A Seamless hybrid Phase II/III design with Bayesian interim subgroup selection

Benjamin Duputel, Nigel Stallard, François Montestruc, Sarah Zohar, Moreno Ursino

A Seamless hybrid Phase II/III design with Bayesian interim subgroup selection

Benjamin Duputel[1,2,3], Nigel Stallard[4], François Montestruc[3], Sarah Zohar[1,2,†] and Moreno Ursino[1,2,5,†]

[1]Université Paris Cité, Sorbonne Université, Inserm, Centre de Recherche des Cordeliers, F-75006 Paris, France

[2]Inria, HeKA, F-75015 Paris, France

[3]eXYSTAT, 92240 Malakoff, France

[4]Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK

[5]Inserm CIC-EC 1426, F-75019 Paris, France

[†]Authors made equal contribution

September 13, 2024

Abstract

Population selection is a crucial subject in clinical development nowadays as personalized medicine is growing interest. Evolution on biomarker scanning techniques allow for composition and detection of subpopulation of interest when analysing new drug responses in a disease. Seamless adaptive trials could allow for subgroup analysis with selection of the most promising population at interim analysis. We propose a hybrid Bayesian design in two stage for seamless phase II/III trials with binary and time-to-event outcomes for the first and second phase, respectively. In this work, at interim analysis several prior distributions including shrinkage prior are compared to possibly select/discard a population, and

a final test using conditional error function as a combination method testing procedure, to control the frequentist type I error, is used. Simulation studies showed that the logistic regression model performs better than frequentist testing for population selection problem when the subgroup should be selected. Shrinkage prior distributions tends to be more conservative than simpler normal distributions as studies that would have ended positive are stopped at interim analysis.

1 Introduction

For the last decades, interest on flexible designs for clinical trials has grown significantly. Bothwell et al. [4] reported in their review paper that almost no adaptive clinical trials were found in the 1980-1990's whereas since 2010 more than twenty a year are recorded. Adaptive designs (AD) are appealing because they respond to ethics requirements as they propose faster ways to propose an efficient treatment to the patients and stop the inefficient ones. They can be more powerful than classical studies (by sample size re-estimation by example) or use a smaller sample size (seamless trials and combination testing) while controlling error rates. They also offer logistical and economical benefits. The counterpart is that those methods need more upfront planing and modeling.

Usually AD are constructed around interim analyses (IA) that help to decide applying or not a predetermined adaptation. The most common and simple adaptation is the group sequential design (GSD) as proposed, for example, by Pocock [16], in which repeated analysis of an outcome allow for early stopping once a significance threshold is reached. The idea has been extended in Jennison and Turnbull [11] or Whitehead [21] books allowing for early futility stopping of the trial if efficacy analyses reach a futility threshold. Other adaptations consider sample size reassessment method [7], dropping ineffective treatment or dose arms [18, 17], or selection of previously identified population [13].

For population selection problems, trials can be constructed as seamless studies (two phases combined in a unique protocol), where the Phase II, that aims at selecting the population that could benefit from the therapy, is combined with the confirmatory Phase III. Pooling both phases together has the benefit of reducing the global sample size while controlling type one error rate. Seamless PhaseII/III trials with population selection has been studied in a few

papers. Jenkins et al.[10] or Spiessens and Debois [19] for example proposed combination test methods to deal with the problem. Those methods select the population on the base of interim responses, and use combined values of statistical tests from both stages at final analysis. Friede et al.[8] proposed a conditional error procedure, which similarly to the previous procedures make interim decision based on statistical tests and conclude based on all studies values while taking into account the correlation between normally distributed statistical tests. This correlation structure has been studied and extended to multi populations, multi arms trials by Chen et al. [6]. Miao et al. [14] recently proposed a gated subgroup analysis using progression free survival and overall survival in a combination test for their seamless phase II/III trial. Bayesian theoretic designs have also been proposed as in Brannath et al. [5], Kimani et al. [12], or more recently, Ballarini et al. [3]. Those methods use maximization of utility functions (responding to a specific need as safety and efficacy, or logistical costs, etc.) to decide if the trial should go to the second stage with the subgroup or the full population.

In this paper we propose to use a Bayesian selection step based on posterior distribution from parameters of a logistic regression. A binary survival rate endpoint is used for the first stage of the study and overall survival for the final analysis. To account for the selection step and try to avoid multiplicity issues, we use the same conditional error function of Friede et al. [9] with the statistical test values from the first and the second steps in the final analysis. Comparison between multiple prior distributions is made and a frequentist two stage design based on statistical test for selection is also used for comparison. The proposed method uses a logistic regression and the selection step is based on posterior distribution rather than an utility function.

In the next section, we briefly present the real clinical trial Atalante-1 that serves as a motivational case study for this research. In Section 3, the model and design are described and illustrated. In Section 4 the parameters and the scenarios of simulations studies are presented along with the results for two distinct prevalence of the subgroup as well as for another selection threshold for the Bayesian designs. Finally the benefits and limitations of the proposed method are discussed.

2 Motivating study

The Atalante-1 clinical trial (NCT02654587) used a seamless Phase II/III design with a Fleming single arm method [8] for first stage and a stratified log-rank test for the second one. The study compared the experimental treatment (Tedopi) to the best standard of care (Docetaxel or Pemetrexed) in terms of survival rate at one year (stage 1) and overall survival (stage 2) on non-small-cell lung cancer. The operating characteristics of this trial included a type I error rate of 2.5% and a power of 80%. The sample size required was 84 evaluable patients for the first stage of the trial, under the null hypothesis of a 12-month Overall Survival rate of 25% in the treatment group, and an alternative hypothesis of 40%. At the interim analysis, the study could be either stopped for futility or proceeded to Phase III to compare overall survival with the control using a frequentist approach. For the second phase, aiming for a power of 80% to detect a significant difference with a bilateral significance level of 5%, 363 new patients were planned to be evaluated. This was under a 2:1 randomization scheme to detect a Hazard Ratio (HR) of 0.7, assuming a median survival of seven months in the control arm, which would imply a ten-month median survival in the treatment arm. At the time of the planned interim analysis when the first 103 patients reached 12 months of follow-up, decision was taken by the sponsor to prematurely stop the accrual due to the coronavirus disease 2019 (COVID-19) pandemic which was rapidly expanding with a strong concern about its impact on patient safety and data integrity. Thereafter, treatment and follow-up continued for the ongoing 219 patients already randomized. Due to this early accrual discontinuation, the data were unblinded and analyzed in the first 103 patients. A subgroup of interest from a stratification factor was identified based on a clinical and biological rationale [2]. At the time of final analysis, this subgroup was analyzed in the overall population of 219 patients as post-hoc analysis.

A confirmatory study (Artemia) is ongoing to confirm the treatment effect in this population of interest. Motivating by this experience, our research focused on exploring new Bayesian method for seamless design where a subgroup is pre-identified, at the design stage, and prospectively selected or discarded during the trial.

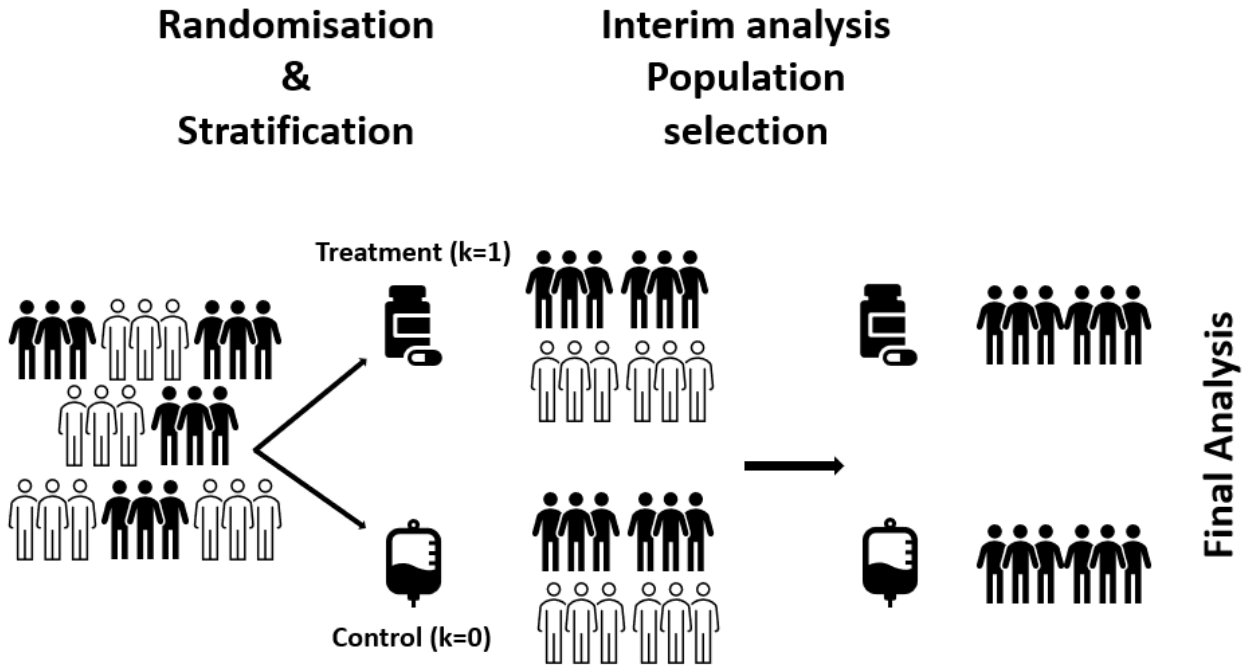

Figure 1: Trial scheme with selection of the relevant population at interim analysis based on binary survival rate endpoint. Then, the final analysis is based on a time-to-event endpoint. In this example, subpopulation is selected at interim and the treatment is found superior to the control at final analysis.

3 Model and Design

Our method shares the same endpoints with the case-study trial used. The mortality rate at one year is used for decision in the first step of the seamless trial, and overall survival is considered for final analysis. At interim analysis, a Bayesian approach, as described in the following subsections, is used as selection tool for the determination of the most promising population (possibly both) for the second step of the study. In the phase III of the trial, we use frequentist testing that accounts for the first stage data via conditional error function for final efficacy analysis. In addition, a closed testing procedure is considered when F and S populations are simultaneously analyzed at the end of the trial. The overall trial design is illustrated in Figure 1.

3.1 Notations

Let $k = 0$ and $k = 1$ be the indicator of control and treatment group respectively. In this work, only two arms are considered. Let N_{II} and N_{III} be the fixed sample size of the Phase II and

Phase III in each group. Let the full population be denoted by F, and the subgroup of interest be denoted by S (and its complementary S^C, $F = S \cup S^C$). Let $t_{i,k}$ and $c_{i,k}$ be the time to event (death in our motivating example) and time of censoring for patient i in treatment group k. Let $y_{i,k}$ be the first occurrence between event and censor (lost to follow up or alive at the time of analysis) $y_{i,k} = \min(t_{i,k}, c_{i,k})$ with $t_{i,k}, c_{i,k} > 0$, and let $\nu_{i,k}$ be the event indicator meaning that $\nu_{i,k} = 1$ if $y_{i,k} = t_{i,k}$ or $\nu_{i,k} = 0$ if $y_{i,k} = c_{i,k}$. Let $\mathcal{D}_k^p$ denote the data from group k at phase $p = II$ or III, $\mathcal{D}_k^p = \{N_k, \mathbf{y}_k, \boldsymbol{\nu}_k\}$, where N_k is the sample size of the group k at time of data collection (i.e. $N_k = N_{II}$ or $N_k = N_{III}$), and, $\boldsymbol{\nu}_k$ and $\mathbf{y}_k$ are vectors of length N_k containing all values of $y_{i,k}$ and $\nu_{i,k}$.

3.2 Stage 1 - Phase II

A dichotomous endpoint, the survival rate at t^* (1 year) is used for interim analysis decision. A patient can either be dead or alive at t^*, patients censored before observation of their status at t^* are excluded from the survival rate comparison. Interim analysis takes place when N_{II} are recruited in both group and have finished the follow-up period. We denote by $y_{i,k}^*$ the observation at t^*

$$y_{i,k}^* = \begin{cases} 1 \text{ if } y_{i,k} \geq t^* \\ 0 \text{ if } y_{i,k} < t^*, \end{cases} \tag{1}$$

Let $p_{i,k} \in [0, 1]$ represent the probability of being alive at time t^*, then for patient i in the k group, we have $y_{i,k}^* \sim \text{Bernoulli}(p_{i,k})$. We assume that $p_{i,k}$ depends on the sub-population for patient i and for k and we use the logit link function, as recommended by Albert and Hu [1], to link $p_{i,k}$ to patient's covariates, that is:

$$\text{logit}(p_{i,k}) = \theta^T X_{i,k}, \tag{2}$$

with $\theta = (\theta_0, \theta_S, \theta_T, \theta_{TS})$ and $X_{i,k}$ the covariate indicator matrix containing information on patient's group (F or S and control or treatment group). θ_0 is the intercept and represents the control effect on the Full population, θ_S is the shift from θ_0 of the global (control and treatment) sub-population group, θ_T the shift of the treatment group in the Full population, and θ_{TS} is the interaction term between the treatment arm and the sub-population. Prior distributions on these parameters are introduced in Section 3.5.

Table 1: Interim analysis decision relative to observed outcomes at the end of the Phase II.

$\mathbb{P}(\theta_{TS} > \zeta_e \mid \mathcal{D}_k) > \tau_S$	$\mathbb{P}(\theta_T > \zeta_e \mid \mathcal{D}_k) > \tau_F$	Population studied in Phase III
Yes	Yes	$F\&S$
	No	S if $\mathbb{P}(\theta_T + \theta_{TS} > \zeta_e \mid \mathcal{D}_k) > \tau_1$
		Futility if $\mathbb{P}(\theta_T + \theta_{TS} > \zeta_e \mid \mathcal{D}_k) \leq \tau_1$
No	Yes	F
	No	Futility

3.3 Interim selection rules

At the end of the Phase II, the promising population (possibly both) is selected for the second stage of the study. Decision rules for each situation are reported in Table 1. After the computation of posterior distributions of all parameters, posterior samples are used to decide if the trial will continue with F, S, or declared futile. In the case of the treatment showing efficacy in both S and F population, the trial would continue in F, but the final analysis would consider both populations. In this case, treatment benefits the full population, but the subgroup treatment effect is then stronger. The interim analysis uses θ_T and θ_{TS} as parameters of interest, meaning that the decision to continue or stop the study will only consider those two parameters as we expect the control to have already proven efficacy in the global population. Therefore as summarized in Table 1, the first step consists on looking at θ_{TS} posterior distribution and its probability of being higher than a fixed threshold ζ_e (for simplicity, ζ_e is set to zero) of being higher than an interim limit threshold $\mathbb{P}(\theta_{TS} > 0|\mathcal{D}_k) > \tau_S$. Given the observed outcomes, analysis of the θ_T parameter follows the same rule. If $\mathbb{P}(\theta_{TS} > 0|\mathcal{D}_k) > \tau_S$ and $\mathbb{P}(\theta_T > 0|\mathcal{D}_k) > \tau_F$, then the trial continues with an analysis in $S\&F$. If $\mathbb{P}(\theta_T > 0|\mathcal{D}_k) < \tau_F$, but $\mathbb{P}(\theta_T + \theta_{TS} > 0|\mathcal{D}_k) > \tau_1$ then the trial continue in the subgroup as a treatment effect is found in the subgroup, otherwise the trial stops for futility. If $\mathbb{P}(\theta_{TS} > 0|\mathcal{D}_k) < \tau_S$, if $\mathbb{P}(\theta_T > 0|\mathcal{D}_k) > \tau_F$ then the trial continues with an interest on F only (see below), if $\mathbb{P}(\theta_T > 0|\mathcal{D}_k) < \tau_F$, the trials stops for futility.

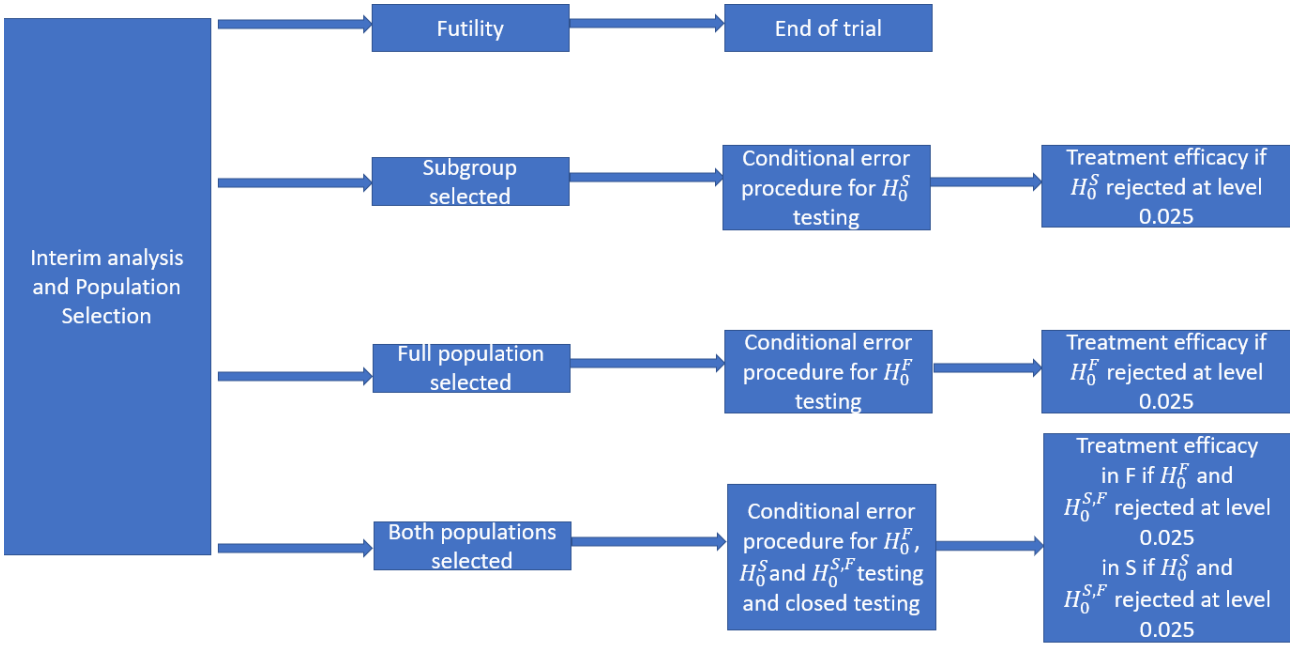

Figure 2: Representation of the trial tests at the end of the trial given the outcome of interim analysis.

3.4 Stage 2 - Phase III

If the trial continues to the second stage, there are three possible situations, either the trial continues with F, either it continues with S, or with both $F\&S$ for analysis as shown in Figure 2. An one-sided Log Rank Test (LRT) is considered for the final analysis: $H_0 : S_0(t) \geq S_1(t)$ vs $H_1 : S_0(t) < S_1(t)$, with S_0 and S_1 representing the survival function of the control and treatment group respectively, in the selected population. To account for the selection step and control the type-one error rate, we use the conditional error function approach along with Spiessens and Debois [19] testing method to account for correlation of F and S tests. If both population are selected to be studied in the final analysis, we also use a closed test procedure to ensure that the type-I error rate is controlled at the desired level (0.025 unilateral in our application).

3.4.1 The Conditional Error Function (CEF) accounting for correlation

The CEF approach used for the final analysis is proposed by Friede et al. [8] and used in [9] or Stallard et al. [20]. It is known to have good properties (better power and comparable type-I error rate control) compared to other combination approaches. The use of the Spiessens and

Debois [19] testing procedure allow for correlation between F and S tests to be accounted for. This method also considers the existing correlation between different tests statistics as well as correlation between S and F denoted Z^S and Z^F. The use of a combination approach supposes that the Phase II data is stored using the late time-to-event endpoint, therefore first stage patients of the selected group continue follow-up until the final analysis.

Let H_0^F be the null hypothesis for F, H_0^S the one of S, and $H_0^{\{S,F\}} = H_0^F \cap H_0^S$ the intersection null hypothesis. Let Z_1^F and Z_1^S be the observed normalized tests statistics of Phase II time-to-event data in F and S, and similarly $Z_{1,b}^F$ and $Z_{1,b}^S$ be the tests statistics of Phase II binary data in F and S. We also assume that $corr(Z_1^F, Z_{1,b}^F) = \rho$ (similarly for S). Let $S_1^F = w_1 Z_{1,b}^F$ (resp. $S_{1,b}^S$) be the weighted tests where $w_1 = \sqrt{n_{phII}/n_{total}}$ represents the ratio of patients of phase III over the total sample size. For the Phase III data, let Z_2^F (resp. Z_2^S) be the statistic tests and $S_2^F = w_1 Z_1^F + w_2 Z_2^F$ (resp. S_2^S), $w_2 = \sqrt{1 - w_1}$ represents the ratio of patients of phase III over the total sample size. As described by Friede et al. [8] under $H_0^{\{S,F\}}$, the weighted test statistics are correlated and follow a Multi-Normal distribution, that is in our situation:

$$\begin{pmatrix} S_1^F \\ S_1^S \\ S_2^F \\ S_2^S \end{pmatrix} \sim \mathcal{MN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} w_1^2 & w_1^2\sqrt{\tau} & \rho w_1^2 & \rho w_1^2\sqrt{\tau} \\ w_1^2\sqrt{\tau} & w_1^2 & \rho w_1^2\sqrt{\tau} & \rho w_1^2 \\ \rho w_1^2 & \rho w_1^2\sqrt{\tau} & 1 & \sqrt{\tau} \\ \rho w_1^2\sqrt{\tau} & \rho w_1^2 & \sqrt{\tau} & 1 \end{pmatrix} \right) \tag{3}$$

At interim analysis, early termination of the trial can be considered if the observed test statistic exceeds a fix threshold, similarly to group sequential designs. In our design, we do not consider early termination for efficacy at interim analysis, computation of the threshold value for the final analysis depends only on the observed value at the first stage and the wanted overall type-one error rate. To be declared positive at final analysis, the final statistical test must reach at least a c_2 threshold computed depending on first stage results. This threshold is computed using the multinormal quantile function with probability $= \alpha$ the type-one error rate, and correlation matrix $\begin{pmatrix} 1 & \sqrt{\tau} \\ \sqrt{\tau} & 1 \end{pmatrix}$, where τ represents the prevalence of the subgroup as in Friede et al. [8]. As described in Figure 2, at interim analysis, four decisions can be taken, either the trial stops for futility, or it continues only in F or only in S, or both population are analyzed at the end of the trial. The first stage data is introduced in the final test thanks to

the described method.

If only a population is selected for the final analysis, the individual population null hypothesis H_0^F or H_0^S is then rejected if it individual test statistics values is higher than the corrected critical value that is $Z_2^F > \frac{c_2 - s_1^F \rho}{w_2}$ or $Z_2^S > \frac{c_2 - s_1^S \rho}{w_2}$.

If both population are selected for final analysis, then the intersection hypothesis also needs to be rejected. For the intersection hypothesis, the critical p-value threshold p_{thresh} is computed through the following equation given stage 1 test results s_1^F and s_1^S: $p_{thresh} = 1 - P(Z_2^F < \frac{c_2 - s_1^F \rho}{w_2}, Z_2^S < \frac{c_2 - s_1^S \rho}{w_2})$. Letting $Z_{max} = \max(Z_2^F, Z_2^S)$ be the maximum observed test statistic, the intersection p-value is $p^{S,F} = 1 - \int_{-\infty}^{Z_{max}} \Phi(\frac{Z_{max} - \sqrt{\tau} z}{\sqrt{1-\tau}}) \phi(z) dz$. The intersection hypothesis is then rejected if $p^{S,F} < p_{thresh}$. Finally, a closed test procedure is considered for rejecting H_0^F or H_0^S, meaning that individual and intersection hypothesis need to be rejected to conclude to a significant treatment effect in either of the populations.

3.4.2 Frequentist design

For comparison to Bayesian methods, a frequentist trial is also run. We consider the above method of conditional error function as presented in Friede *et al.* [9] with the *asd* **R** package. Selection is also based on a threshold limit, using statistical test values to select the population. To mimic the Bayesian method, S is selected if $Z_{1,b}^S > \tau_S^*, Z_{1,b}^F < \tau_F^*$, F is selected for a single test on F if $Z_{1,b}^S < \tau_S^*, Z_{1,b}^F > \tau_F^*$, F and S are tested if $Z_{1,b}^S > \tau_S^*, Z_{1,b}^F > \tau_F^*$, and futility is declared at interim analysis if none of the thresholds are crossed. Final analysis is the same as for the Bayesian methods, using CEF and combining statistical tests from both stages.

3.5 Prior distribution and parameter choice

In this work we wanted to evaluate several prior distributions. We first chose two horseshoe prior parametrization [15]. The horseshoe prior is know for shrinking posterior distributions of non influential covariates toward zero. For $k = \{S, T, TS\}$, $\theta_k | \lambda_k, \tau_k \sim \mathcal{N}(0, \tau_k^2 \lambda_k^2)$, with $\lambda_k, \tau_k^2 \sim Half-T()$. We compare two different parametrization of the horseshoe prior distribution, one with a prior distribution giving more weight to zero, and the other being more spread. We refer as the centered on zero horseshoe prior as "peaked horseshoe" while the other one is referred as "flatter horseshoe". The peaked horseshoe prior as $\lambda_k \sim Half-T(0, 1, 1), \tau_k \sim Half-T(0, 1, 1)$,

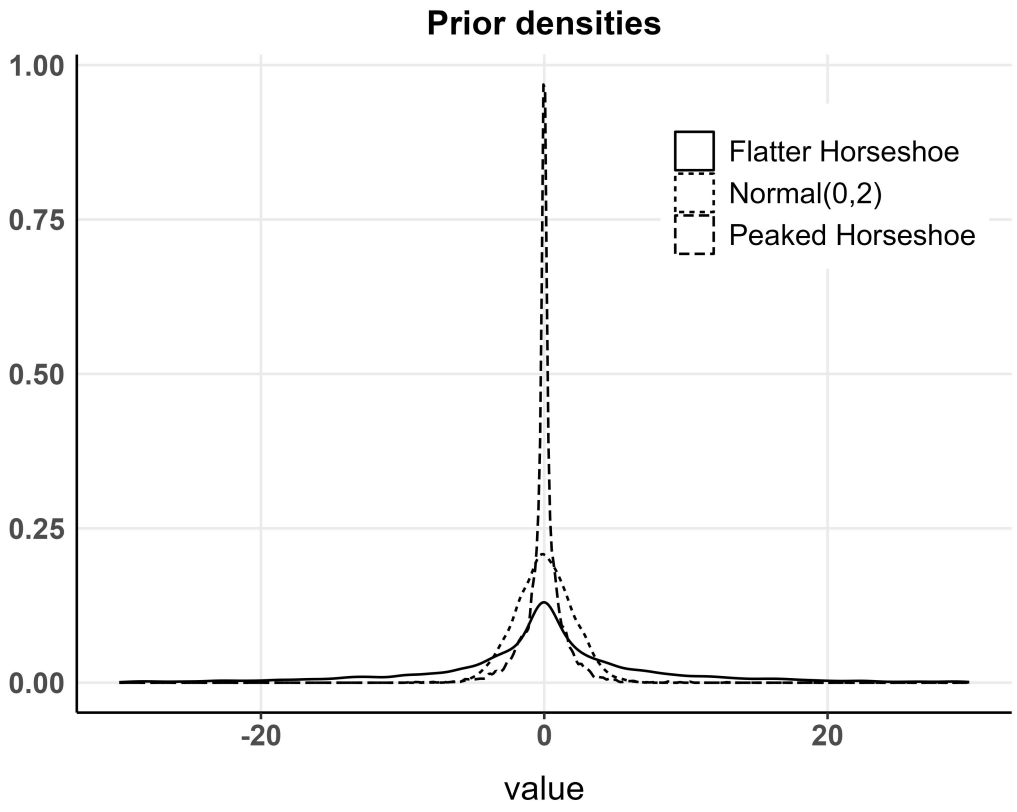

Figure 3: Prior distribution of regression parameter for all θ in the Bayesian design

whereas $\tau_k \sim Half - T(0, 10, 1)$ in the "flatter horseshoe". We also checked another Bayesian prior, that is a Normal non informative $\mathcal{N}(0, 2)$ prior distribution for regression parameters. Samples from the 3 prior densities are plotted in described in Figure 3. Finally, for the sake of simplicity we assumed that ρ is known and equal to 0.8 (the correlation we saw in preliminary sensitivity analysis).

4 Simulation study and application to the case study

4.1 Simulation setting

We evaluate the operating characteristics of the seamless hybrid design and its ability of correctly choosing the population that benefits the most from the treatment. Let N_F, N_S and N_{S^C} be the sample size for F, S and S^C populations. A global sample size of $N_F = 190$ is used along with a prevalence of 0.7 for the subgroup S that is $N_S = .7 * 190 = 133, N_{S^C} = 57$. In the case study trial, the identified subgroup had a prevalence around 0.6. Therefore, we restricted our research to high prevalence simulations, as we wanted to respect that S has a bigger sample

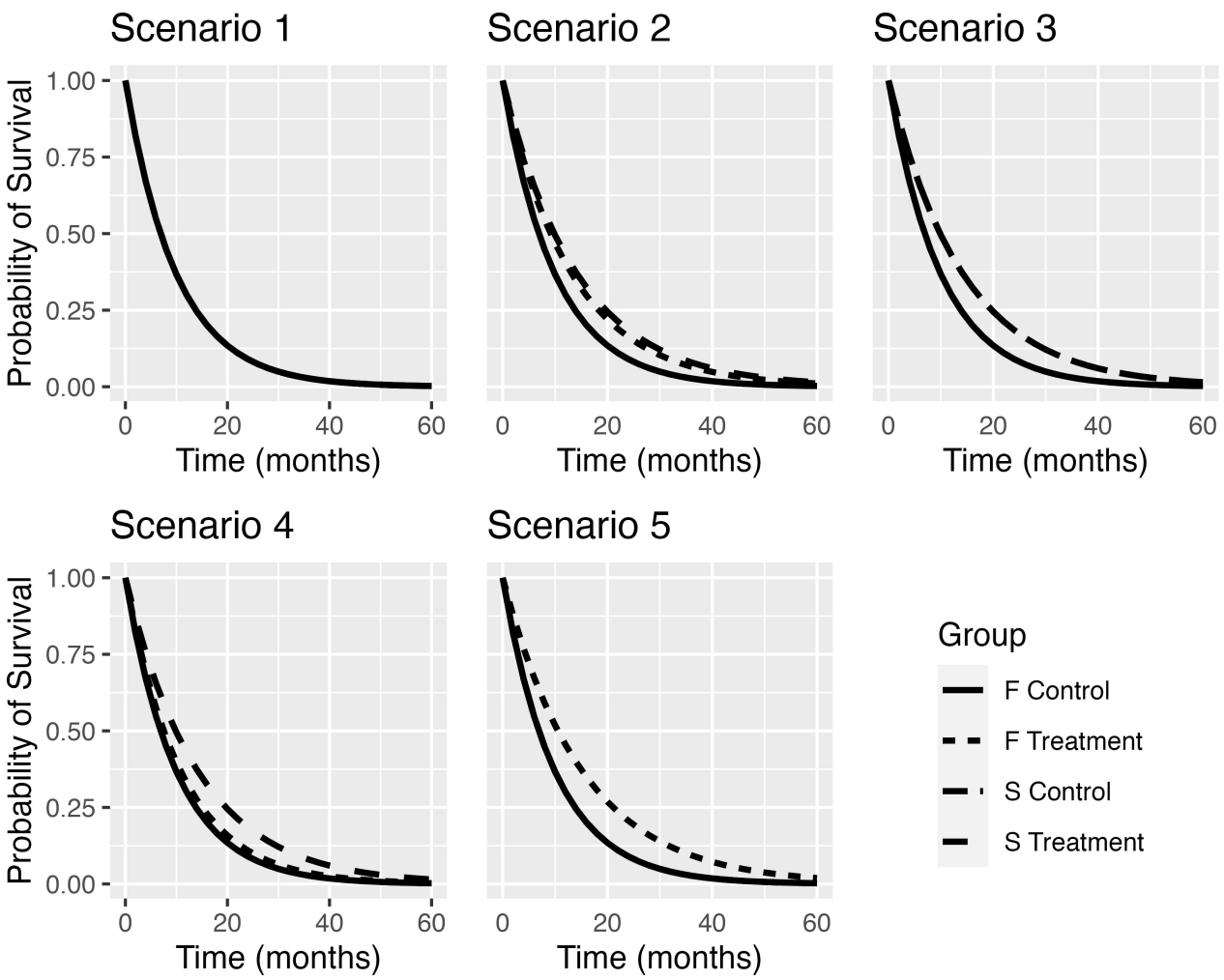

Figure 4: Survival Times Distribution for F and S population in all scenarios

size than S^C. We present results for a S prevalence of 0.7, since other studies of interests, not described here had similar prevalence values, but results for a lower prevalence of the subgroup are presented in appendix. Sample size was computed by simulation for a False Positive rate controlled at 0.05 (bilateral) and a power of 0.8 under the second scenario in which S should be selected with an Hazard ratio of 0.7 for frequentist analysis, considering $\rho = 1$. Hazard ratio were set for both population of the treatment group in comparison to the control. As risk are proportional in the S subgroup as well as in its complementary, it is not expected to observe risk proportionality between S and its complementary in the control group.

As described in the figure 4, the first scenario, of "Futility" scenario, was set up to study the ability of the design to conclude to a futile study when the treatment arm does not benefit to F nor S. In a second scenario, the S subgroup has an increased benefit from the treatment compared to its complementary, $HR(S) = 0.7$, $HR(F) = 0.76$. The third scenario has the F

Table 2: Simulations Scenarios. In column the scenario number, the Hazard ratio Treatment vs control of the Full population and the HR for the subgroup, where PST, PFT, PSC, PFC are respectively the one year survival rates in Treatment Subgroup, Treatment Full population and Control Subgroup and Full population.

Scenario	HR(F)	HR(S)	PST	PFT	PSC	PFC	Population to select
1	1	1	30%	30%	30%	30%	Futility
2	0.76	0.7	39%	42%	30%	30%	S, F or S & F
3	0.7	0.7	42%	42%	30%	30%	F & S
4	1.05	0.7	29%	42%	30%	30%	S
5	0.75	1	39%	30%	30%	30%	F

population having the same treatment effect in both subgroups, that is $HR(F) = HR(S) = 0.7$. In the fourth one, the hazard ratio of the full population $HR(F) = 1.05$ and $HR(S) = 0.7$, with a strong negative effect in the complementary and still the same effect in the subgroup. In that case, since S has a high prevalence, in order reach a hazard ratio in superior to one in F (meaning that the non subgroup patients a high HR). Finally in the last scenario the S population did not benefit from the treatment, but its complementary did, that would be a case where the population was badly identified. $HR(S) = 1$, and $HR(F) = 0.75$.

4.2 Results

Results of all methods are presented in Tables 3, and 4. In Table 3 results are shown for a prevalence of 0.7 of the S subpopulation, the τ_1 futility threshold is set to 0 and threshold for F and for S are respectively $\tau_F = 0.7$ and $\tau_S = 0.5$. In the Table 4, we add a stricter rule to Bayesian methods as the τ_1 threshold that is set to 0.4.

For each scenario, the proportion of futile studies at interim analysis is presented among with the percentage of selection of each population in the 1000 simulated trials. The Ftot and Stot columns represent the proportion of studies in which F and S were selected and were significant at the final analysis, finally the overall power column is the global percentage of positive studies, defined as trials in which the final analysis detects a significant effect of the treatment in (at least one of) the selected population(s). In scenario 1, we see that the type one error rate is only controled by the frequentist methods as they often declare more futility at interim and selects both $F\&S$ population when the study continues, leading to a more strict testing procedure at final analysis. Bayesian methods show a slight inflation (up to 1.5%) of the false positive rate compared to frequentist. For scenario 2, frequentist methods perform

less well than Bayesian horseshoe prior model with around 5% less overall power. Both type of methods are outperformed by the Bayesian Normal model, as this method is less strict for futility at interim analysis, and selects F for the phase III. In the third scenario, frequentist methods with no threshold on F are better in terms of power than both horseshoe Bayesian designs, and more generally frequentist and horsehoe tend to have comparable power. As previously, Normal prior Bayesian model is the most powerful. The last two scenarios, in which treatment effect is only simulated in one studied of the population, we observe a large gap between Bayesian and frequentist methods. In the fourth scenario for instance power drops significantly for all frequentist methods. When the threshold is put on F, we observe higher power for the frequentist design. Bayesian methods all select the right population with high accuracy and are equivalent to each other. Finally in scenario 5 we observe the opposite effect of scenario 4 as the frequentist model with threshold on S is better than the two others and Bayesian methods almost always manage to reject the null hypothesis by selecting F

From Table 4, we can see that having a more stringent futility rule by setting $\tau = 0.4$ rises the proportion of futile studies. While it has a very limited impact on most of the scenario, it is interesting to note that for scenario 4, where effect is large on S, the peaked horseshoe prior model is the best. That is because $P(\theta_T > 0)$ is unlikely to reach the thresholds with the posterior being concentrated around zero, while it is reached for $P(\theta_{TS} > 0)$ and $P(\theta_T + \theta_{TS} > 0)$ in a little bit more cases, leading to more selection of S and then to a little increase in terms of power. The scenario 3 does not change a lot with the addition of τ_1, as both population are simulated efficient, it leads to good posterior probability of being superior to zero for all terms and therefore this scenario is not affected by the threshold as the last one which has such a strong effect that the threshold is easily reached.

Table 3: Results of the simulated scenarios for a prevalence of 0.7 of the Subgroup. $\tau_F = 0.7, \tau_S = 0.5, \tau_1 = 0$. The $\%$ columns indicate the percentages of selection at interim analysis for Futility, $F, S, S\&F$ with the percentage (of them) of positive studies at final analysis in parentheses. The Ftot and Stot columns respectively represent the percentage of studies with selection and significant testing in F and S. The False Positive Rate refers to the percentage of studies erroneously declared as positive while the Overall Power column is the global percentage of studies with at least one right population selected with positive result.

Scenario	Model	Futility	F	S	S&F	Ftot	Stot	False Positive Rate
	Peaked horseshoe	**48.1**	5.7(9%)	40.1(5)	6.1(8)	1	2.5	3
	Flatter Horseshoe	**45.5**	8.8(9%)	39.8(5%)	5.9(5)	1.1	2.3	3.1
	Normal(0,2)	**27**	26.9(6%)	43.1(5%)	3(7)	1.8	2.3	3.9
1	Freq($\tau_F^* = \tau_S^* = 0$)	**44.3**	9(4%)	8.3(4%)	38.4(4%)	1.9	1.8	2.2
	Freq($\tau_F^* = 0, \tau_S^* = 0.5$)	**49.6**	20.2(5%)	3(3%)	27.2(5%)	2.4	1.5	2.5
	Freq($\tau_F^* = 0.5, \tau_S^* = 0$)	**50.5**	2.8(4%)	18.3(3%)	28.4(5%)	1.5	1.9	2

Scenario	Model	Futility	F	S	S&F	Ftot	Stot	Overall Power
	Peaked horseshoe	12.5	6.6(89%)	**54.7**(**80**%)	26.2(83%)	27.7	65.4	71.3
	Flatter Horseshoe	11.8	9.9(88%)	**53.4**(**80**%)	24.9(84%)	29.6	63.5	72.2
	Normal(0,2)	7.1	28.2(83)	**50.2**(**80**%)	14.5(88%)	36.1	52.9	76.3
2	Freq($\tau_F^* = \tau_S^* = 0$)	9.9	4.6(72%)	**4.7**(**64**%)	80.8(76%)	64.7	64.4	67.7
	Freq($\tau_F^* = 0, \tau_S^* = 0.5$)	13.4	13(68)	**1.2**(**67**%)	72.4(78)	65.2	57.2	66
	Freq($\tau_F^* = 0.5, \tau_S^* = 0$)	13.2	1.3(77%)	**15.9**(**70**%)	69.6(79%)	56	66.2	67.2
	Peaked horseshoe	14.2	17.9(94%)	34.8(78%)	**33.1**(**90**%)	46.8	57	73.9
	Flatter Horseshoe	12.5	25.4(93%)	33.7(79%)	**28.4**(**90**%)	49.2	52.2	75.7
	Normal(0,2)	7.2	43.5(92%)	31.5(79%)	**17.8**(**93**%)	56.6	41.5	81.6
3	Freq($\tau_F^* = \tau_S^* = 0$)	7.5	7(89%)	1.9(68%)	**83.6**(**81**%)	74.3	69.4	75.6
	Freq($\tau_F^* = 0, \tau_S^* = 0.5$)	8.8	17.6(83%)	0.6(67%)	**73**(**84**%)	76	61.8	76.4
	Freq($\tau_F^* = 0.5, \tau_S^* = 0$)	11.4	3.1(90%)	8.9(64%)	**76.6**(**84**%)	66.8	69.7	72.5
	Peaked horseshoe	2.7	0(0)	**95.8**(**79**%)	1.5(33%)	0.5	76.4	76.4
	Flatter Horseshoe	2.4	0.3(33%)	**95.4**(**79**%)	1.9(42%)	0.9	76.2	76.3
	Normal(0,2)	2	0.8(25%)	**95.6**(**79**%)	1.6(31%)	0.7	76.5	76.7
4	Freq($\tau_F^* = \tau_S^* = 0$)	14.3	0.2(0)	**23.6**(**74**%)	61.9(26)	16.4	33.8	33.8
	Freq($\tau_F^* = 0, \tau_S^* = 0.5$)	24.3	2.1(5%)	**13.6**(**79**%)	60(27)	16.4	27.1	27.2
	Freq($\tau_F^* = 0.5, \tau_S^* = 0$)	14.5	0	**41.9**(**77**%)	43.6(31%)	13.7	46	46
	Peaked horseshoe	2.6	**96.1**(**100**%)	0.2(0)	1(18%)	96.1	0.2	96.1
	Flatter Horseshoe	1.7	**97.1**(**100**%)	0.3(33%)	0.9(22%)	97.1	0.3	97.2
	Normal(0,2)	0.1	**99.4**(**100**%)	0.2(0)	0.3(67%)	99.3	0.2	99.3
5	Freq($\tau_F^* = \tau_S^* = 0$)	4.5	**48.8**(**100**%)	0	46.7(6%)	51.5	2.9	51.5
	Freq($\tau_F^* = 0, \tau_S^* = 0.5$)	4.5	**65.3**(**100**%)	0	30.2(8%)	67.5	2.4	67.5
	Freq($\tau_F^* = 0.5, \tau_S^* = 0$)	11	**42.3**(**100**%)	0.2(0)	46.5(6%)	45.2	2.9	45.2

Table 4: Results of the simulated scenarios for a prevalence of 0.7 of the Subgroup. $\tau_F = 0.7, \tau_S = 0.5, \tau_1 = 0.4$. The % columns indicate the percentages of selection at interim analysis for Futility, $F, S, S\&F$ with the percentage (of them) of positive studies at final analysis in parentheses. The Ftot and Stot columns respectively represent the percentage of studies with selection and significant testing in F and S. The False Positive Rate refers to the percentage of studies erroneously declared as positive while the Overall Power column is the global percentage of studies with at least one right population selected with positive result.

Scenario	Model	Futility	F	S	S&F	Ftot	Stot	False Positive Rate
1	Peaked horseshoe	**63.5**	6.3(9%)	11.3(6%)	18.9(8%)	1	2.5	3
	Flatter Horseshoe	**51.3**	8.8(9%)	34(6%)	5.9(5%)	1.1	2.3	3.1
	Normal(0,2)	**38.6**	26.9(6%)	31.5(6%)	3(7%)	1.8	2.2	3.8

Scenario	Model	Futility	F	S	S&F	Ftot	Stot	Overall Power
2	Peaked horseshoe	14	6.6(89%)	**53.2**(**81**%)	26.2(83%)	27.7	64.8	70.7
	Flatter Horseshoe	13.9	9.9(88%)	**51.3**(**81**%)	24.9(84%)	29.6	62.6	71.3
	Normal(0,2)	10.3	28.2(83%)	**47**(**83**%)	14.5(88%)	36.1	51.8	75.2
3	Peaked horseshoe	14.4	17.9(94%)	34.6(78%)	**33.1**(**90**%)	67.4	64.9	73.7
	Flatter Horseshoe	12.8	25.4(93%)	33.4(79%)	**28.4**(**90**%)	49.2	52	75.5
	Normal(0,2)	8.8	43.5(92%)	29.9(82%)	**17.8**(**93**%)	56.6	41.1	81.2
4	Peaked horseshoe	9.9	0(0)	**88.6**(**82**%)	1.5(33%)	0.5	73.3	73.3
	Flatter Horseshoe	10.5	0.3(33%)	**87.3**(**82**%)	1.9(42%)	0.9	72.6	72.7
	Normal(0,2)	11.5	0.8(25%)	**86.1**(**83**%)	1.6(31%)	0.7	71.8	72
5	Peaked horseshoe	2.6	**96.1**(**99.9**%)	0.2(0)	1.1(18%)	96.1	0.2	96.1
	Flatter Horseshoe	1.7	**97.1**(**100**%)	0.3(33%)	0.9(22%)	97.1	0.3	97.2
	Normal(0,2)	0.1	**99.4**(**100**%)	0.2(0)	0.3(67%)	99.3	0.2	99.3

4.3 Application to the Atalante-1 study data

In this subsection, we briefly present and discuss an application of Bayesian and frequentist methods to the Atalante-1 study data. The original trial did not plan for selection at interim analysis and therefore pursued the study in the full population, and did not consider testing the subgroup. When we apply the proposed method to the trial data, each Bayesian method selects the S population at interim analysis while frequentist designs go on in $F\&S$ populations. Bayesian posterior distribution of θ_T and θ_{TS} are plotted in Figure**??**. Posterior probability for θ_T and for θ_{TS} of being superior to zero are reported in Table 5. Frequentist methods all go with $F\&S$ at interim analysis as the value of Z_F reaches a high enough value. With all the designs, the final analysis leads to a futile study with a final p-value of 0.18 in the subgroup. It was anticipated, given that the original design was not calibrated for these methods, resulting in a sample size that was not appropriately tailored.

Table 5: Results obtained using the Atalante-1 dataset.

Method	$\mathbb{P}(\theta_T > 0)$	$\mathbb{P}(\theta_{TS} > 0)$	$Z^F_{1,b}$	$Z^S_{1,b}$	p^F	p^S	$p^{S,F}$
Peaked Horseshoe	0.49	0.51	0.893	1.01	NA	0.18	NA
Flatter Horseshoe	0.49	0.51	0.893	1.01	NA	0.18	NA
Normal(0,2)	0.58	0.74	0.893	1.01	NA	0.18	NA
Frequentist	NA	NA	0.893	1.01	0.4	0.18	0.24

5 Discussion

In this paper, we compared the operating characteristics of three Bayesian models and a Frequentist one with diverse thresholds in the context of seamless phase II/III trial with selection of a population at interim analysis. We compared multiple selection criteria and model while having a similar (frequentist) testing procedure at the end of the trial.

From the results we saw that for Bayesian distributions, the horseshoe prior models are the one declaring the more futile studies at interim. The horseshoe prior is known to be optimal in variable selection when there is an important number of covariates, as only a few are selected and other are shrinked to zero. In this context of sub-population selection horseshoe appears to be less appealing than simpler normal prior distribution as the method is often too conservative declaring futility at interim analysis. Simple prior like the normal prior distribution can be preferable in those situations, satisfying results of this prior distribution also come with a reduced computational time and better MCMC convergence. Model selection method as Bayesian Model Averaging or Weighted AIC were also studied, but were not included as first results showed no gain of use.

On the other hand the normal prior distribution can inflate a type one error rate as it can lack of conservatism at interim analysis, with only a few studies stopped. For efficacy scenarios normal prior tends to select more only F population rather than $F\&S$ as the horseshoe. This results in a higher power as the final test is only based on F tests and is therefore less conservative. More generally, selection of the F population at interim analysis when all patients benefit from the treatment will lead to a higher overall power, as it uses the total number of patients and only apply the testing procedure to one group. Whereas the testing in S suffers from a more restricted sample size, and rejection of $H_0^{\{S,F\}}$ implies rejection of H_0^F and H_0^S.

Simulation studies also showed that in some situations, the use of the frequentist statistical test value to chose between the full population and a subgroup can lead to erroneously chose

to continue with the analysis of both population. In particular when a strong effect of the treatment is observed on the subgroup, leading to an artificially inflate Z^F value. For example, in scenarios 4 and 5, when only one of the studied population benefits from the treatment, we observe an important difference on overall power between both types of methods. In scenario 4, $Z^F_{1,b}$ value manages to reach the τ^*_F threshold as $S \subset F$, this implies an analysis in a non responding population and a more strict rule for final analysis (as tests need to be performed in both population). For the scenario 5, the opposite effect is observed as the $HR = 1$ for the S subgroup, a threshold τ^*_S of zero can be often reach leading to a selection of both population implying a loss of power for the same reasons of scenario 4.

The use of a logistic regression allow to overcome this issue. If a significant effect is found on S in the treatment group, then the coefficient associated to the interaction term will be far from zero while the coefficient of the full population will be close to zero.

Another difference between results from the Bayesian and the frequentist methods lies in the selection rules. Indeed, if we imagine to apply some mathematical transformations we could see the Bayesian thresholds on probabilities transferred on the Z scales, to match the frequentist counterparts. Even if a perfect correspondence between Bayesian and frequentist thresholds may be found, the Bayesian methods adopt an additional level of selection by using τ_1. This could be the reason why the frequentist methods select more often the $F\&S$ populations instead of the S alone.

Other possible designs include continuing with S^C if a null effect where found in S. But in our context, since the subgroup is supposed to be identified before the trial, we did not consider that option. Sample size re-estimation is also a common adaptation in that context. We did not study the optimal sample size as computation for all the methods is time consuming and a lot more simulations would be needed in that case. But the larger the sample size is for interim analysis, the better is the choice of the population. However for fixed sample size designs with no reassessment, having interim analysis done later improve the percentage of correct population selection, but it does not improve the power by much in scenarios where S is selected because the final sample size for S would be the same as when interim analysis is done with a lower sample size. In that case the small gain observed is due to a better selection accuracy.

When planing a seamless adaptive design with population selection, we recommend that the subgroup studied as a large enough prevalence, in particular if no sample size reassessment is done after interim analysis when the subgroup is selected.

References

[1] Jim Albert and Monika Hu. Probability and Bayesian Modeling. CRC Press, Taylor & Francis Group, 2020.

[2] Besse B, Felip E, Garcia Campelo R, Cobo M, Mascaux C, Madroszyk A, Cappuzzo F, Hilgers W, Romano G, Denis F, Viteri S, Debieuvre D, Galetta D, Baldini E, Razaq M, Robinet G, Maio M, Delmonte A, Roch B, Masson P, Schuette W, Zer A, Remon J, Costantini D, Vasseur B, Dziadziuszko R, and Giaccone G. Atalante-1 study group. randomized open-label controlled study of cancer vaccine ose2101 versus chemotherapy in hla-a2-positive patients with advanced non-small-cell lung cancer with resistance to immunotherapy: Atalante-1. Ann Oncol., pages 920–933, 10 2023.

[3] Nicolas Ballarini, Thomas Burnett, Thomas Jaki, Christoper Jennison, Franz König, and Martin Posch. Optimizing subgroup selection in two-stage adaptive enrichment and umbrella designs. Statistics in Medicine, 40, 03 2021.

[4] Laura E Bothwell, Jerry Avorn, Nazleen F Khan, and Aaron S Kesselheim. Adaptive design clinical trials: a review of the literature and clinicaltrials.gov. BMJ Open, 8(2), 2018.

[5] Werner Brannath, Emmanuel Zuber, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, and Amy Racine-Poon. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. Statistics in Medicine, 28(10):1445–1463, 2009.

[6] Ting-Yu Chen, Jing Zhao, Linda Sun, and Keaven Anderson. Multiplicity for a group sequential trial with biomarker subpopulations. Contemporary Clinical Trials, 101:106249, 12 2020.

[7] Lu Cui, HM James Hung, and Sue-Jane Wang. Modification of sample size in group sequential clinical trials. Biometrics, 55(3):853–857, 1999.

[8] Tim Friede, N Parsons, and Nigel Stallard. A conditional error function approach for subgroup selection in adaptive clinical trials. Statistics in Medicine, 31(30):4309–4320, 2012.

[9] Tim Friede, Nigel Stallard, and Nicholas Parsons. Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. Biometrical Journal, 62(5):1264–1283, 2020.

[10] Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharmaceutical Statistics, 10(4):347–356, 2010.

[11] C. Jennison and B.W. Turnbull. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC, 2000.

[12] Peter K Kimani, Ekkehard Glimm, Willi Maurer, Jane L Hutton, and Nigel Stallard. Practical guidelines for adaptive seamless phase II/III clinical trials that use Bayesian methods. Statistics in Medicine, 31(19):2068–2085, 2012.

[13] Baldur P. Magnusson and Bruce W. Turnbull. Group sequential enrichment design incorporating subgroup selection. Statistics in medicine, 32(16):2695–2714, 2013.

[14] Guanhong Miao, Jason Liao, Jing Yang, and Keaven Anderson. A gated group sequential design for seamless phase ii/iii trial with subpopulation selection. BMC Medical Research Methodology, 23, 01 2023.

[15] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics, 11(2), jan 2017.

[16] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. Biometrika, 64(2):191–199, 1977.

[17] Martin Posch, Franz Koenig, Michael Branson, Werner Brannath, Cornelia Dunger-Baldauf, and Peter Bauer. Testing and estimation in flexible group sequential designs with adaptive treatment selection. Statistics in Medicine, 24(24):3697–3714, 2005.

[18] Allan R. Sampson and Michael W. Sill. Drop-the-losers design: Normal case. Biometrical Journal, 47(3):257–268, 2005.

[19] Bart Spiessens and Muriel Debois. Adjusted significance levels for subgroup analyses in clinical trials. Contemporary clinical trials, 31 6:647–56, 2010.

[20] Nigel Stallard, Cornelia Ursula Kunz, Susan Todd, Nicholas Parsons, and Tim Friede. Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial. Statistics in Medicine, 34(23):3104–3115, 2015.

[21] J. Whitehead. The Design and Analysis of Sequential Clinical Trials. Statistics in Practice. Wiley, 1997.