Using dichotomized survival data to construct a prior distribution for a Bayesian seamless Phase II/III clinical trial

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

[1]Universit Paris Cit, Sorbonne Universit 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]Unit of Clinical Epidemiology, Assistance Publique-Hpitaux de Paris, CHU Robert Debr, Inserm CIC-EC 1426, F-75019 Paris, France

[†]Authors made equal contribution

Corresponding author: Sarah Zohar, PariSante Campus, 10 rue Oradour-sur-glane, F-75015, Paris, FR. Email: sarah.zohar@inserm.fr

Abstract

Master protocol designs allow for simultaneous comparison of multiple treatments or disease subgroups. Master protocols can also be designed as seamless studies, in which two or more clinical phases are considered within the same trial. They can be divided into two categories: operationally seamless, in which the two phases are separated into two independent studies, and inferentially seamless, in which the interim analysis is considered an adaptation of the study. Bayesian designs are scarcely studied. Our aim is to propose and compare Bayesian operationally seamless Phase II/III designs using a binary endpoint for the first stage and a time-to-event endpoint for the second stage. At the end of Phase II, arm selection is based on posterior (futility) and predictive (selection) probabilities. The results of the first phase are then incorporated into prior distributions of a time-to-event model. Simulation studies showed that Bayesian operationally seamless designs can approach the inferentially seamless counterpart, allowing for an increasing simulated power with respect to the

operationally frequentist design.

Keywords: Bayesian Confirmatory trials, Different Outcomes, Operating Characteristics, Treatment Selection.

1 Introduction

Master protocols, such as basket, umbrella and platform trials, have revolutionized the way clinical studies are conducted, especially in oncology(Woodcock & LaVange 2017). A recent systematic review by Park et al. (Park et al. 2019) identified 83 master protocols (49 basket, 18 umbrella, and 16 platform trials), and the number of master protocols has been increasing rapidly over the last five years. These studies provide an efficient and flexible methodology for the assessment of multiple interventions in one or more diseases or conditions in a single protocol, possibly in a continuous manner. The benefits of master protocols include the use of a common control, the pooling of resources, and a reduction of the investment required to evaluate additional interventions beyond those initially studied. They also maintain a high quality standard for ongoing evaluations through the establishment of a trial network and may positively impact patient recruitment, as the trials are performed on sites that are continuously learning and improving their performance. However, such studies also present a number of challenges; the upfront investment is large, meaning that the number of platform trials that can be conducted may be relatively limited, and more insight into the prerequisites for the design to become efficient and sustainable is needed. Renfro and Sargent (Renfro & Sargent 2017) have also highlighted the "sample size" limitation versus "effect size", where a small sample size is used for each study arm in order to maintain the overall trial feasibility. This implies that the targeted effect size could be larger than the expected effect size leading to lower power and/or higher type I error than usual Phase II or III trials.

Another interesting design class is composed of the so-called seamless design. Traditionally, separate Phase II and Phase III trials are carried out with specific objectives during clinical drug development. For example, a late Phase II trial may be used to estimate parameters to be used in sample size calculation for a confirmatory Phase III trial. In contrast to traditional clinical trials, seamless clinical trials allow for the combination of multiple trial phases inside a single protocol (Bretz et al. 2006). If trial phases are done sequentially but separately, they are called "operationally seamless". Conversely, in the adaptive seamless design, the final analysis uses data from patients enrolled in all phases, before and after adaptation, in an "inferentially seamless" approach (Maca et al. 2006). Decisions on how to "adapt" the study are made after taking planned interim views of the data. Inferentially seamless designs come with statistical challenges since multiple comparisons arise from both (possible) repeated in-

terim looks at gathered data and the selection process (selected data will also be included in the final analysis) (Stallard 2011). This approach usually requires that the familywise error rate be controlled in the strong sense since pairwise tests are performed between the arms at the selection stage. To address these challenges, many of the methods proposed in the literature are based on group sequential approaches (Stallard 2011, Stallard et al. 2015), on combination test approaches (Cui et al. 1999, Quan et al. 2020) or on the conditional error function method (Friede et al. 2012, 2020). Selection procedures based on utility functions (Aouni et al. 2021), on benefit cost ratio perspective (Sun et al. 2020) and/or on conditional power (Kimani et al. 2009) have also been explored.

Bayesian inference naturally fits seamless (and master) designs, permitting the inclusion of Phase II data into prior distributions for the Phase III model parameters, sharing information between trial phases in operationally seamless designs, or enabling frequentist multiplicity corrections but working on prior distributions Chang & Berger (2021) and/or on thresholds on posterior values Aupiais et al. (2019) in inferentially seamless designs. Moreover, in Bayesian thinking, operationally seamless designs can mathematically approach inferentially seamless ones. The information, whether added in prior distributions, as in the operationally seamless, or in the likelihood function, as in the inferentially seamless, can have the same mathematical role in the Bayes formula. If the power prior approach Ibrahim et al. (2015) is used it is straightforward to incorporate Phase II data into Phase III analysis: if the power prior parameter is set to one, all information is gathered, and if the same outcome is used in both phases, the posterior distribution of an inferentially seamless and of an operationally seamless coincide. However, while Bayesian inference has been widely used in seamless Phase I/II, only a few attempts have been made to include Bayesian inference in the seamless design of Phase II/III (Chapple & Thall 2019). For example, Bayesian tools were proposed to perform treatment(s) (Schmidli et al. 2007, Kimani et al. 2009) or subpopulation(s) (Brannath et al. 2009) selection. In Kimani et al. Kimani et al. (2012), Bayesian estimation of the dose-response curve was adopted at the end of Phase II; however, frequentist analysis was then used to test the Phase III hypotheses on treatment superiority.

Another important feature of seamless studies is the use of a generally shorter term endpoint for the first stage of the study. Usually, this midterm endpoint can be considered as a surrogate endpoint for the Phase III study. A few works have studied the use of different but correlated endpoints in the two phases, under a frequentist paradigm. Jenkins et al. (2010), Stallard (2010)

In this work, we propose and evaluate simple Bayesian operationally seamless Phase II/III designs for survival analysis. As in the master protocol, the design allows for several treatments to be compared to a single control arm. As in seamless design, selection rules are specified, and only the selected arm proceeds to Phase III. Moreover, our work focuses on the situation where two different, but related, endpoints are used in the two phases, that is, when a dichotomized

survival outcome, such as a survival rate at a fixed time point, is used to make decision at the end of the Phase II. In this setting, we have developed two ways of incorporating Phase II information into the Phase III analysis when endpoints differ but are related as in the case study used for this paper. Indeed, we proposed to tune prior distributions based on the Effective Sample Size Morita et al. (2008) or on a likelihood approach. Since seamless designs plan several phases in the same protocol that are usually done in similar populations and environmental settings, we expect prior distributions on the latest phase to include the maximum possible information from previous stages. Adding information into prior distribution is usual in Bayesian framework, even if still not explored in previous work on seamless design. We explore the performance of several Bayesian methods, that is, with weakly informative prior distributions, as well as with informative ones, in terms of frequentist operating characteristics through extensive simulation studies. The aim is also to understand how close Bayesian operationally seamless designs performances, in term of frequentist operating characteristics, can get to Bayesian inferentially seamless ones. Therefore, an inferentially seamless Bayesian design, that uses data from both phases and with the true survival time of Phase II patients not dichotomized for the Phase III analysis, is introduced for comparison only. A frequentist operationally design is also introduced for comparison only.

In the next section, the motivating example is introduced. In Section 3, the methods of each phase, selection rules and final claim rule are described. Simulation design setups and results are summarized in Section 4. A practical example and discussion are then shown in Sections 5 and 6, respectively.

2 Motivating study

This work was motivated by the Atalante-1 clinical trial (NCT02654587). This study on non-small-cell lung cancer was designed as an operationally seamless Phase II/III trial comparing the efficacy of an experimental treatment (Tedopi) against the best standard of care (Docetaxel or Pemetrexed). A frequentist approach was used with a Fleming single-arm design (Fleming 1982) for the Phase II stage, considering only the treatment arm. For the first stage, a binary endpoint, the survival at 12 months, was chosen. With a type I error rate of 2.5% and a power of 80%, the study included 84 patients in its first stage. The null hypothesis H0 was a 25% survival rate, and the alternative hypothesis H1 was a 40% survival rate. If the null hypothesis was rejected at the end of Phase II, the study could continue to the Phase III stage; otherwise, the trial would be stopped due to futility. For the second stage, using a 2:1 randomization and a survival endpoint (overall survival) with a two-sided log-rank test at the 5% significance level with a power of 80%, 363 new patients were planned to be included in the trial, and 278 events were needed assuming median OS of 7 months in the control group and 10 months in the experimental group under

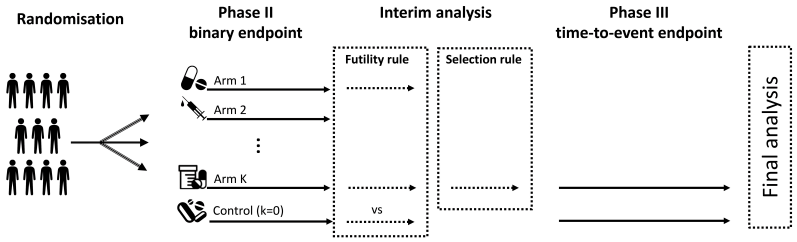

Figure 1: Scheme of the seamless design, with a Phase II involving binary outcomes, that is, mortality rate at a prespecified time, and a Phase III with a time-to-event endpoint.

the alternative hypothesis. The Atalante-1 study was discontinued due to the COVID-19 pandemic Kunz et al. (2020) based on the recommendations of the DSMB, when 219 patients had been randomized and enrolled. Besse (n.d.) The design proposed in this manuscript is based on a Bayesian generalization of the Atalante-1 trial.

3 Methods

Based on the Atalante-1 case study, our proposed Bayesian designs use the same endpoints, that is, binary outcomes at the Phase II stage and survival outcomes at the Phase III stage. However, we propose extending the design to allow for multiple treatment arms at the first stage, as shown in Figure 1. For the sake of simplicity, we assumed a 1:1 randomization ratio in both stages. In the following, notation is introduced along with the proposed mathematical models and statistical rules at each stage.

3.1 Notation

Let k, $k = 0, ..., K$, be the index of the K treatments involved in the trial, with $k = 0$ representing the control group. Let N_k be the maximum sample size of the Phase II stage for each group and $M_{\tilde{k}}$ the maximum sample size of Phase III for the selected arm, $\tilde{k} = 1$ and for the control arm, $\tilde{k} = 0$. Let $t_{i,k}, i = 1, ..., N_k$, and $t_{i,\tilde{k}}, i = 1, ..., M_{\tilde{k}}$ be the time to event (death) for Phase II and Phase III, respectively, and $c_{i,k}$ and $c_{i,\tilde{k}}$ be the censoring time for each individual. For each patient in the trial, we will observe either the event or the censoring time (if the patient is lost to follow-up or alive at the end of the observation window). For each individual in Phase II, $y_{i,k}$ denotes the time of the first occurrence of the event or censoring, and $\nu_{i,k}$ is the event indicator, that is, $y_{i,k} = \min(t_{i,k}, c_{i,k})$

with $t_{i,k}, c_{i,k} > 0$ and

$$\nu_{i,k} = \begin{cases} 1 \text{ if } y_{i,k} = t_{i,k} \Longleftrightarrow t_{i,k} \leq c_{i,k} \\ 0 \text{ if } y_{i,k} = c_{i,k} \Longleftrightarrow t_{i,k} > c_{i,k} \end{cases}.$$

Let $\mathcal{D}_k$ be the data from group k, $\mathcal{D}_k = \{N_k, \mathbf{y}_k, \boldsymbol{\nu}_k\}_{k=0,\dots K}$, where $\boldsymbol{\nu}_k$ and $\mathbf{y}_k$ are vectors of length N_k containing all values of $y_{i,k}$ and $\nu_{i,k}$.

Similarly, $\mathcal{D}_{\tilde{k}}$ is defined for $\tilde{k} = 0, 1$ in Phase III.

3.2 Stage 1 - Phase II

A binary primary endpoint (the survival at t^*, with $t^* = 12$ months in our example) is considered at the first stage. The two possible outcomes are dying before t^* or still being alive at that time. For simplicity, patients censored before t^* are excluded from the survival rate calculation.

Phase II analysis is planned when N_k patients are recruited in each k group and have finished the follow-up period. In practice, the first stage sample size, N_k, will depend on the maximum sample allowed for the study (usually depending on external criteria) and on the design performance. After setting the maximum N_k allowed for each arm, the sample size could be calibrated using simulations to derive operating characteristics, of the stage and of the whole trial, such as simulated type I error, power, percentage of correct arm selection, etc.. Then, N_k could be chosen to maximize desired criteria, such as power subject to sufficiently small simulated type I error.

Let $y^*_{i,k}$ be the indicator of survival at time t^* for patient i in group k, that is,

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

with $i \in 1, .., n_k$, where $n_k \leq N_K$ denotes the number of noncensored patients in group k.

Let $p_k \in [0, 1]$ be the probability of being alive at time t^* in group k; then, we have $\sum y^*_{i,k} \sim \text{Binomial}(n_k, p_k)$. We use the logit link function to study the inference on p_k as recommended by Albert and Hu Albert & Hu (2020). The model can be written as $p_k = \text{logit}^{-1}(\theta_k)$, with $\theta_k = \theta_0 + \mu_k$, $k = 1, \dots, K$, and where θ_0 is the parameter associated with the control arm ($p_0 = \text{logit}^{-1}(\theta_0)$). To complete the Bayesian model, normal prior distributions can be given with $\mu_k \sim \mathcal{N}(\tilde{\mu}_k, \sigma_k^2)$, and $\theta_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Hyperparameter values can be set using historical data available for the control group, and $\tilde{\mu}_k$ could be set to 0 if a conservative no-effect prior is preferred. Notably, a simpler beta-binomial model could also be used for each arm at this stage. Similar results were found using

a beta-binomial conjugate distribution for the first stage of the study. However, the proposed model enables a faster interpretation of the treatment effect and for possible extensions (e.g., hierarchical model, adding doses within treatment groups and a dose-response underlying curve for some groups, etc.), as discussed in Section 6. All Bayesian methods presented in the following sections of this paper use the same prior distribution for the Phase II stage and differ only in the final analysis where prior and data specifications depend on the model used.

3.3 Interim selection rules

At the end of Phase II, we are interested in selecting the most promising arm among the treatment groups. We propose a two-step algorithm: (i) applying a futility rule and (ii) selecting the most promising arm among those retained in the previous step. In the first step, all arms that did not reach a threshold of τ_1 for the posterior probability of having a higher survival rate than the control, that is, $\mathbb{P}(p_k - p_0 > 0|\mathcal{D}_k) < \tau_1$, were excluded from the study. In the second step, among all remaining treatment arms, the one with the highest predictive probability of success is selected for the survival part of the study. This probability is defined as the probability that a future patient reaches at least one year of survival in the arm (posterior expected value of p_k). If none of the arms is selected because of the futility rule, the trial is stopped. The futility rule cannot change the best arm selection, but it is used to stop the whole trial if no evidence of benefit is gathered at the Phase II stage. Moreover, the two steps can be reversed without changing the results.

3.4 Stage 2 - Phase III

After selection of the most promising arm, the study continues to Phase III with a survival endpoint. As an example, we use the Weibull distribution to model the survival time. Other parametric distributions could alternatively be used. For regression purposes, we adopted the shape α and scale $\gamma_{\tilde{k}}$ parametrization, $W(\alpha, \gamma_{\tilde{k}})$, that is:

$$f(t|\alpha, \gamma_{\tilde{k}}) = \frac{\alpha}{\gamma_{\tilde{k}}} \left(\frac{t}{\gamma_{\tilde{k}}}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\gamma_{\tilde{k}}}\right)^{\alpha}\right), \text{ with } \alpha > 0,\ \gamma_{\tilde{k}} > 0,\ \tilde{k} = 0, 1.$$

Here, $\tilde{k} = 0$ and $\tilde{k} = 1$ denote the control arm and the selected treatment arm, respectively. As usual in Weibull survival regression, the shape parameter is considered shared by both arms, while each arm is associated with its corresponding scale parameter, $\gamma_{\tilde{k}}$. For each arm, the associated survival function is $S(y_{i,\tilde{k}}|\alpha, \gamma_{\tilde{k}}) = \exp\left(-\left(\frac{y_{i,\tilde{k}}}{\gamma_{\tilde{k}}}\right)^{\alpha}\right)$, and the likelihood of $M_{\tilde{k}}$ accrued patients in the $\tilde{k}$-th arm, including censored data, can be written as $L(\alpha, \gamma_0, \gamma_1|\mathcal{D}) =$

$\prod_{\tilde{k}=0}^{1}\prod_{i=1}^{M\tilde{k}} f(y_{i,\tilde{k}}|\alpha,\gamma_{\tilde{k}})^{\nu_{i,\tilde{k}}} S(y_{i,\tilde{k}}|\alpha,\gamma_{\tilde{k}})^{(1-\nu_{i,\tilde{k}})}$, where $y_{i,\tilde{k}}$, and $\nu_{i,\tilde{k}}$ respectively represents the time to event or censor of individual i in the $\tilde{k}$ group and its event indicator. To make a stable inference, $\beta_{\tilde{k}} = \log(\gamma_{\tilde{k}})$ is estimated for the $\tilde{k}$-th arm.

To complete the Bayesian model, prior distributions on α, β_0, β_1 are needed. In the following, two approaches that use Phase II results are proposed. In the first one, called the ESS approach, following the suggestion of Ibrahim et al. Ibrahim et al. (2001), $\pi(\alpha) = \text{InverseGamma}(\rho, \kappa)$ with $\rho, \kappa > 0$ and $\pi(\beta_{\tilde{k}}) = \mathcal{N}(\mu_{\tilde{k}}, \sigma_{\tilde{k}})$ with the standard deviation $\sigma_{\tilde{k}} > 0$ are chosen as prior distributions. The hyperparameter $\mu_{\tilde{k}}$ is derived from the Phase II results (at time t^*) using the survival function relationship $\tilde{p}_{\tilde{k}} = \exp\left(-\left(\frac{t^*}{\gamma_{\tilde{k}}}\right)^{\alpha}\right)$, where $\tilde{p}_{\tilde{k}}$ denotes the estimated survival rate of the $\tilde{k}$-th arm (usually the posterior mean or median). Considering $\alpha = \tilde{\alpha}$, with $\tilde{\alpha}$ denoting the expected value of the prior distribution on α (usually set equal to 1),

$$\mu_{\tilde{k}} = \log\left(\frac{-t^*}{\log\left(\tilde{p}_{\tilde{k}}\right)^{1/\tilde{\alpha}}}\right).$$

Standard deviation parameters, $\sigma_{\tilde{k}}$, can be set corresponding to a desired effective sample size (ESS) (Morita et al. 2008). The ESS quantifies the amount of information given by the prior distribution in terms of the number of potential additional patients added to the analysis. Several ESS definitions are available in the literature; see, for example, Morita et al. Morita et al. (2008), Neuenschwander et al. Neuenschwander et al. (2020) or Wiesenfarth and Calderazzo Wiesenfarth & Calderazzo (2020); however, their application to time-to-event models is challenging. Therefore, we suggest that a new ESS idea be built on the unit information concept (Liang et al. 2008). The unit information, $\mathcal{I}_u$, in a frequentist setting, can be defined as the information contributed by a single subject, that is, the Fisher information matrix divided by the sample size. In the Bayesian setting, we can adopt a similar definition, that is, the inverse of the variance of the posterior distribution divided by the sample size. Then, to have the ESS equal to n_*, we simply have to multiply the obtained $\mathcal{I}_u$ by n_* and return to the standard deviation scale, that is, set $\sigma_{\tilde{k}} = (n_*\mathcal{I}_u)^{-2}$. The value of $\mathcal{I}_u$ is linked to the statistical model and, in our case, with a binary outcome at Phase II to be translated into a prior distribution for Weibull parameters, was obtained by studying the (expected) linear relationship between sample sizes and posterior variance. However, in the survival setting, the censoring rate (c_r) can also impact the value of $\mathcal{I}_u$, making $\sigma_{\tilde{k}} = (\mathcal{I}(n^*, c_r))^{-2}$ a function of both sample size and censoring rate. Details on the computation are given in Appendix A1.

In the second approach, called the likelihood approach, the joint prior distribution on α, β_0, β_1 is computed using the Weibull binary likelihood based on the

Phase II data, that is,

$$\pi(\alpha, \beta_0, \beta_1) \propto \prod_{\tilde{k}=0}^{1} S(t^*|\alpha, \exp(\beta_{\tilde{k}}))^{y_{\tilde{k}}} \left(1 - S(t^*|\alpha, \exp(\beta_{\tilde{k}}))\right)^{n_{\tilde{k}}} - y_{\tilde{k}}\pi(\alpha)\pi(\beta_{\tilde{k}}), \tag{2}$$

with $\pi(\alpha)$ and $\pi(\beta_{\tilde{k}})$ denoting noninformative priors (for example, in the same families as in the ESS approach), and $n_{\tilde{k}}$ and $y_{\tilde{k}} = \sum y^*_{i,\tilde{k}}$ the total number of patients analyzed at Phase II and the number of survivors at t^* in arm $\tilde{k}$, respectively.

3.5 Final claim

The final claim is made considering the parameter associated with the treatment effect, $\Delta = \beta_1 - \beta_0$. The treatment is considered superior to the control arm if the posterior probability of $\Delta > 0$ exceeds a prespecified τ_2 threshold, that is,

$$P(\Delta > 0) > \tau_2. \tag{3}$$

Alternatively, if a stronger treatment effect is desired or a noninferiority margin is accepted, we could have $P(\Delta > \zeta_u) > \tau_2$. Details on how to fix the value of ζ_u are given in Appendix A2.

4 Simulation Study

4.1 Simulation settings

We evaluated the operating characteristics of five designs via an extensive simulation study. A frequentist operationally seamless method (F-OP), that is, with two stages written in the same protocol, done sequentially but with only Phase III data used for the final claim (therefore, no type I error adjustment is planned), is considered the main reference method and is included for comparison. The F-OP method uses a similar decision process as the Bayesian ones. At the time of interim analysis, no test is planned, and treatments are selected based on observed survival rates using a threshold approach (Friede et al. 2020), that is, the arm with the highest survival rate is selected to proceed to the phase III stage so long as the estimated difference to control is at least ζ_T, with the trial stopped otherwise. For the final analysis, F-OP uses a one-sided Wald test from a Weibull regression to estimate the impact of treatment, with a significance level of 0.025. The analysis was performed using the flexsurvreg() function in the `flexsurv` R package.

The Bayesian operational counterpart, which uses a weakly informative prior at the second stage, with an ESS = 1 and $\rho = 10$, $\kappa = 9$, will be denoted B-

OPwinf. An alternative Bayesian method with an informative prior based on the ESS approach, B-OPinfESS, where $\sigma_{\tilde{k}} = \left(\mathcal{I}(n_{\tilde{k}}, \hat{c}_{r,\tilde{k}})\right)^{-2}$ is computed using all $n_{\tilde{k}}$ Phase II patients who are not censored before 12 months in arm $\tilde{k}$ and the empirical censoring rate, $\hat{c}_{r,\tilde{k}}$, of Phase III of the same arm, is also considered. As shown in Appendix 1, we used $\mathcal{I}(n_{\tilde{k}}, \hat{c}_{r,\tilde{k}}) = 0.932n_{\tilde{k}} - 10.105\hat{c}_{r,\tilde{k}} - 0.747n_{\tilde{k}}\hat{c}_{r,\tilde{k}}$ (truncated at zero when negative values are given) when $\hat{c}_{r,\tilde{k}} = 0$, making it a conservative method. The fourth method, denoted B-OPinLIK, shares the same ideas as B-OPinfESS, but the informative prior is built using the likelihood-approach. Finally, the fifth method, added for comparison only, is a Bayesian inferential design, B-INFER, where at the end of the trial true survival times from all patients in both phases are analyzed; therefore, we assume to have followed and stored time of death information for each patient from Phase II in the control and selected treatment arm and not only the dichotomised outcomes used at the Phase II selections stage. Weakly informative priors, as for B-OPwinf, are used in B-INFER. For all Bayesian designs, we used an efficacy threshold $\tau_2 = 0.975$. This threshold value is widely used when dealing with Bayesian efficacy proof. Even if it is not completely analogous to the frequentist 0.025 1-sided counterpart, it often has very similar properties, as shown by comparing Bayesian non-informative and frequentist operational analyses. If one desires a full equivalence between Bayesian and frequentist threshold, extensive simulations are needed to find a threshold value that gives the same properties as the frequentist significance level. Several futility thresholds, τ_1, are evaluated along with corresponding thresholds (obtained by simulations, not shown) for the F-OP.

Four main scenarios were selected and are presented in Figure 2 to evaluate the operating characteristics (simulated type I error and power), and 1000 trials per scenario were simulated. In the main simulation set, we simulated nine candidate therapies and the control group. Twenty patients per arm were simulated at the Phase II stage, that is $N_k = 20$, $\forall k$, and 150 per arm were simulated at Phase III, that is $M_{\tilde{k}} = 150$, $\forall \tilde{k}$. The sample size at the second stage was computed from a frequentist point of view to have 80% power and a maximum of 5% type one error rate when only Phase III was performed with two arms (treatment vs. control). Patient survival times were drawn from a Weibull distribution, and censoring time was drawn from a uniform distribution.

In the first scenario, no arm is better than the control arm; that is, for all treatment arms, the hazard ratio (HR) is higher than or equal to one (equality only for one arm). This scenario helped us to evaluate the simulated type I error. In scenario 2, only one arm is truly superior to the control group, with a HR of 0.7 corresponding to the initial assumption of the Atalante-1 study. All other groups have a HR > 1. In the third scenario, two treatment arms are truly superior to the control, with one having a HR of 0.7 and the other having a HR of 0.8. Regarding the fourth scenario, only one arm is truly superior to the control group, as in scenario 1, but the real HR of 0.75 is smaller than the

expected HR = 0.7. The last three scenarios allow us to evaluate the simulated power in several situations.

Simulations are carried out using R software, and the Bayesian model is written using the Stan programming language Stan Development Team (2022). MCMC chains are then computed using the `rstan` R package. We run four MCMC chains using 4000 samples with a warm-up of 2000.

4.2 Simulation results

Figure 3 shows the results of the five designs in scenario 1. When a futility threshold $\tau = 0.4$ or $\zeta_T = -0.025$ is used at the selection stage, 22.1% and 26.7% of the trials are stopped at the interim analysis when using a Bayesian (all Bayesian designs share the same Phase II outcome) or the frequentist design, respectively. The simulated one-side type I error rate ranges from 1.8% for B-OPwinf to 2.4% for B-OPinfESS when no futility analysis is performed (corresponding to $\tau = 0$ or $\zeta_T = -1$). Otherwise, the type I error results are lower, ranging from 1.4% to 2.2%.

Regarding the simulated power, Figure 4 shows the results in scenarios 2, 3 and 4. In the first column, the percentage of simulations in which the correct arm is selected is given for the Bayesian and frequentist methods for the same two futility thresholds as in scenario 1. Generally, futility stopping reduces the percentage of correct claim with this reduction being of 5 points on average. This is linked to the percentage of early stopped trials. In scenario 2, of 1000 studies with a futility threshold $\tau = 0.4$ or $\zeta_T = -0.025$, 102 studies are stopped for all the Bayesian methods vs. 129 for the frequentist one, which results in a slight loss of power in the frequentist analysis compared to the Bayesian ones. B-OPwinf shows similar performances to F-OP, as expected. B-OPinfESS has an equal or slightly lower percentage of correct claims than B-INFER, while B-OPinfLIK performances are between the other two operationally Bayesian designs.

We then investigated the situation when the total sample size, that is, that of Phase II plus that of Phase III for the selected arms, was equal to 150. Figure 5 shows the results at the Phase III stage: the results at the first steps are identical to Figures 3 and 4 since the same sample size and setting are adopted. For all methods and scenarios, we observe a reduction of positive claims between 5% and 10% due to the reduction of gathered data for the final analysis.

In the Supplementary material, we have also evaluated and reported the methods-performance when the sample size at Phase II is increased up to 40 patients, when a higher censoring rate (20%) is considered, or with other futility thresholds.

Table 1: Interim analysis results. $\hat{p}^F$ refers to the frequentist estimation of survival rate, Δ_k^F to the difference of frequentist point estimations $\hat{p}_k^F - \hat{p}_0^F$, $\tilde{p}$ to the Bayesian posterior mean of survival rate, Δ to the Bayesian mean posterior difference of survival rates with respect to the control arm, $P(\Delta_k > 0)$ to the posterior probability that Δ is higher than 0. CI and CrI denote the confidence interval and the credible interval, respectively.

Arm	$\sum y^*_{i,k}/n_k$	p_k^F 95%[CI]	$(p_k^F - p_0^F)$	$\tilde{p}_k$ 95%[CrI]	Δ_k	$P(\Delta_k > 0)$
Control	5/18	0.278 [0.107-0.536]		0.267 [0.096-0.481]		
Arm 1	1/17	0.059 [0.003-0.308]	-0.219	0.060 [0.002-0.207]	-0.204	0.029
Arm 2	1/20	0.050 [0.003-0.269]	-0.228	0.052 [0.002-0.177]	-0.212	0.019
Arm 3	3/20	0.150 [0.040-0.389]	-0.128	0.151 [0.035-0.333]	-0.113	0.173
Arm 4	3/18	0.167 [0.044-0.423]	-0.111	0.167 [0.039-0.362]	-0.096	0.212
Arm 5	6/18	0.333 [0.144-0.588]	0.056	0.334 [0.139-0.563]	0.071	0.676
Arm 6	11/17	0.647 [0.386-0.847]	0.369	0.645 [0.408-0.847]	0.385	0.990
Arm 7	0/17	0.000 [0.000-0.229]	-0.278	0.006 [0.000-0.055]	-0.258	0.001
Arm 8	2/17	0.118 [0.021-0.377]	-0.160	0.119 [0.016-0.305]	-0.145	0.115
Arm 9	1/16	0.062 [0.003-0.323]	-0.215	0.064 [0.002-0.221]	-0.201	0.034

5 Illustration

In this section, we present as an illustration a simulated study extracted from scenario 2. Therefore, the sixth treatment is simulated as the efficacious one with HR = 0.7, while all the others are simulated with HR> 1, as illustrated in Figure 2. Ten percent was chosen as the censoring rate for each arm, and twenty patients were randomly allocated to each of the ten arms (control and nine treatment arms) at the Phase II stage. Table 1 shows the number of survivors ($\sum y^*_{i,k}$) among the noncensored patients analyzed in Phase II (n_k); the frequentist proportion estimation along with its 95% confidence interval; the frequentist proportion difference; the Bayesian posterior mean of $\tilde{p}_k$ along with its 95% credible interval; the mean of the posterior difference of each arm vs. the control (Δ); and the probability that Δ is higher than zero.

According to the frequentist design, if a futility step is added at the end of Phase II, only Arm 5 and Arm 6 would be considered for the selection (all arms with $(p_k^F - p_0^F) < -0.025$ will be excluded), and then Arm 6 should be selected. The same arm is selected if no futility step is planned. Similar results are achieved by the Bayesian methods since Arm 5 and Arm 6 are considered for the selection ($P(\Delta_k > 0) > 0.4$) and Arm 6 is selected for Phase III.

For Phase III, 150 patients were randomly allocated to the control arm and to Arm 6, called the treatment arm. The survival times of the patients are shown in a Kaplan-Meier plot, Figure 6.

Figure 7 shows the informative and weakly informative (winf) distributions derived using the ESS approach (infESS) and the likelihood approach (infLIK) for the selected treatment arm. Table 2 summarizes the results for all designs:

in the second column, either the posterior mean of the treatment estimate β_1 for Bayesian models or the `flexsurv` estimation of the treatment regression factor for the frequentist model is presented; and in the third column, Δ refers to either the posterior difference with respect to the control for Bayesian models or the `flexsurv` estimated difference for the frequentist method. The results are relatively close for all the methods, but we observe narrower credible intervals for Bayesian methods using Phase II information. B-OPinfLIK, B-OPinfESS and B-INFER designs conclude on a positive treatment effect since the lower bound of their credible interval is higher than 0. The lower bounds of the B-OPwinf and F-OP designs are very close to but lower than 0; therefore, those designs do not provide informative conclusions on the treatment effect, illustrating the benefit of incorporating the phase II results in the final analysis via the prior distribution.

6 Discussion

In this work, we explored the use of the Bayesian framework in seamless designs and how information could be transferred in the case of different but related endpoints, taking an example from the Atalante-1 case study.

As inferential seamless clinical trial is not always feasible in practice, we proposed two ways to set informative prior distributions (that can also be derived from external data in non seamless trials). In the first one, the ESS approach, we evaluated how much the unit information should be in a Weibull survival regression. In our case, a linear relationship was found, even if, in principle, the ESS could vary according to the survival rate in other models. In the second approach, the likelihood approach, we added the information at the binomial scale and not directly at the survival scale. As shown in Figure 7, using the informative ESS approach for the prior usually leads to less dispersed, and therefore more informative, prior distributions. Moreover, since possible loss to

Table 2: Final Analysis. β_1 denotes either the posterior mean of the treatment estimate for Bayesian models or the flexsurv estimation of the treatment regression factor for the frequentist model, and Δ to either the posterior difference with respect to the control for Bayesian models or the flexsurv estimated difference for the frequentist method.

Design	β_1 [CrI or CI]	Δ [CrI or CI]
B-INFER	2.515 [2.358-2.672]	0.271 [0.057-0.491]
B-OPinfESS	2.528 [2.372-2.691]	0.277 [0.054-0.503]
B-OPinfLIK	2.501 [2.343-2.663]	0.246 [0.027-0.467]
B-OPwinf	2.454 [2.291-2.618]	0.200 [-0.028-0.428]
F-OP	2.446 [2.058-2.835]	0.194 [-0.032-0.194]

follow-up is only estimated using Phase III data, it can produce stronger information than the information brought by real Phase II patients, as in the 150 patients simulation setting. This is because a censored patient comes with less information than a patient who had the event in the survival likelihood. While it can be useful to increase the power, it could come with a possible type one error inflation with respect to the inferentially seamless designs. To be more conservative, a possible action consists in decreasing the number of patients n_* used to build the prior distribution. To note, B-INFER uses Phase II patients in Phase III analysis. Since this method was added to comparison purpose only, we did not consider any further correction in posterior distribution thresholds to ensure type I error control. However, it should be considered in real practice along with simulation studies.

The likelihood approach can be seen as an intermediate method between the ESS informative and the ESS weakly informative approaches, which was expected since the information brought by a binary variable is known to be lower than that of a continuous variable. Notably, while the binary-Weibull likelihood could be seen as a nonwell-posed problem, it produces a proper prior when coupled with noninformative proper distributions, as done in our application. This prior can be seen as a power prior Ibrahim et al. (2015) where the power parameter is set to 1 and the likelihood of historical trial, here the Phase II data, and the actual trial, the Phase III, differs. When full data are available for Phase II, that is, the actual survival times are known rather than just whether they are less than or greater than t^*, the power prior approach that shares the same likelihood across studies can be used. In this case, setting the power parameter equal to 1 leads to a Bayesian operationally seamless design that is exactly equal to a Bayesian inferentially seamless design. Even if the Phase II likelihood appears in different parts of the Bayes formula, that is, in the prior for the operationally seamless design and in the likelihood part for the inferentially seamless one, the two final formulas coincide. Therefore, B-INFER can also be seen as an operationally seamless design that uses a power prior approach with no discounting in historical data. As for the ESS approach, to be more conservative, a possible action consists in generalising the prior distribution using a power prior approach, that is $\pi(\alpha, \beta_0, \beta_1) \propto \prod_{\tilde{k}=0}^{1} \left[S(t^*|\alpha, \exp(\beta_{\tilde{k}}))^{y_{\tilde{k}}} \left(1 - S(t^*|\alpha, \exp(\beta_{\tilde{k}}))\right)^{n_{\tilde{k}} - y_{\tilde{k}}} \right]^{q_{\tilde{k}}} \pi(\alpha)\pi(\beta_{\tilde{k}})$, with $0 \leq q_{\tilde{k}} < 1$. Setting the $q_{\tilde{k}}$ parameter lower than 1 discounts the likelihood term.

Obviously, using a similar survival endpoint, even if not completely identical if the follow-ups differ, for the two steps of the analysis benefits the prior construction, making it easier to create informative prior distributions. In this case, the posterior distributions of Phase II can be directly used as priors for Phase III parameters. For the frequentist F-OP design to control the type I error the phase III part needs to include no patients with data used in the phase II part even as censored observations. The advantage of the methods we proposed is that they can also be used when Phase II is external to the trial (and therefore not a seamless master protocol), and data can be found in the literature. In

this case, if the trial conditions are not exactly the same between studies, we suggest decreasing the ESS of the prior distributions or checking for prior-data conflict Ollier et al. (2020), Wiesenfarth & Calderazzo (2020).

Increasing the sample size of the Phase II stage will increase, as expected, the probability of selecting the correct arm and, therefore, the power of the whole seamless study, as we can see in the Supplementary material. If one wants to include the censored patients of the first stage in the survival rate computation, multiple imputations methods could be considered. However, we did not consider them relevant at this early stage of the trial. Another option, would be to consider patients censored before t^* as dead, but this choice would be, in general, too conservative. In this case, a low and homogeneous censoring rate among treatment groups should not affect the results of interim analysis while a high censoring rate should increase the number of studies stopped for futility. The futility step, which can be done before or after the selection step, can benefit patients by terminating early studies that would have ended negatively. However, since the Phase II stage involves a small sample size in each arm, it could also stop a few studies that would have ended with confirmatory results at the final analysis. Therefore, a proper tradeoff can be evaluated via simulations at the protocol writing stage. In our example, we did not use a strong threshold; for example, the Bayesian threshold on posterior probabilities is lower than 0.5 and therefore lower than targeting at the median. For the sake of comparability and interpretability, the threshold for the frequentist analysis was fixed by previous simulations, trying to have similar selection results for the different methods. However, a perfect threshold was not found, resulting in slightly different results between Bayesian and frequentist methods.

Since Bayesian selection uses predictive probabilities via the MCMC approach, heavy tails impacting the posterior expected value could occur, leading to a different arm selection with respect to a simpler frequentist threshold approach.

A Phase II model was constructed to allow for possible modifications and extensions. We focused on a simple setting; however, several arms can be linked to each other if the same treatment but different doses or regimens are evaluated. In this case, if an efficacy-dose relation shape is expected, it can be added to the Phase II model. Moreover, if the arms do not represent different treatments but different populations or disease types, a correlation structure can be added as an additional model level.

In conclusion, the Bayesian framework provides a powerful tool to transfer information between trial phases, and it is particularly adapted to a seamless design. In our setting, where Phase II only aims at selecting the best treatment arm or stopping the trial, we showed how using Phase II data can increase the simulated power of the seamless design relative to an operationally seamless approach that uses only the phase III data in the final analysis, while still having acceptable simulated type I error, provided a binding futility threshold is used. In future work, the sample size needed to achieve a prespecified power while

controlling the type I error should be evaluated in this Bayesian setting.

Supplementary information

All the codes will be available at the first author's GitHub repository.

Acknowledgments

The work of the Benjamin Duputel was partially funded by a grant from the Association Nationale de la Recherche et de la Technologie, with eXYSTAT, Convention industrielle de formation par la recherche number 2019/1364. The authors would like to thank the anonymous reviewers, for their insightful comments and suggestions, Silvia Calderazzo, for the constructive exchange on the ESS topic, and Berangere Vasseur and OSE laboratories for the use and brainstorming on Atalante-1.

Declarations

The authors declare no conflicts of interest.

References

Albert, J. & Hu, M. (2020), 'Probability and Bayesian Modeling', CRC Press, Taylor & Francis Group .

Aouni, J., Bacro, J., Toulemonde, G., Colin, P., Darchy, L. & Sébastien, B. (2021), 'On the use of utility functions for optimizing phase II/phase III seamless trial designs', Journal of Clinical Trials **10**(415).

Aupiais, C., Alberti, C., Schmitz, T., Baud, O., Ursino, M. & Zohar, S. (2019), 'A Bayesian non-inferiority approach using experts margin elicitation–application to the monitoring of safety events', BMC Medical Research Methodology **19**(1), 1–13.

Besse, B. (2021), 'Activity of OSE-2101 in HLA-A2+ non-small cell lung cancer (NSCLC) patients after failure to immune checkpoint inhibitors (IO): Final results of phase III Atalante-1 randomised trial', ESMO 2021.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. & Racine-Poon, A. (2009), 'Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology', Statistics in Medicine **28**(10), 1445–1463.

Bretz, F., Schmidli, H., König, F., Racine, A. & Maurer, W. (2006), 'Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts', Biometrical Journal: Journal of Mathematical Methods in Biosciences **48**(4), 623–634.

Chang, S. & Berger, J. O. (2021), 'Comparison of Bayesian and Frequentist Multiplicity Correction for Testing Mutually Exclusive Hypotheses Under Data Dependence', Bayesian Analysis **16**(1), 111 – 128.

Chapple, A. G. & Thall, P. F. (2019), 'A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III', Biometrics **75**(2), 371–381.

Cui, L., Hung, H. J. & Wang, S.-J. (1999), 'Modification of sample size in group sequential clinical trials', Biometrics **55**(3), 853–857.

Fleming, T. R. (1982), 'One-sample multiple testing procedure for phase II clinical trials', Biometrics **38**(1), 143–151.

Friede, T., Parsons, N. & Stallard, N. (2012), 'A conditional error function approach for subgroup selection in adaptive clinical trials', Statistics in Medicine **31**(30), 4309–4320.

Friede, T., Stallard, N. & Parsons, N. (2020), 'Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R', Biometrical Journal **62**(5), 1264–1283.

Ibrahim, J. G., Chen, M.-H., Gwon, Y. & Chen, F. (2015), 'The power prior: theory and applications', Statistics in Medicine **34**(28), 3724–3749.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001), 'Bayesian survival analysis', Springer .

Jenkins, M., Stone, A. & Jennison, C. (2010), 'An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints', Pharmaceutical Statistics **10**(4), 347356.

Kimani, P. K., Glimm, E., Maurer, W., Hutton, J. L. & Stallard, N. (2012), 'Practical guidelines for adaptive seamless phase II/III clinical trials that use Bayesian methods', Statistics in Medicine **31**(19), 2068–2085.

Kimani, P. K., Stallard, N. & Hutton, J. L. (2009), 'Dose selection in seamless phase II/III clinical trials based on efficacy and safety', Statistics in Medicine **28**(6), 917–936.

Kunz, C. U., Jrgens, S., Bretz, F., Stallard, N., Van Lancker, K., Xi, D., Zohar, S., Gerlinger, C. & Friede, T. (2020), 'Clinical Trials Impacted by the COVID-19 Pandemic: Adaptive Designs to the Rescue?', Statistics in Biopharmaceutical Research **12**(4), 461.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008), 'Mixtures of g priors for Bayesian variable selection', Journal of the American Statistical Association **103**(481), 410–423.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P. & Krams, M. (2006), 'Adaptive seamless phase II/III designsbackground, operational aspects, and examples', Drug Information Journal **40**(4), 463–473.

Morita, S., Thall, P. F. & Müller, P. (2008), 'Determining the effective sample size of a parametric prior', Biometrics **64**(2), 595–602.

Neuenschwander, B., Weber, S., Schmidli, H. & O'Hagan, A. (2020), 'Predictively consistent prior effective sample sizes', Biometrics **76**(2), 578–587.

Ollier, A., Morita, S., Ursino, M. & Zohar, S. (2020), 'An adaptive power prior for sequential clinical trials–Application to bridging studies', Statistical Methods in Medical Research **29**(8), 2282–2294.

Park, J. J. H., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K. & Mills, E. J. (2019), 'Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols', Trials **20**(1), 572.

Quan, H., Luo, X., Zhou, T. & Zhao, P.-L. (2020), 'Seamless phase II/III/IIIb clinical trial designs with different endpoints for different phases', Communications in Statistics-Theory and Methods **49**(22), 5436–5454.

Renfro, L. A. & Sargent, D. J. (2017), 'Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples', Annals of Oncology **28**(1), 34–43.

Schmidli, H., Bretz, F. & Racine-Poon, A. (2007), 'Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint', Statistics in Medicine **26**(27), 4925–4938.

Stallard, N. (2010), 'A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information', Statistics in Medicine **29**(9), 959971.

Stallard, N. (2011), 'Group-sequential methods for adaptive seamless phase II/III clinical trials', Journal of Biopharmaceutical Statistics **21**(4), 787–801.

Stallard, N., Kunz, C. U., Todd, S., Parsons, N. & Friede, T. (2015), '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.

Stan Development Team (2022), 'Stan modeling language users guide and reference manual, version 2.29.0'.
URL: *http://mc-stan.org/*

Sun, L. Z., Li, W., Chen, C. & Zhao, J. (2020), 'Advanced utilization of intermediate endpoints for making optimized cost-effective decisions in seamless phase II/III oncology trials', Statistics in Biopharmaceutical Research **12**(2), 224–233.

Wiesenfarth, M. & Calderazzo, S. (2020), 'Quantification of prior impact in terms of effective current sample size', Biometrics **76**(1), 326–336.

Woodcock, J. & LaVange, L. M. (2017), 'Master protocols to study multiple therapies, multiple diseases, or both', New England Journal of Medicine **377**(1), 62–70.

Appendix

A.1 Information computation

To compute $\mathcal{I}(n_{\tilde{k}}, \hat{c}_{r,\tilde{k}})$, we studied the relationship between the expected inverse of the variance of the β posterior distribution, the sample size and the censoring rate. We first set the mortality rate estimated from Phase II, $\tilde{p}$, and then compute the corresponding true β, when α is set equal to 1, using the following relationship:

$$\tilde{p} = \exp\left(-\left(\frac{t^*}{\exp(\beta)}\right)^{\alpha}\right).$$

The resulting values are then used as true scenario parameters. We simulated time-to-event (with and without censoring) datasets drawing from the Weibull distribution associated with the true scenario parameters, with increasing sample size, from 5 to 100 patients. For each sample size, 1000 replications were simulated, and for each replication, a Bayesian Weibull analysis, as described in Section 3.4, was performed using noninformative prior distributions, that is, $\pi(\alpha) = \text{InverseGamma}(\rho, \kappa)$, and $\pi(\beta) = \mathcal{N}\left(\log\left(\frac{-t^*}{\log(\tilde{p})^{1/\tilde{\alpha}}}\right), 10\right)$. The expected inverse of the variance of β was approximated by the sampling mean over the 1000 replications. We studied three values for $\tilde{p}$, $\tilde{p} \in \{0.3, 0.5, 0.8\}$, and since all results were similar, we continued with only $\tilde{p} = 0.5$, and we performed linear regressions without intercepts. In our exploratory analysis, two regression lines, depending on the sample size and censoring rate, were found, one when no patient was censored and the other for all other cases, as shown in Figure 8. Since $\mathcal{I}(n_{\tilde{k}}, \hat{c}_{r,\tilde{k}})$ cannot be negative, we truncated the equation to zero.

For other time-to-event parametrizations, $\mathcal{I}_u$ could depend on the scenario parameters, such as $\tilde{p}$. In this case, a linear regression should be performed for each

possible $\tilde{p}$. Even if it could be quite challenging in a simulation study, in real practice, only the estimated Phase II $\tilde{p}$ would be analyzed, reducing the computational effort. Using another parametric distribution for the phase III would, in general, require modifications of some specific formulas in the manuscript. It is straightforward to adapt the OPinfLIK method to the new parametric distribution, since it would only change the survival distribution ($S(t|\nu_k)$, where ν_k refers to the model parameters of the kth arm) that is used in the likelihood at the prior level. Regarding the OPinfESS method, the process to obtain prior distribution is the same but formula modifications needed will depend on the number of model parameters ν_k of the survival function. For example, using an exponential distribution, with only one parameter, we could directly solve equations shown in appendix A.1, without having to fix any other parameter. If the parametric model presents more than one parameter, as in the Weibull model, one or more parameters need to be fixed to create informative prior distribution on the parameter of interest (usually a location parameter). The choice of the best parametric model to use could be done using some model selection criteria, or multiple models can be introduced into a Bayesian Model Averaging approach.

A.2 Computation ofζ_u

The ζ_u value is linked to the minimal improvement we wish to observe in terms of survival in the treatment group. The median survival times of the control group m_0 and the treatment group m_1 are given by $m_0 = \log(2)^{1/\alpha}\exp(\beta_0)$ and $m_1 = \log(2)^{1/\alpha}\exp(\beta_1)$, respectively.

We can find the minimum ζ_u that improves by q times the median survival of m_0 by solving the following steps:

$$m_1 = qm_0$$

$$\log(2)^{1/\alpha}\exp(\beta_1) = q(\log(2)^{1/\alpha}\exp(\beta_0))$$

$$\exp(\beta_1) = q\exp(\beta_0)$$

$$\beta_1 = log(q) + \beta_0$$

$$\beta_1 - \beta_0 = log(q).$$

Therefore, $\zeta_u = log(q)$.

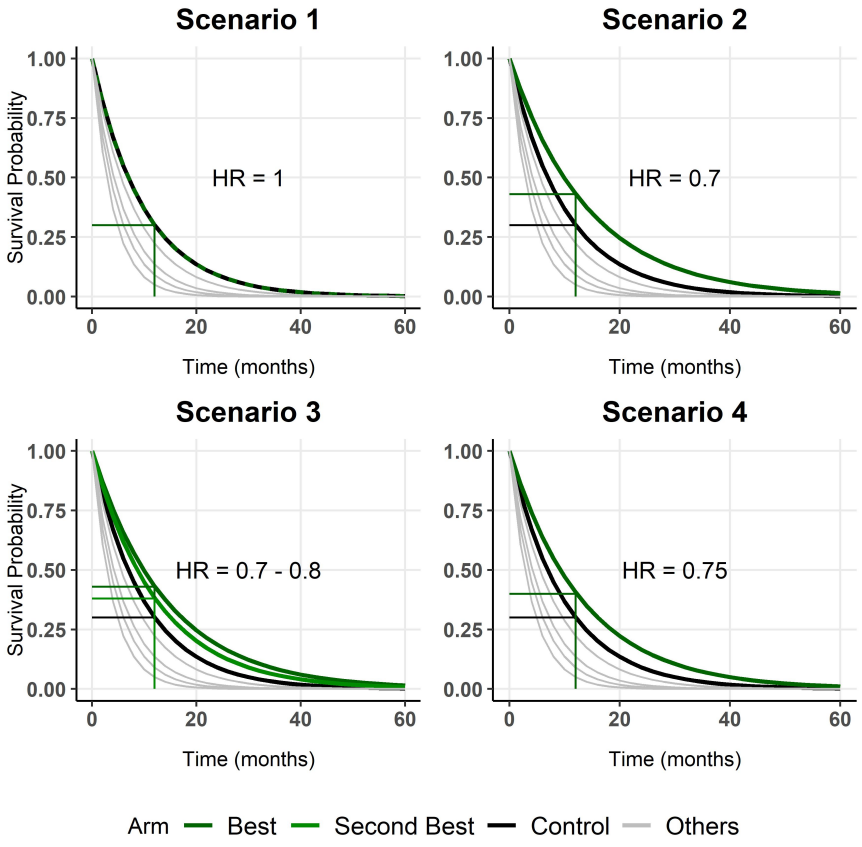

Figure 2: Survival probability function of all arms in each scenario. Only the control arm and the best arm (and the second-best arm for scenario 3) are highlighted. The corresponding survival probability at 12 months is indicated by horizontal lines.

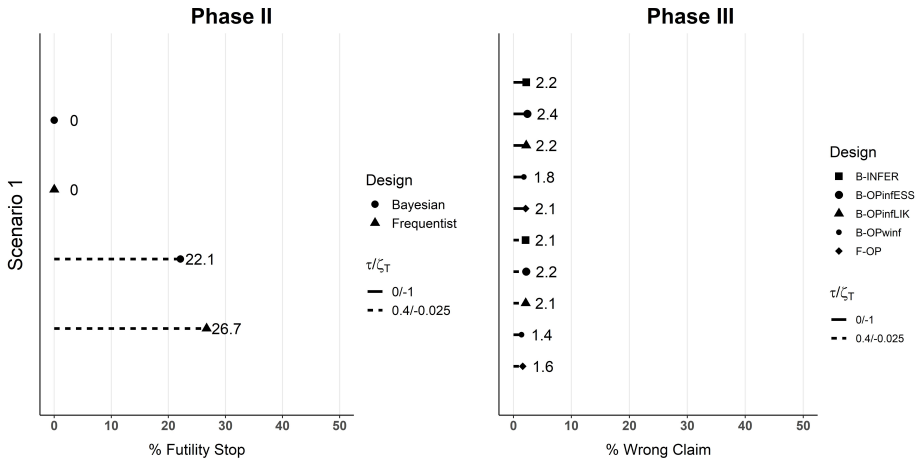

Figure 3: Results in terms of simulated type one error. On the left-hand side, the percentage of trials that were stopped at the futility analysis at the end of Phase II in scenario 1 is shown. On the right-hand side, the percentage of trials where the treatment arm was erroneously claimed to be superior is shown. Straight lines refer to no futility rule applied, that is, $\tau = 0$ and $\zeta_T = -1$, while dashed lines refer to $\tau = 0.4$ and $\zeta_T = -0.025$.

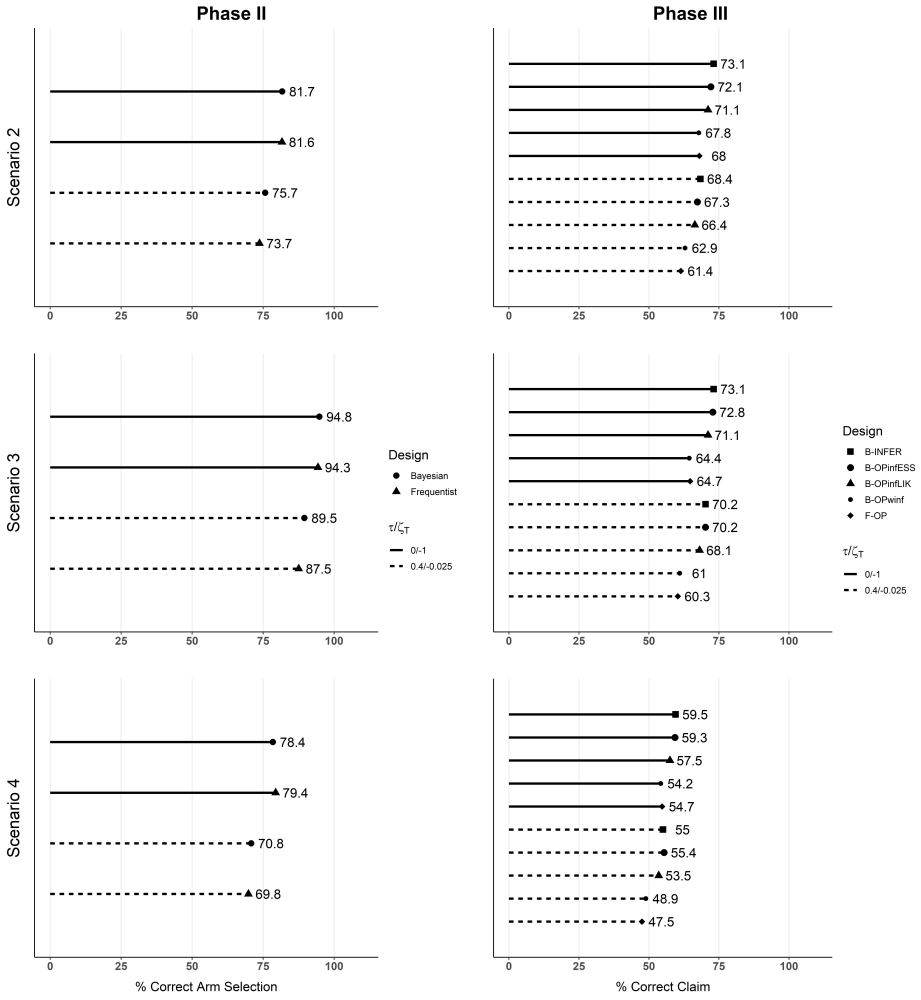

Figure 4: Results in terms of simulated power. Each plot line represents one scenario. In the first column, the percentages of correct arm selection at the end of Phase II are shown for the Bayesian methods and the frequentist one. In the second column, the percentages of final Phase III correct claims associated with each design are given. Straight lines refer to no futility rule applied, that is, $\tau = 0$ and $\zeta_T = -1$, while dashed lines refer to $\tau = 0.4$ and $\zeta_T = -0.03$.

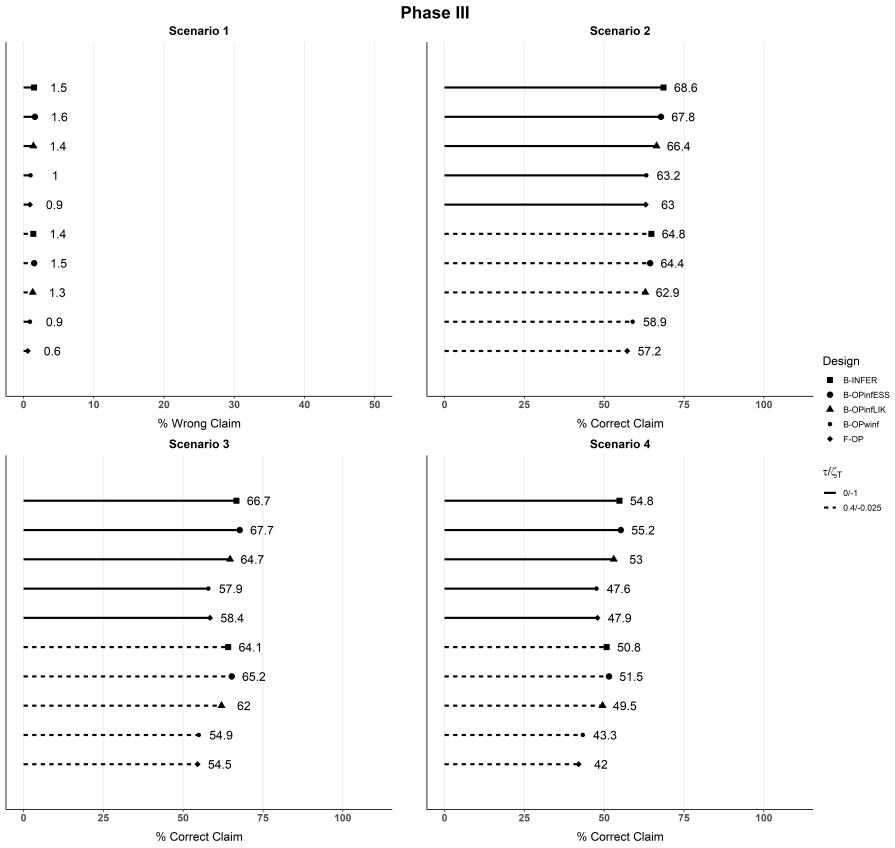

Figure 5: Results in terms of simulated power and type I error for a total sample size of 150 patients with 20 used for Phase II. Each plot represents a scenario, and the percentages of final Phase III correct and incorrect claims associated with each design are shown. Straight lines refer to no futility rule applied, that is, $\tau = 0$ and $\zeta_T = -1$, while dashed lines refer to $\tau = 0.4$ and $\zeta_T = -0.03$.

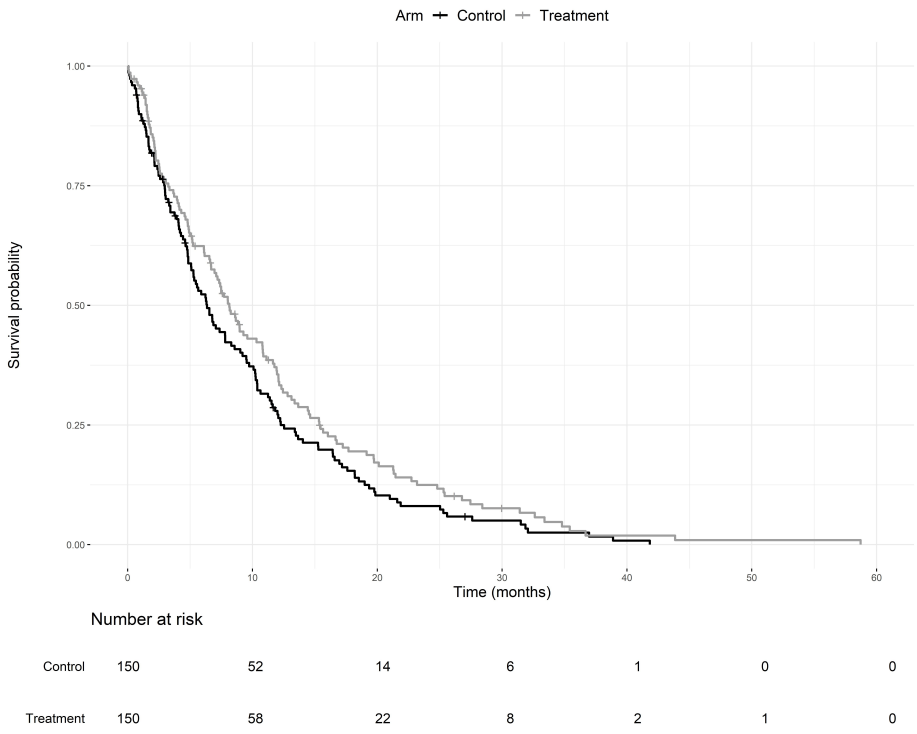

Figure 6: Kaplan-Meier survival curves of the control and treatment arms for the illustrative example.

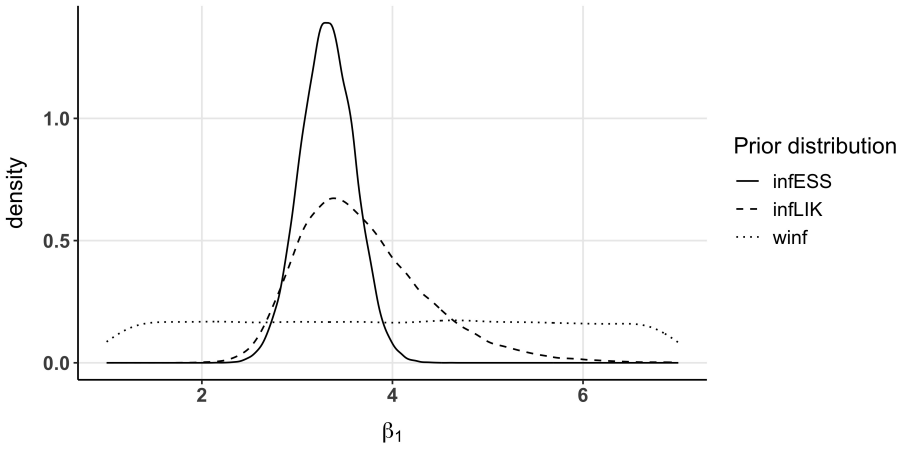

Figure 7: Marginal prior distributions obtained using Phase II data for the treatment arm in the illustrative example. The infESS was computed using $n_* = 17$ and a censoring rate of 14%.

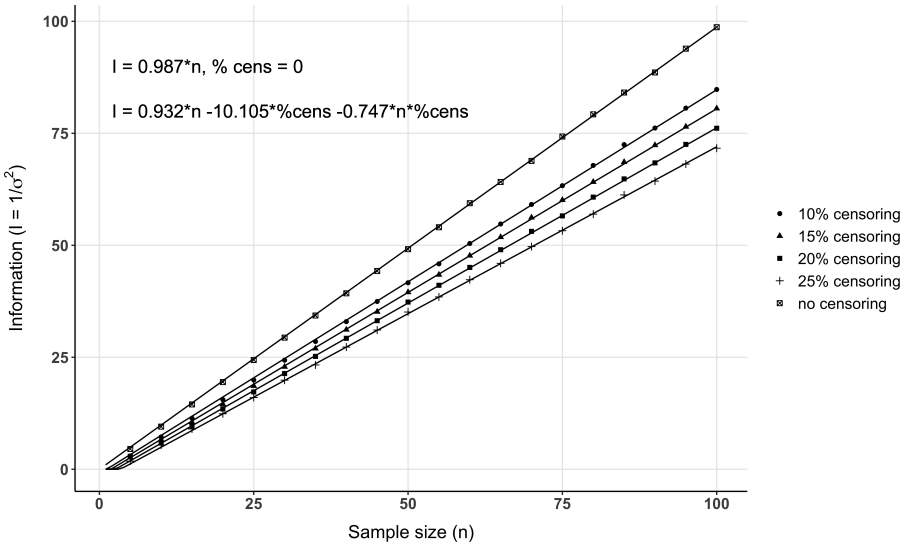

Figure 8: Information vs. sample size and censoring rate and the estimated regression lines.