, α, as 0.025 and set the sample size for each scenario so that the power of the design is $1-\beta = 80\%$. Our alternative hypothesis is that the treatment has an effect of 0.5 in S_1 while the effect of the treatment is zero for all other subgroups. Therefore the power aims to detect the non-zero effect in S_1 (that is to reject H_{01}) once the first subgroup is selected. The assumed common variance across subpopulations, σ^2, is set to 1 and we use 1,000,000 simulation runs.

The designs we consider are: a single-stage design with two subpopulations (**Design 1**), a single-stage design with three subpopulations (**Design 2**), a two-stage design with two subpopulations and three subpopulations (**Design 3** and

Design 4, respectively), with an O'Brien Fleming (OBF) upper stopping boundary and a fixed lower boundary of zero is used. We calculate the stopping boundaries and the total sample sizes for F based on (2) and (3) for single-stage designs (and (5) and (6) for multi-stage designs). The sample sizes and critical values for each of the designs are given in Appendix A. Based on these four designs, several scenarios are investigated altering the design features such as prevalence.

Denote $\hat{\theta}$ as the naive MLE (that is not accounting for selection) for the parameter θ, then $\hat{\theta}_f$ and $\hat{\theta}_s$ represent the MLEs for the treatment effect of F and S_s, respectively. The estimates can be calculated by $Z_s^{(k)}/I_s^{(k)} = \bar{Y}_{s,T}^{(k)} - \bar{Y}_{s,C}^{(k)}$, where $s \in \{1, f\}$ in scenarios for **Design 1** and $s \in \{1, 1+2, f\}$ in scenarios for **Design 2** . In multi-stage scenarios, the MLE estimates of $\hat{\theta}_f$ and $\hat{\theta}_1$ are calculated by $Z_s^{1:M}/I_s^{1:M} = \bar{Y}_{s,T}^{1:M} - \bar{Y}_{s,C}^{1:M}$, where $s \in \{1, f\}$ and M is the stage at which the study stops.

We define bias as $\text{bias}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta$ and the mean squared error (MSE), $\text{MSE}(\hat{\theta}) = \text{E}((\hat{\theta} - \theta)^2)$ as performance measures for estimation assessment. As the sample size for the full population satisfies the above power requirement and varies across different prevalence, a standardized scale is used in the assessments (readers are referred to Supplementary Materials S.7 for details on the standardization). In our subsequent evaluations we will consider three situations. Firstly, we consider the treatment effect estimator regardless of the population being selected or the hypothesis test being significant. Secondly we consider only the estimators of the selected populations which is expected to result in *selection bias*. The third situation considers *reporting bias* and for this we only consider only treatment effect estimates of the selected population if the corresponding hypothesis test is significant. Implicitly we are therefore considering that the outcome of a study is only reported (published) if it was significant. Note that in the evaluations to follow we refer to the selection bias as $Select\, S_w$ and the reporting bias as $Select\, S_w \,+\, Reject\, H_{0w}$, where w in $\mathcal{S}$ specifies the population chosen through a selection rule. In addition to the bias and MSE depending on which subgroup has been selected, we also report the family-wise (FW) bias and MSE, i.e. the bias and MSE averaged over all possible selections.

3.2. Scenarios for Design 1

Scenarios here cover different prevalence values of S_1, λ_1 varying from 0.05 to 0.95 in increments of 0.05. We illustrate the assessments for the scenarios under three configurations of different values of θ_1 and θ_2 in Figure 3-4. Their horizontal axes are for the prevalence of S_1, λ_1, and the vertical axes of the row-wise panels are for standardized bias, standardized $\sqrt{\text{MSE}}$ and simulation proportions (%).

Figure 2 presents the estimation assessment of $\hat{\theta}_f$ and $\hat{\theta}_1$ under the assumption of $\theta_1 = 0$ and $\theta_2 = 0$. As expected we do not see any bias when no selection is undertaken as well as constant standardized MSE - a pattern that is repeated throughout all other simulations. Additionally the selection probability is constant at 50% due to the equal effect in both subgroups. The selection bias is largest when the prevalence in the subgroup is smallest with a matching pattern for the standardized MSE. The reporting bias and MSE follow the same pattern although at a markedly increased level.

Figure 3 considers the case when $\theta_1 = 0.5$ and $\theta_2 = 0$. Considering the selection probabilities first, we find that, as per design, there is a 80% chance to select population 1 correctly and reject the corresponding hypothesis. The selection probability of the full population increases as the prevalence increases as the effect in the full population gets larger as the subpopulation contributes more towards it. At the same time the chance to also reject the hypothesis also increases. The selection and reporting bias in the full population estimate is largest when the prevalence in the subpopulation is smallest and then steadily decreases towards zero. The size of the bias is well over 0.5 standard errors for almost all prevalences and hence should be considered important although incorrect selection in itself is not very common in this case. For the full population the bias dominates the MSE and hence the MSE follows the same pattern.

Focusing attention on subpopulation 1, we find that bias is present, although it is of much smaller magnitude (selection bias at most 0.1 and reporting bias at most 0.35 standard errors) than for the full population (up to over 2 standard errors). The selection bias is maximised at a prevalence of around 0.75 while it is largest for a small prevalence for the reporting bias.

When both treatment groups have the same effect, $\theta_1 = \theta_2 = 0.5$ (Figure 4) we observe that almost always the full

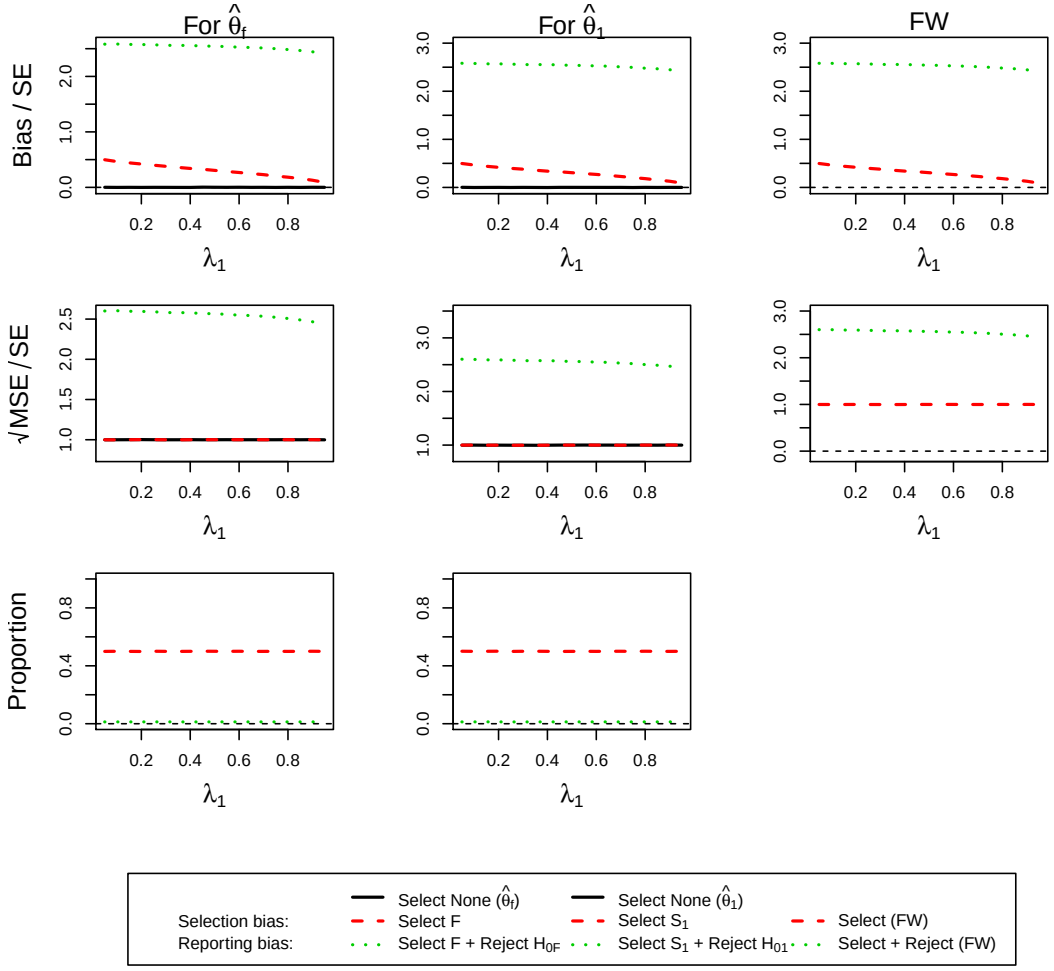

Figure 2. (For **Design 1**, $\theta_1 = 0$ and $\theta_2 = 0$) the standardized bias and standardized $\sqrt{\text{MSE}}$ of MLEs $\hat{\theta}_f$, $\hat{\theta}_1$ and the simulation proportions for different circumstances against the prevalence of subpopulation 1, λ_1.

population is selected and only for large prevalences of the subpopulation ($> 50\%$) we obtain notable selection probability for the subpopulation (up to 20%). As a consequence of this we obtain no estimate of the bias and MSE for the subpopulation for low prevalences. The bias in the estimate in this population is potentially very large (> 3 standard errors) but drops quickly towards zero as the prevalence increases. In this setting it is also notable, that the selection bias is virtually identical to the reporting bias as very large observed effects are necessary to select the subpopulation in the first place.

The patterns for the full population are somewhat more distinct as no bias is observed for small prevalences, since it is always the full population that is selected. The bias in this case is, however, very small even in the worst case situation (prevalence of around 0.75) where the reporting bias is less than 0.1 standard errors and the selection bias is even smaller.

3.3. Scenarios for Design 2

Scenarios for **Design 2** regard to select a population among S_1, S_{1+2} and F under different configurations of θ_1, θ_2 and θ_3. Our focus here is to assess the MLEs $\hat{\theta}_1$, $\hat{\theta}_1$ and $\hat{\theta}_f$ under $\theta_1 = 0.5, \theta_2 = 0, \theta_3 = 0$ under the population selection rule

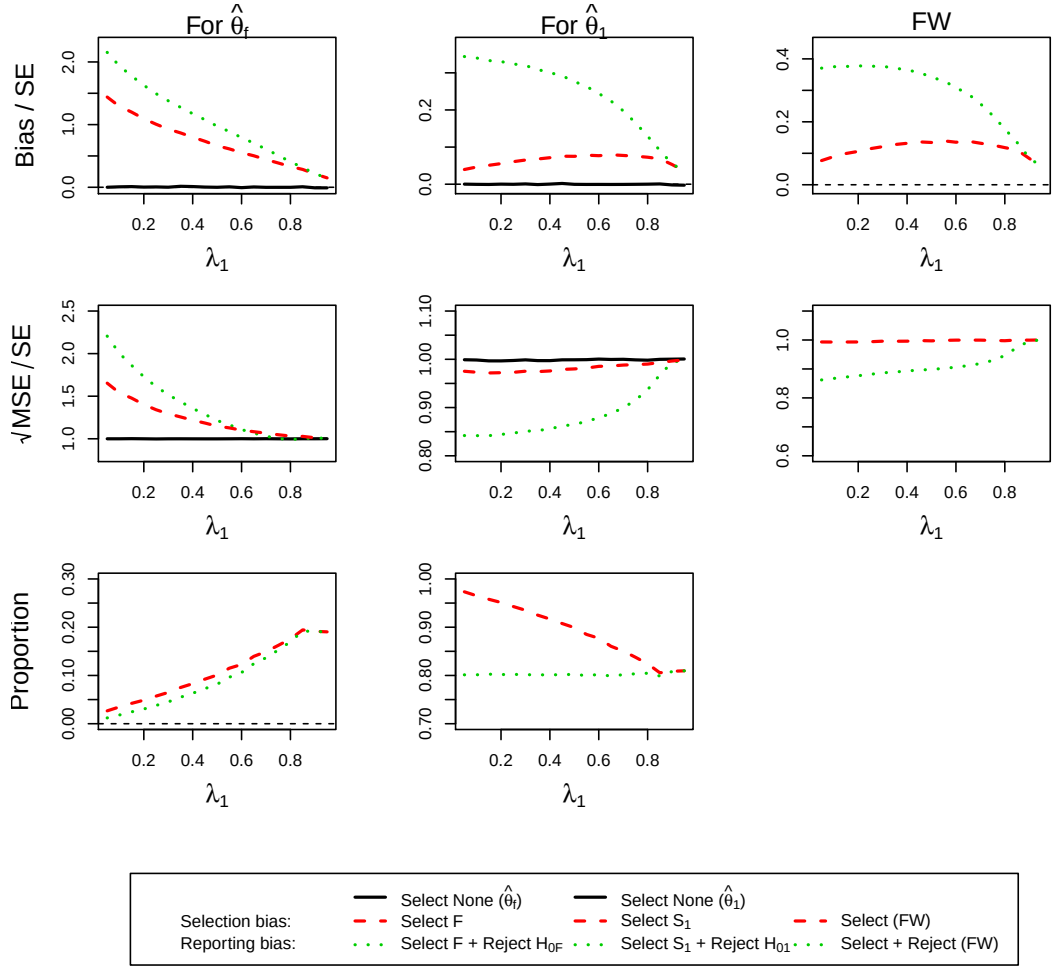

Figure 3. (For **Design 1**, $\theta_1 = 0.5$ and $\theta_2 = 0$) the standardized bias and standardized $\sqrt{\text{MSE}}$ of MLEs $\hat{\theta}_f, \hat{\theta}_1$ and the simulation proportions for different circumstances against the prevalence of subpopulation 1, λ_1.

given by

$$\begin{cases} \text{select } S_1 & \text{if } Z_1^{(k)} > \max(Z_f^{(k)}, Z_{1+2}^{(k)}) \\ \text{select } S_{1+2} & \text{if } Z_1^{(k)} \not> \max(Z_f^{(k)}, Z_{1+2}^{(k)}), \text{and } Z_{1+2}^{(k)} > Z_f^{(k)} \\ \text{select } F & \text{if } Z_1^{(k)} \not> \max(Z_f^{(k)}, Z_{1+2}^{(k)}), \text{and } Z_{1+2}^{(k)} < Z_f^{(k)}, \end{cases} \tag{9}$$

This rule is one variant of the maximum statistic rule and sequentially decides which population to be selected. The results for other configurations of θ_1, θ_2 and θ_3 are provided in Tables S.1-S.3 in Supplementary Materials S.8. Note that for all the scenarios simulations are run under the same stopping boundaries and sample sizes ($n_f^{(1)} = 576$) found based on **Design 2** with the maximum statistics selection rule, $\theta_1 = 0.5, \theta_2 = 0, \theta_3 = 0$ and equal subgroup prevalence.

The results in Table 2 shows that in this case the correct population is selected most of the time ($> 80\%$) due to the design constraint to obtain 80% power. The selection bias when selecting the correct population is small at < 0.1 standard errors and even the reporting bias is only modest at 0.27 standard errors. The selection and reporting bias when selecting the incorrect population are notably larger in this instance resulting in biases up to 1.3 standard errors. The bias is largest for the full population as the true underlying effect in this group is at 0.167 smallest amongst all populations and hence a rather unusual sample is required for its MLE to be largest.

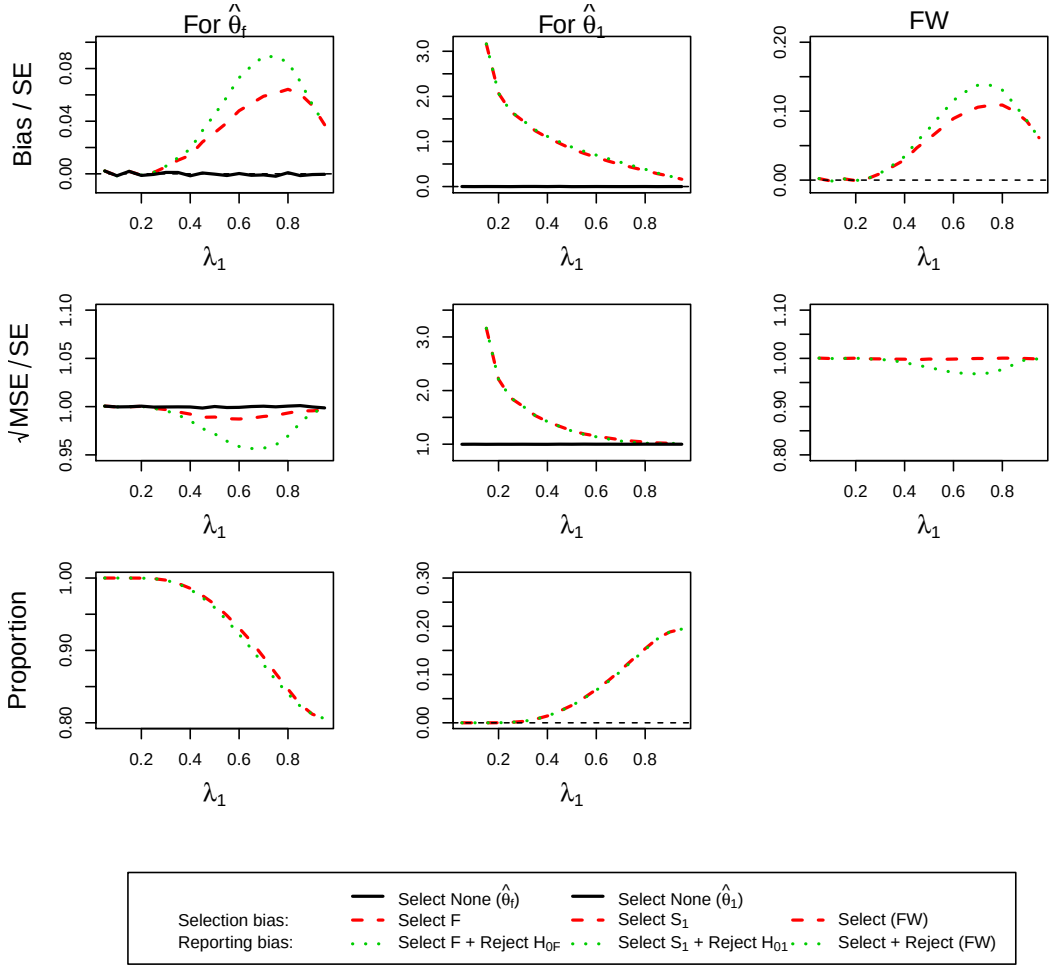

Figure 4. (For **Design 1**, $\theta_1 = 0.5$ and $\theta_2 = 0.5$) the standardized bias and standardized $\sqrt{\text{MSE}}$ of MLEs $\hat{\theta}_f$, $\hat{\theta}_1$ and the simulation proportions for different circumstances against the prevalence of subpopulation 1, λ_1.

	Bias/SE	$\sqrt{\text{MSE}}$/SE	Prop.(%)
$\hat{\theta}_f$ (Select None)	-0.00186	0.99849	
$\hat{\theta}_f$ (Select F)	0.96546	1.32104	3.74
$\hat{\theta}_f$ (Select F + Reject H_{0F})	1.31217	1.47472	2.91
$\hat{\theta}_1$ (Select None)	-0.00151	1.00004	
$\hat{\theta}_1$ (Select S_1)	0.09094	0.97526	88.58
$\hat{\theta}_1$ (Select S_1 + Reject H_{01})	0.27068	0.87036	80.20
$\hat{\theta}_{1+2}$ (Select None)	-0.00118	0.99884	
$\hat{\theta}_{1+2}$ (Select S_{1+2})	0.76128	1.19617	7.68
$\hat{\theta}_{1+2}$ (Select S_{1+2} + Reject $H_{0,1+2}$)	1.02579	1.26021	6.47
Family-wise Select	0.17516	1.00518	
Family-wise Select + Reject	0.35902	0.91814	

Table 2. (For **Design 2**, $\theta_1 = 0.5$, $\theta_2 = 0$ and $\theta_3 = 0$) Standardized bias and standardized $\sqrt{\text{MSE}}$ of the MLEs where the prevalence rates of three subgroups are 1/3. In addition, Proportion (Prop.) stands for how often the corresponding circumstance occurs.

3.4. *Scenarios for Design 3*

The investigation presented here concerns Design 3, a two-stage design and we focus on $\theta_1 = 0.5$ and $\theta_2 = 0$ here while the results for other configurations are given in Figures S.1-S.6 of Supplementary Materials S.8.

Figure 5 shows the results of the estimator for the full population. The top row corresponds to standardized bias, middle row to standardized $\sqrt{\text{MSE}}$ and the bottom row to the probability of selecting the full population. The first column is associated with the estimators that stop at Stage 1, the second considers only trials that reach Stage 2 while the final column corresponds to the estimator irrespective of when the trial was stopped. In addition to the selection bias and the reporting bias, we also consider the estimator irrespective of the reason for stopping (green triangle) in the figure.

The reporting bias is potentially very large (up to 3 standard errors for stage 1 only and up to 2 standard errors for stage 2) and is largest when the prevalence of the subgroup is small and subsequently decreases. When only considering studies that select the full population and stop at stage 1 it approaches zero while the bias does in fact become negative for trials that stop at the second stage. The overall estimator is, however, always positively biased showing a very similar pattern as the stage 1 cases only. The selection bias overall and for stage 2 only follows the same pattern as the reporting bias while it does show an inverted U-shape for stage 1 only which is maximised at a prevalence of around 0.5. The bias in the estimator that only considers stopping at stage 1 for any reason follows the same pattern as the selection bias although the bias is smaller. It is noteworthy that, although substantial bias is exhibited under some situation, the probability of reaching these (e.g. selecting the full population and stopping at stage 1) are very rare. The standardized $\sqrt{\text{MSE}}$ appears like that in standardized bias except for the second stage. In those exceptional cases, the MSE (for selection, reporting and regardless of selection) decreases at a different rate before inflating substantially at a prevalence of 0.8.

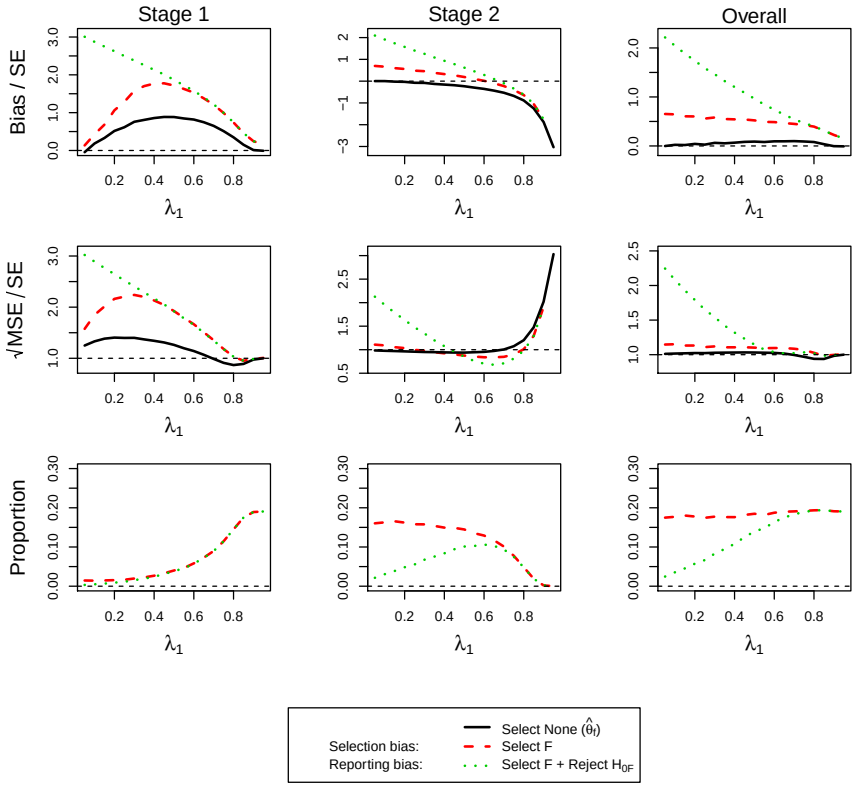

Figure 5. (For **Design 3**, $\theta_1 = 0.5$ and $\theta_2 = 0$) standardized bias and MSE of $\hat{\theta}_f$ and simulation proportions for different circumstances at stopping stage 1, 2 and overall, against the prevalence of subpopulation 1, λ_1.

Considering the findings for the estimator of the first subpopulation, $\hat{\theta}_1$ (Figure 6), the results exhibit similar patterns in

many circumstances in Figure 5. When stopping the trial at the first stage, the estimator is largely biased for prevalences up to 0.6. The reporting bias subsequently decreases from 2 standard errors while the selection bias is more moderate at around 1 SE. All the MSE (regardless of any circumstances) decreases to 0.9 SE from 2 and is close one for larger prevalances larger 0.7. As most of the time the subpopulation is selected correctly, the selection bias and the bias considering all studies that stopped at stage 1 are very similar and the MSE, meanwhile, is near 1 standard error. The estimators considering only trials that stop at stage 2 are almost unbiased for small and moderate prevalence but can exhibit a large negative bias when the prevalence is large. The MSE is close to 1 SE for most of prevalences but becomes very large beyond a prevalence of 0.7. The overall estimator is, however, positively biased (for both selection and reporting) for all prevalences and shows an inverted U-shape with a maximum bias of about 0.3 SEs for a prevalence of 0.6. Its MSE conditional on selection or no-selection appears different from that considering reporting before a prevalence of 0.7. The estimator thereafter performs similarly in MSE with a small U-shape under 1 SE.

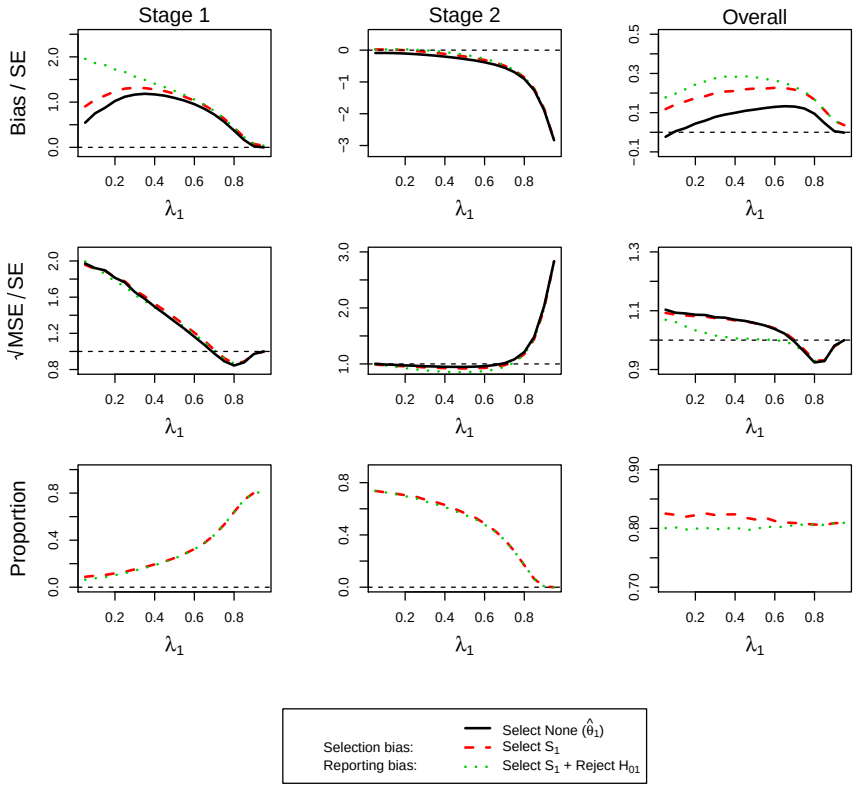

Figure 6. (For **Design 3**, $\theta_1 = 0.5$ and $\theta_2 = 0$) standardized bias and MSE of $\hat{\theta}_1$ and simulation proportions for different circumstances at stopping stage 1, 2 and overall, against the prevalence of subpopulation 1, λ_1.

The family-wise bias and MSE for this design with $\theta_1 = 0.5$ and $\theta_2 = 0$ is given in Figure 7.

3.5. Scenarios for Design 4

Scenarios for **Design 4** is the two-stage counterpart of **Design 2** for selecting a population among S_1, S_{1+2} and F under different configurations of θ_1, θ_2 and θ_3. The investigation here focus on assessing the MLEs $\hat{\theta}_1$, $\hat{\theta}_{1+2}$ and $\hat{\theta}_f$ under $\theta_1 = 0.5, \theta_2 = 0, \theta_3 = 0$ under the population selection rule given in the equation 9. The results for other configurations of θ_1, θ_2 and θ_3 are provided in Tables S.4-S.6 in Supplementary Materials S.8. All the simulations are run under the same stopping boundaries and sample sizes ($n_f^{(1)} = 335$) found based on **Design 4** with the maximum statistics selection rule, the configuration of treatment effects ($\theta_1 = 0.5, \theta_2 = 0, \theta_3 = 0$) and subgroup prevalences being 1/3.

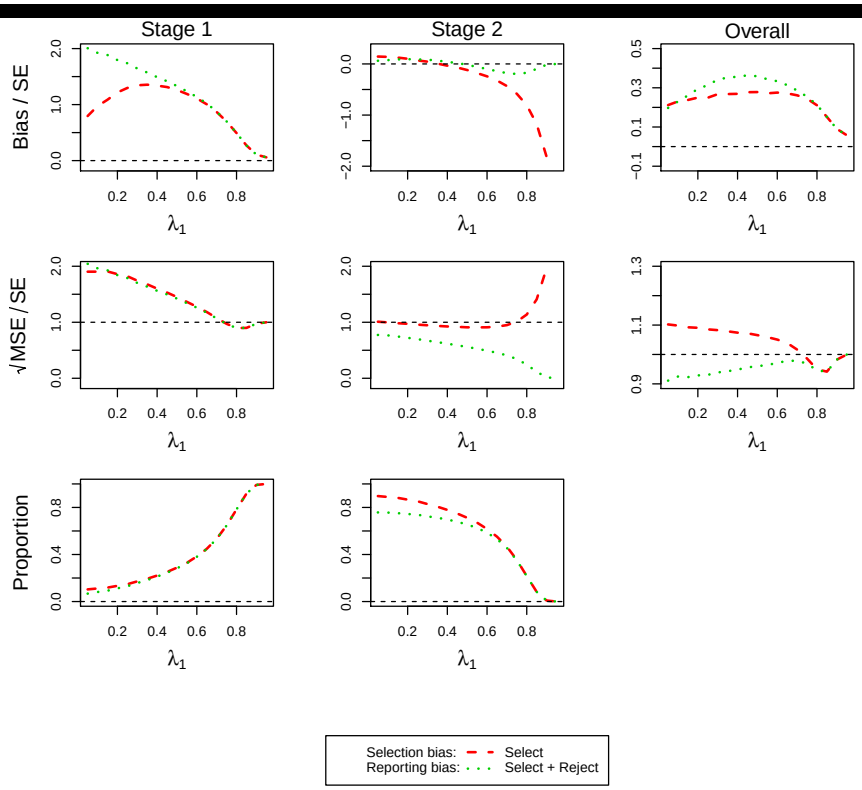

Figure 7. (For **Design 3**, $\theta_1 = 0.5$ and $\theta_2 = 0$) family-wise bias and MSE and simulation proportions for different circumstances at stopping stage 1, 2 and overall, against the prevalence of subpopulation 1, λ_1.

	Stop at Stage 1			Stop at Stage 2			Overall		
	Bias/SE	$\sqrt{\text{MSE}}$/SE	Prop.(%)	Bias/SE	$\sqrt{\text{MSE}}$/SE	Prop.(%)	Bias/SE	$\sqrt{\text{MSE}}$/SE	Prop.(%)
$\hat{\theta}_f$ (Select None)	0.67715	1.13140		-0.26340	0.96585		0.05614	1.02209	
$\hat{\theta}_f$ (Select F)	2.02110	2.14318	1.54	0.36337	0.92461	5.98	0.70283	1.17414	7.52
$\hat{\theta}_f$ (Select F + Reject H_{0F})	2.10568	2.14995	1.51	0.85268	1.01727	3.89	1.20283	1.33379	5.39
$\hat{\theta}_1$ (Select None)	1.03255	1.22555		-0.29968	0.97840		0.15293	1.06237	
$\hat{\theta}_1$ (Select S_1)	1.12932	1.27694	29.13	-0.20496	0.94416	51.17	0.27906	1.06487	80.29
$\hat{\theta}_1$ (Select S_1 + Reject H_{01})	1.14828	1.26420	28.99	-0.20175	0.93853	51.11	0.28684	1.05639	80.11
$\hat{\theta}_{1+2}$ (Select None)	0.81892	1.16958		-0.26809	0.96219		0.10120	1.03265	
$\hat{\theta}_{1+2}$ (Select S_{1+2})	1.78983	1.88884	3.31	0.17732	0.89687	8.88	0.61488	1.16604	12.18
$\hat{\theta}_{1+2}$ (Select S_{1+2} + Reject $H_{0,1+2}$)	1.82834	1.88619	3.27	0.41744	0.81323	7.57	0.84338	1.13715	10.85
Family-wise Select	1.23403	1.37576	33.97	-0.10207	0.93603	66.03	0.35185	1.08542	
Family-wise Select + Reject	1.25693	1.36402	33.77	-0.06135	0.92826	62.57	0.40076	1.08101	

Table 3. For **Design 4**, $\theta_1 = 0.5$, $\theta_2 = 0$ and $\theta_3 = 0$) Standardized bias and standardized $\sqrt{\text{MSE}}$ of the MLEs where the prevalence rates of three subgroups are 1/3. In addition, Proportion (Prop.) stands for how often the corresponding circumstance occurs.

Table 3 shows the results of the estimators for the first subgroup, the combined subgroup and the full population. The standardized bias, standardized $\sqrt{\text{MSE}}$ and simulation proportions are presented in the trials that stop at Stage 1, reach Stage 2 and are irrespective of which stopping stage.

Considering the trials irrespective of stopping, we observed the correct population is selected in the 80% of simulations due to the design requirement of 80% power. The bias is found positive for all the overall estimators and varies widely (smallest at 0.05 and maximum up to 1.2 standard errors). The selection and reporting bias when selecting the correct

population are the smallest (less than 0.3 standard errors), but larger when selecting the incorrect population (particularly for the full population). All the standardized MSE are larger than 1 standard errors but only up to a moderate size of around 1.3. While selecting the correct population or rejecting the null hypothesis the estimator for the first subgroup has a smaller standardized MSE (around 1.06 standard errors) than its counterparts.

The results at different stages show a contrary picture. More trials stop at stage 2 than at Stage 1 and each stage has a higher proportion of selecting the correct population (around 30% and 50% at Stage 1 and Stage 2, respectively). The bias is large at Stage 1. The selection and reporting bias are smaller when selecting S_1 (around 1.1 standard errors) than those when selecting S_{1+2} or F (around 1.8 and 2, respectively). A moderate bias is observed at Stage 2 (up to 0.85 standard errors). In particular, the selection and reporting bias are found negative in the estimator for the first subgroup. The standardized MSE of all the estimators at Stage 1 are much larger than 1 SE but those at Stage 2 show the opposite pattern being less than 1 (between 0.8 and 1).

4. Discussions and concluding Remarks

In this paper we have discussed general design considerations for clinical trials with subpopulation selection and illustrate how such studies can be designed. The design framework described can be viewed as an extension of group-sequential methods [42] and therefore requires the same types of assumptions and specifically we do assume an independent increment structure of the data. In our evaluations we have assumed that the primary endpoint is available immediately or at least before the next patient is recruited to the trial. While the general results in the paper will remain to hold if the endpoint is available only after some time, patients may still be recruited from a subpopulation that is subsequently not selected. Different approaches to deal with delayed responses have been proposed (e.g. [43]) in the context of group-sequential trials have been proposed. As a general rule, however, it is clear that the efficiency of selection is reduced if the time to observe the endpoint is long in comparison with the recruitment speed. Other assumptions made within this framework are common to most adaptive designs. Most notably we are assuming that there are differences in the population before and after interim analysis and in particular that no time trends are present.

In this work we only consider designs with normally distributed endpoints, although they can easily be extend to other types of endpoints via the efficient scores framework [42, 44]. Note, however, that particular care is required when using time to event endpoints - see [45] for a more detailed challenges of adaptive trials with time to event endpoints. Moreover we assume that the subgroup prevalence is known although clearly specifying this parameter correctly in the design will be crucial for the designs operating characteristics. A consequence of the assumed known prevalence is that we only present the estimation assessment of the MLE where subgroup sample sizes are fixed according to the respective prevalence in designs. Further simulations (not shown), however, suggest that random sample sizes of populations only alter the findings marginally.

Selection based on the maximum test statistics is the main focus throughout the paper and an R package implementing this design is currently under development. While this selection rule is simple and intuitive, it may not be optimal in certain circumstances. It makes sense to adopt the rule when some subgroup treatment effects have been identified as being positive and difference between test statistics across subgroups are reasonably large. However, when the test statistic for S_s and F are close but the former is larger, applying this rule leads to ethical issues that selecting only part of the population rather than the whole population although they could benefit from the treatment. Therefore, other options for selection rules should be considered for similar situations and investigation.

One alternative, which is also considered for designs with treatment selection (e.g. [46]), can be to introduce a threshold in the selection rule. This allows all the subgroups whose effect sizes are similar to the best one (their absolute difference is within a threshold) to be united so that the pooled population can continue to the next stage. Meanwhile, it also permits

to select a population whose effect size is above a threshold plus the effect size from the others.

Another option that has been used in the context of treatment selection (e.g. [35, 16]) is simply to select a population whose efficacy exceeds a certain value at stage 1. This selection rule was used in [16] and integrates population selection and hypothesis testing at the first stage. Their designs considering a prior ordering on underlying effect sizes of all individual subgroups somehow connect to ours where the target subpopulations for selection has a nested structure. It is noted that the mathematical expression of $p_{Z_W^{(1)},W}(\cdot,\cdot)$ in (1) will be different if the above selection rules are used. We provide the required modifications to the design framework in the Supplementary Materials for illustrative purposes.

In term of estimation we have assessed the bias of the MLE under various scenarios. We find that almost always bias is positive leading to an over-enthusiastic estimate of the true treatment effect. While for some settings the size of the bias can be viewed as negligible it can become large under other situations. The challenge clearly being that one will usually not know if one is in one of these extreme situations. Another observation we make is that although bias is introduced by selecting the population, the bias gets markedly increased (often more than doubled) when only significant results are reported highlighting the effect of reporting bias which may be even more problematic than the bias introduced by selection.

Our results suggest the MSE of the overall MLEs performs quite well (around 1 standard error) in many circumstances and scenarios. We find whether selecting the correct population or not impacts the size of MSE for the corresponding estimator. The extent can be more substantial when further reporting significant results. The same finding is even observed in the extreme scenario, where no correct population is defined because the underlying effect of each subgroup is assumed none.

Future work will consider estimators that are unbiased (or have smaller bias) while maintaining comparable MSE. Conditional bias-adjusted estimator following the ideas in [28] appear most promising. One extension to the case of multiple-stage designs given the process continues to the final stage can be naturally achieved. However, whether the derived estimators have less MSE should be verified in further investigations.

References

1. Kent D M, Rothwell P M, Ioannidis J P, Altman D G, Hayward R A. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010; **11**(1):85.
2. Varadhan R, Segal J B, Boyd C M, Wu A W, Weiss C O. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of clinical epidemiology* 2013; **66**(8):818-825.
3. Basu S, Sussman J B, Hayward R A. Detecting heterogeneous treatment effects to guide personalized blood pressure treatment: a modeling study of randomized clinical trials. *Annals of internal Medicine* 2017; **166**(5):354-360.
4. Cuzick J. Forest plots and the interpretation of subgroups. *The Lancet* 2005; **365**(9467):1308.
5. Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, Stallard N, Posch M. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review.*Journal of Biopharmaceutical Statistics* 2016; **26**(1): 99-119.
6. Alosh M, Huque M, Bretz F, D'Agostino, R.B. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Statistics in Medicine* 2017; **36**(8):1334-1360.
7. Lipkovich I, Dmitrienko A, D'Agostino R.B. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine* 2017; **36**(1): 136-196.
8. Placzek M, Friede T. Clinical trials with nested subgroups: Analysis, sample size determinatino and internal pilot studies. *Statistical Methods in Medical Research* 2017; first published date: March-14-2017, 10.1177/0962280217696116.
9. Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials.*Contemporary Clinical Trials* 2010; **20**: 331-335.
10. Song Y, Chi G. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 2007; **26**(1): 3535-3549.
11. Alosh M, Huque M. A flexible strtegy for testing subgroups and overall population. *Statistics in Medicine* 2009; **28**(1): 3-23.
12. Graf A C., Posch Martin, Koening F. Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal* 2015; **57**(1): 76-89.
13. Jennison C, Turnbull B W. *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC Boca Raton, FL, USA, 2000.

14. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**: 1833-1848.
15. Ghosh P, Liu L.Y., Senchaudhuri P., Gao P., Mehta C. Design and Monitoring of Multi-Arm Multi-Stage Clinical Trials. *Biometrics* 2017; first published: 27 March 2017. DOI: 10.1111/biom.12687.
16. Magnusson B, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine* 2013; **32**(16): 2695-2754.
17. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 2011; **10**: 347-356.
18. Stallard N, Hamborg T, Parsons N, Frided T. Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics* 2014; **24**(1): 168-187.
19. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias–Two hostile brothers. *Statistics in Medicine* 2010; **29**(1): 1-13.
20. Shen L. An improved method of evaluating drug effect in a multiple dose clinical trial. *Statistics in Medicine* 1999; **20**: 1913-1929.
21. Stallard N, Todd S, Whitehead J. Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference* 2008; **138**: 1629-1638.
22. Cohen A, Sackrowitz H. Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* 1989; **8**: 273-278.
23. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **50**(4): 515-527.
24. Sill M W, Sampson A R. Extension of a two-stage conditionally unbiased estimator of the selected population to the bivariate normal case. *Communications in Statistics - Theory and Methods* 2007; **36**(**4**): 801-813.
25. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine* 2013; **32**(10): 1677-1690.
26. Kimani P, Todd S, Stallard N. A comparison of methods for constructing confidence intervals after phase II/III clinical trials. *Biometrical Journal* 2014; **56**(1): 107-128.
27. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika* 2013; **100**(4): 985–996.
28. Stallard N, Todd S. Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference* 2005; **135**: 402-419.
29. Brckner M, Titman A, and Jaki T. Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint. *Statistics in Medicine* 2017; published online ahead of print.
30. Rosenkranz G K. Bootstrap corrections of treatment effect estimates following selection. *Statistics in Medicine* 2014; **69**: 220-227.
31. Kimani P, Todd S, Stallard N. Estimation after subpopulation selection in adaptive seamless trials. *Statistics in Medicine* 2015; **34**(18): 2581-2601.
32. Wang Y, Leung D. Bias reduction via resampling for estimation following sequential tests. *Sequential Analysis* 1997; **16**(3): 249-267.
33. Davison A, Hinkley D. *Bootstrap Methods and Their Application.* Cambridge University Press: Cambridge, UK, 1999.
34. Dmitrienko A, Tamhane A.C., Bretz F., editors. *Multiple Testing Problems in Pharmaceutical Statistics.* Chapamn & Hall/CRC biostatistics series, Boca Raton, 2010.
35. Magirr D, Jaki T., Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**: 494-501.
36. O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. *Biometrics* 1979; **35**(3): 549-556.
37. Pavord I D, Korn S, Howarth P, Bleecker E R, Buhl R, Keene O N, Ortega H, Chanez P. Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial. *The Lancet* 2012; **380**(9842): 651-659.
38. Ortega H G, Yancey S W, Mayer B, Gunsoy N B, Keene O N, Bleecker E R, Brightling C E, Pavord I D. Severe eosinophilic asthma treated with mepolizumab stratified by baseline eosinophil thresholds: a secondary analysis of the DREAM and MENSA studies. *The Lancet Respiratory Medicine* 2016; **4**(7): 549-556.
39. Yancey S W, Keene O N, Albers F C, Ortega H, Bates S, Bleecker E R, Pavord I. Biomarkers for severe eosinophilic asthma. *Journal of Allergy and Clinical Immunology* 2017; **140**(6): 1509-1518.
40. Santanello N C, Zhang J, Seidenberg B, Reiss T F, Barber B L. What are minimal important changes for asthma measures in a clinical trial?. *European Respiratory Journal* 1999; **14**(1): 23-27.
41. Marcus R., Peritz E., Gabriel K.R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**(3): 655-660.
42. Whitehead J. *The Design and Analysis of Sequential Clinical Trials.*Revised 2nd edn. Wiley: Chichester, 1997.
43. Hampson L V, Jennison C. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2013;**75**(1):3-54.
44. Jaki T, Magirr D. Considerations on covariates and endpoints in multiarm multistage clinical trials selecting all promising treatments. *Statistics in Medicine* 2013; **32**(7): 1150-1163.
45. Magirr D, Jaki T, Koenig F, Posch M. Sample size reassessment and hypothesis testing in adaptive survival trials. *PLOS one*. 2016; **11**(2):e0146465.
46. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Tutorial in Biostatistics - Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2009; **28**: 1181-1217.

A. Design specifications

A.1. Design 1

λ	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
c	2.232	2.228	2.223	2.217	2.212	2.206	2.200	2.193	2.186	2.178
N	3070	1546	1040	788	638	539	469	418	380	351

λ	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
c	2.170	2.160	2.150	2.139	2.126	2.111	2.094	2.072	2.042
N	329	313	303	298	302	318	363	493	943

Table 4. Design specifications for different values of λ for Design 1. c is the critical value and N is the total sample size.

A.2. Design 2

λ	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
c_1	3.018	3.031	3.037	3.039	3.039	3.037	3.034	3.029	3.023	3.016
c_2	2.134	2.143	2.147	2.149	2.149	2.148	2.145	2.142	2.138	2.133
N	719	401	298	251	224	207	196	188	183	181

λ	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
c_1	3.008	2.999	2.989	2.977	2.964	2.948	2.930	2.907	2.875
c_2	2.127	2.121	2.114	2.105	2.096	2.085	2.072	2.055	2.033
N	181	184	192	205	229	269	342	491	943

Table 5. Design specifications for different values of λ for Design 2 and an interim analysis after half the patients have been observed. c_1 is the upper stopping boundary at stage 1, c_2 the final stage critial value and N the total sample size. The lower bound at stage 1 is fixed at zero for all λ.

A.3. Design 3

For this single stage design with three subgroups, the prevalance of each subgroup is equal to one third resulting in a critical value of $c = 2.289$ and a total sample size of $N = 575$.

A.4. Design 4

The two stage design with three subgroups uses equal prevalance of each subgroup and an interim analysis after half the patients have been observed. The critical value at the first stage is $c_1 = 3.119$ while the final critical value is $c_2 = 2.205$. A fixed futility bound of zero is used and the total sample size is $N = 335$.