SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(4) 1272–1289
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280217747310
journals.sagepub.com/home/smm
SAGE

Bayesian Phase II optimization for time-to-event data based on historical information

Anja Bertsche,[1,2] Frank Fleischer,[1] Jan Beyersmann[2] and Gerhard Nehmiz[1]

Abstract

After exploratory drug development, companies face the decision whether to initiate confirmatory trials based on limited efficacy information. This proof-of-concept decision is typically performed after a Phase II trial studying a novel treatment versus either placebo or an active comparator. The article aims to optimize the design of such a proof-of-concept trial with respect to decision making. We incorporate historical information and develop pre-specified decision criteria accounting for the uncertainty of the observed treatment effect. We optimize these criteria based on sensitivity and specificity, given the historical information. Specifically, time-to-event data are considered in a randomized 2-arm trial with additional prior information on the control treatment. The proof-of-concept criterion uses treatment effect size, rather than significance. Criteria are defined on the posterior distribution of the hazard ratio given the Phase II data and the historical control information. Event times are exponentially modeled within groups, allowing for group-specific conjugate prior-to-posterior calculation. While a non-informative prior is placed on the investigational treatment, the control prior is constructed via the meta-analytic-predictive approach. The design parameters including sample size and allocation ratio are then optimized, maximizing the probability of taking the right decision. The approach is illustrated with an example in lung cancer.

Keywords

Proof-of-concept, Go–NoGo decision, Bayes, time-to-event, operating characteristics, meta-analytic-predictive prior distribution

1 Introduction

One of the key steps in the drug development process is the decision for a Go or NoGo after the exploratory phase. Go in this context means that one or more subsequent confirmatory trials are initiated. NoGo usually relates to the stop of the development for this compound at least in the investigated indication or even in total. From a company's perspective, there are two risks or possibilities for error involved here. On the one hand, it should be avoided to stop the development of an effective drug before reaching the confirmatory phase. On the other hand, the development of an insufficiently efficacious drug should be stopped as soon as possible in order to avoid subsequent costs and to free resources for more promising compounds. In recent years, companies are often in the comfortable position of having many targets and potential compounds under investigation.[1] Therefore the latter case has become more and more important. For reaching such a Go–NoGo decision or for prospectively defining a Go–NoGo criterion, almost exclusively statistical significance testing has been used in the past. It has become obvious that such an approach is suffering from some major shortcomings. The most prominent ones are that a statistical significant treatment effect alone is not sufficient for decision making as well as the inability to smoothly and systematically incorporate historical information. Quantitative criteria that relate to a clinically relevant observed effect size are needed as a small observed improvement may not be considered as worthwhile, even if

[1]Biost. and Data Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach/Riss, Germany
[2]Institute of Statistics, Ulm University, Ulm, Germany

Corresponding author:
Anja Bertsche, Biost. and Data Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, BDS D123-00-14, Birkendorfer Str. 65, 88397 Biberach/Riss, Germany.
Email: anja.bertsche@boehringer-ingelheim.com

proven to be statistically significant. On the other hand, possibly promising large effects could have been dismissed. Whereas this problem might be solved by considering point estimates and associated confidence intervals,[2] the second shortcoming is usually difficult to overcome and has led to the development of Bayesian approaches. The systematic and pre-defined inclusion of historical information is continuously gaining in importance as historical data are becoming more and more available, e.g. based on earlier in-house trials but also on publicly available resources like clinicaltrials.gov or transparency initiatives by major pharmaceutical companies.[3] Moreover, a systematic process for including additional historical information is useful for enhancing the decision process in at least two aspects. The bigger amount of overall information should increase the precision of the estimate and hence decrease error rates in both directions. Additionally, the decision making process should become more objective by quantitatively incorporating the historical knowledge. Note that these arguments may in particular hold in the early phase setting as the newly generated information might be relatively small compared to the already available one.

In our article, we aim to optimize the design of a proof-of-concept (PoC) trial with respect to decision making for a randomized parallel comparison between the experimental treatment and either placebo or active control. Specifically, we consider a time-to-event setting and investigate exponentially distributed data. We consider the situation where prior information is only available on the control treatment and will be included as a meta-analytic-predictive (MAP) prior.[4,5] The control treatment group in the new Phase II trial is assumed exchangeable with the previous control treatment groups. The posterior distribution of the effect size, expressed as hazard ratio (HR), is then derived. We define and optimize a quantitative Go criterion based on the effect size. It is assumed that the company's decision is made after completion of the Phase II PoC trial, not at an interim time-point. The overall approach discussed in this article is illustrated in Figure 1.

The idea to include historical information into the planning and evaluation of a new trial is actually not new. Already in 1976, Pocock[6] describes the combination of randomized and historical control groups. Lecoutre et al.[7] consider the posterior distribution of Weibull parameters and the resulting prediction, but do not optimize further

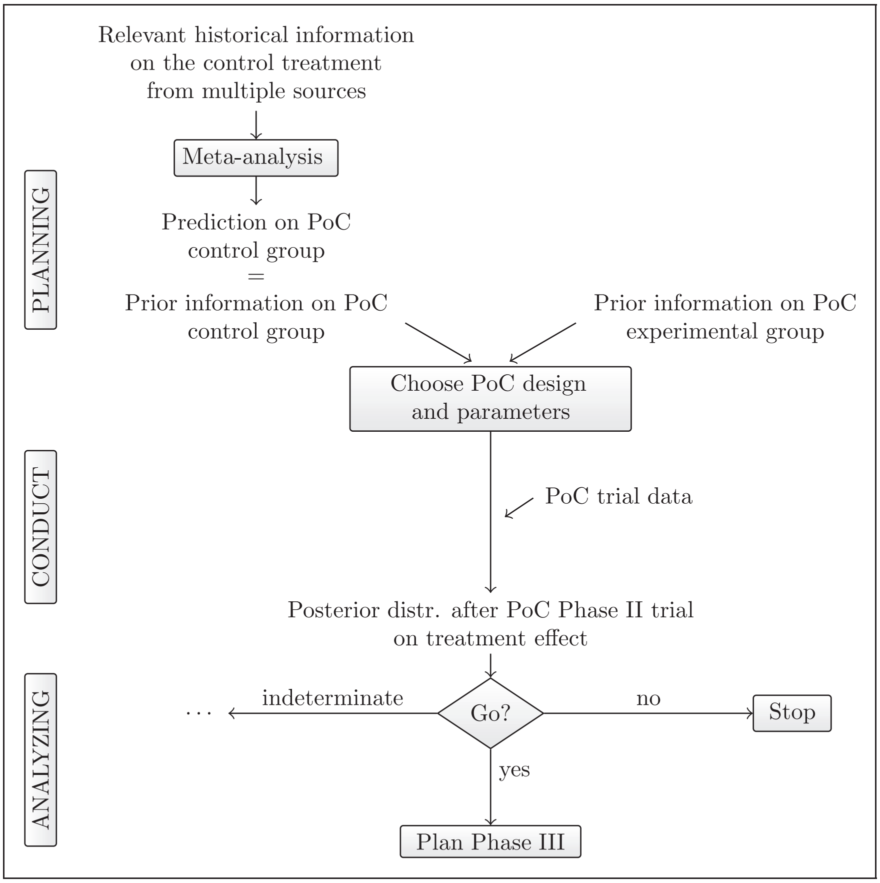

PLANNING
Relevant historical information on the control treatment from multiple sources
Meta-analysis
Prediction on PoC control group
=
Prior information on PoC control group
Prior information on PoC experimental group
Choose PoC design and parameters
CONDUCT
PoC trial data
Posterior distr. after PoC Phase II trial on treatment effect
ANALYZING
...
indeterminate
Go?
no
Stop
yes
Plan Phase III

Figure 1. Methodological approach.

trials on that basis. Thall[8] describes problems arising from collecting time-to-event data at interim time-points in general. Moreover, they investigate robustness against deviations from the exponential model with a conjugated Gamma prior for the hazard λ. The article of Neuenschwander et al.[4] is a central reference. The authors introduce the MAP distribution for normally distributed data and calculate the effective sample size thereof. Moreover, they provide an extensive list of the older literature. Viele et al.[9] also consider normal or binary data in a PoC trial and describe several methods to set up the prior distribution for the control treatment.

Several authors have discussed triage approaches for defining Go–NoGo criteria for PoC between Phase II and III. In the article of Grieve,[10] interesting thoughts regarding the interplay of Bayesian and frequentist approaches for decision making can be found. Frewer et al.[11] discuss a frequentist concept for decision making that includes three possible decisions: Go, NoGo and indeterminate, where for the indeterminate zone the next steps need to be defined. The most common approach for a systematic inclusion of historical information into a new trial is however via the usage of a prior distribution. Fisch et al.[12] describe Bayesian concepts in Go–NoGo decision making based on significance as well as relevance for normally distributed and binary data. Walley et al.[13] optimize the PoC decision, with inclusion of prior information, for normally distributed data. Musser et al.[14] utilize after Phase II the information from two biomarkers, in addition to normally distributed endpoint data, to select a Phase III dose through a pharmacokinetic/pharmacodynamic (PK/PD) model. However, they find only a small benefit even in favorable scenarios. Nikolakopoulou et al.[15] consider a network-meta-analytic prior and subsequent decisions about performing further trials; they are however restricted to asymptotic normal priors and likelihoods, and they do not make reference to the MAP prior distribution. Schmidli et al.[5] do actually combine the MAP approach with network-meta-analysis. Kirchner et al.[16] optimize the allocation of sample size between Phase II and III on the basis of assurance of the Phase III success, so that the difference between expected gain and deterministic cost is maximized. Götte et al.[17] consider possibly censored time-to-event data with constant HR between treatments and optimize the assurance for Phase III success in dependence of the size of Phase II; they do however only use stylized prior distributions. Cotterill and Whitehead[18] follow a similar approach for arbitrary or Weibull distributed time-to-event data within groups.

To achieve the inclusion of more recent data in the decision making, interim analyses and adaptations in Phase II trials have been considered by several authors. Cellamare and Sambucini[19] investigate a 2-stage trial with binary observations, and optimize the decision to perform stage 2. Neuenschwander et al.[20] include not only prior information but also interim "co-data" in the final analysis of a trial. Moreover, they discuss the case of relevant time trends (non-exchangeability between older and newer data). Gsponer et al.[21] include prior information in the monitoring of a group-sequential trial with normal approximations for the prior and the interim effect likelihoods. Yuan et al.[22] compare an adaptively selected set of treatments with a permanent control group and investigate the operating characteristics for the correct selection. Götte et al.[23] optimize interim decisions about population enrichment, with inclusion of prior information. Ohwada and Morita[24] include biomarker information in the development of decision rules on whether or not to confine an investigation to a subgroup.

This article shows a systematic approach for optimizing the design and the decision criteria of a PoC trial based on given maximum error rates and historical information for the control group. To the best of our knowledge, decision criteria for exponential event data based on the control group MAP, read from the effect size between trials, has not been investigated up to now.

Our article is structured as follows. In Section 2, the approach regarding the between-trial decisions as well as the underlying Bayesian modeling for the time-to-event data is introduced. Section 3 explains the MAP approach taken. Aspects of different censoring types and parametric density estimate representation of the prior are also discussed here. Section 4 deals with design optimization with respect to the PoC criterion chosen and the operating characteristic (OC) leading to the optimized result. In Section 5, a real trial data example is considered that is based on data from a compound in second line non-small-cell lung cancer. For this example, it is shown how to derive the MAP prior as well as how to optimize the design of the PoC trial. A reality check based on the actual data is performed. The article finishes with a discussion. Some specific calculations and program code are provided in the Supplemental Material. For this work, R version 3.2.2 and WinBUGS 1.4.3 are used.

2 Decision after Phase II

In this section, we introduce decision criteria based on the treatment effect size on the HR scale. These criteria are Bayesian in so far as the HR is treated as an unknown random parameter $\theta \in (0, \infty)$. It is defined as the ratio of hazard rates of experimental and control treatment. Thus, the smaller the observed HR, the more beneficial is the experimental treatment. For decision making, the parameter space of the HR is partitioned into two or three regions: a success ($\theta < \theta_L$), a futile ($\theta > \theta_U$), and an optional indeterminate region ($\theta_L \leq \theta \leq \theta_U$), where θ_U and θ_L are pre-specified effect thresholds ($\theta_L \leq \theta_U$). From a Bayesian perspective, decisions can be based on the posterior probability of the HR falling into one of these regions, denoted by $\mathbb{P}$. A Go decision will be taken if

$$\mathbb{P}(\theta < \theta_L | \text{data and prior information}) > \gamma_s \tag{1}$$

and accordingly a NoGo decision if

$$\mathbb{P}(\theta > \theta_U | \text{data and prior information}) > 1 - \gamma_f \tag{2}$$

where γ_s and γ_f are pre-specified probability thresholds. If none of the criteria is fulfilled, we get an indeterminate result and no decision can be taken so far. As a rule for success, γ_s will be large, whereas γ_f will be much smaller. For a meaningful relevance criterion, θ_L is typically chosen smaller than 1 (no treatment effect) whereas θ_U might be set to 1 in the sense of a classical significance criterion. An important task in planning the Phase II trial is to find optimal probability and effect thresholds in the sense that the risk of a false decision and the required resources are well-balanced. This is outlined in Section 4. Moreover, we would like to stress that these posterior probabilities contain all information on θ up to now. This includes not only the data emerging from the current PoC trial, but also the prior distribution. That is why Bayesian decision criteria become particularly attractive when historical knowledge is available.

To evaluate the decision criteria, we will model the treatments separately and combine the treatment-specific posterior distributions in terms of a ratio distribution. It is supposed that the event times of the patients, e.g. death or progression-free survival (PFS), are exponentially distributed for both treatments. Particularly, $T_{ij} \sim Exp(\lambda_j)$ for each patient i in group j, $i = 1, \ldots, n_j$, $j = E, C, \lambda_j > 0$ with survival function $S(t|\lambda_j) = \exp(-\lambda_j t)$ and constant hazard rate λ_j. It holds that the median survival is given by $\log(2)/\lambda_j$. A sample of observed time-to-event data from n_j patients consists at least of the minimum of the event or right censoring time $t_{ij} > 0$ and corresponding event indicator

$$d_{ij} = \begin{cases} 0 & \text{if } t_{ij} \text{ is the censoring time} \\ 1 & \text{if } t_{ij} \text{ is the event time} \end{cases} \tag{3}$$

In pharmaceutical studies, it is a common practice to wait until a pre-specified number of events have been observed (type II censoring) rather than pre-fixing the termination date (type I censoring). Although the trial duration is not known in advance, the information content of the trial is known and quantified by the number of observed events across both groups.

Lawless[25] shows (see p. 125f) that the form of the group-specific likelihood function is given as follows, although it is strictly speaking a partial likelihood in case of type II censoring

$$\mathscr{L}(\lambda_j) \propto \lambda_j^{d_j} \exp(-\lambda_j \mathrm{T}_{+j}) \tag{4}$$

where $d_j = \sum_{i=1}^{n_j} d_{ij}$ is the total number of events in group j and $\mathrm{T}_{+j} = \sum_{i=1}^{n_j} t_{ij}$ the total observation time including also the censoring times. The observation times are subject to independent uniformly distributed recruitment. The likelihood (4) also applies in this case, the reason being that the individual counting processes of an observed event still have the desired intensity, i.e., an individual at-risk indicator times λ_j. The individual at-risk status accounts for both type II censoring and staggered study entry, while the momentary event risk is still λ_j for an individual at risk.

In a Bayesian sense, we place prior distributions on each λ_j. The gamma distribution defined with two hyper-parameters, a scale parameter $\alpha_j > 0$ and a rate parameter $\beta_j > 0$, constitutes a conjugate prior. The density is given by

$$\pi(\lambda_j) \propto \lambda_j^{\alpha_j - 1} \exp(-\lambda_j \beta_j) \tag{5}$$

Taking $\lambda_j \sim Gam(\alpha_j, \beta_j)$ a priori, the resulting posterior distribution is

$$\pi(\lambda_j | d_j, \mathrm{T}_{+j}, \alpha_j, \beta_j) \propto \pi(\lambda_j) \mathscr{L}(\lambda_j) = \lambda_j^{\alpha_j + d_j - 1} \exp(-\lambda_j(\beta_j + \mathrm{T}_{+j})) \tag{6}$$

which is again the kernel of a gamma distribution, in fact $Gam(\alpha_j + d_j, \beta_j + \mathrm{T}_{+j})$. Therefore, the posterior parameters are obtained by adding the number of new events to the prior shape parameter and the total time on study to the prior rate parameter. This shows that α_j has the dimension of a number of previous events and β_j has the dimension of a previous total time on study for group j.

It can be shown that the decision criteria (1) and (2) that involve the ratio distribution of two independent gamma distributions can be evaluated analytically. Therefore, note that the group-specific posterior distribution (6) can be expressed equivalently as

$$(\beta_j + \mathrm{T}_{+j})\lambda_j | d_j, \mathrm{T}_{+j}, \alpha_j, \beta_j \sim Gam\big(\alpha_j + d_j, 1\big) \tag{7}$$

Comparing equation (7) for $j = E$ and $j = C$ and assuming λ_E and λ_C to be independent a priori, the posterior random variable

$$X := \frac{(\beta_C + \mathrm{T}_{+C})\lambda_C}{(\beta_C + \mathrm{T}_{+C})\lambda_C + (\beta_E + \mathrm{T}_{+E})\lambda_E} \tag{8}$$

follows a beta distribution with parameters $\alpha_C + d_C$ and $\alpha_E + d_E$. Moreover, the HR can be rewritten as

$$\theta = \frac{(\beta_C + \mathrm{T}_{+C})(1 - X)}{(\beta_E + \mathrm{T}_{+E})X} \tag{9}$$

It follows that for example the Go criterion, i.e. the posterior probability that the HR is smaller than some relevant threshold, is given by

$$\begin{aligned} \mathbb{P}(\theta < \theta_L | \text{data and prior information}) &= \mathbb{P}\left(\frac{1 - X}{X} < \frac{\theta_L(\beta_E + \mathrm{T}_{+E})}{\beta_C + \mathrm{T}_{+C}}\right) \\ &= \mathbb{P}\left(X > \frac{\beta_C + \mathrm{T}_{+C}}{\theta_L(\beta_E + \mathrm{T}_{+E}) + \beta_C + \mathrm{T}_{+C}}\right) \end{aligned} \tag{10}$$

which is the survival function of a beta distribution.

3 Meta-analytic-predictive approach

In the situation where group-specific prior information is available from preceding trials, we utilize the MAP approach[4,5] to derive an informative prior distribution on the corresponding rate parameters λ_j, $j = E, C$. In the case where only little prior knowledge is available, flat priors will be used, e.g. for the experimental group $\lambda_E \sim Gam(0.01, 0.01)$. In contrast, at least some information will be available for C. This is typically the case when the control treatment is an already established treatment. The MAP approach combines prior knowledge from multiple sources in a prior distribution. The fundamental idea is to perform a meta-analysis of the historical data taking the heterogeneity of the historical sources into account and predicting the future observations. In more detail, the predictive posterior distribution is utilized as the MAP prior distribution for the matching future treatment group.

3.1 Hierarchical model

Suppose that there are H historical data sets similar to our control group C that we want to incorporate in the Go–NoGo decision after the Phase II PoC trial. If there is also historical data available on the experimental group, one can proceed in the same manner, but we do not follow this up here. For the notation of the control groups, we use the index h for the historical control groups and keep C for the future Phase II control group. So each historical data set consists of n_h event or censoring times t_{ih} with $n_h - d_h$ censored observations, $i = 1, \ldots, n_h,\ h = 1, \ldots, H$. Furthermore, it is assumed that the event times are realizations of exponentially distributed independent random variables, i.e. $T_{ih} \sim Exp(\lambda_h), \quad \forall i$ in group h. When the underlying censoring scheme is type II censoring, the exact property (see Lawless,[25] p. 153)

$$\mathrm{T}_{+h} \sim Gam(d_h, \lambda_h) \tag{11}$$

can be utilized to specify the assumption on the sampling distribution in a hierarchical model. We assume that this property holds also for the future control group in an exchangeable manner, i.e. $\mathrm{T}_{+C} \sim Gam(d_C, \lambda_C)$. While all the rate parameters λ_h and λ_C are modeled as random, the number of events is pre-fixed at the planning stage. The relationship of the historical parameters and the future rate parameter is expressed by exchangeable and independent parameters (random effects) under the same normal distribution

$$\log(\lambda_1), \ldots, \log(\lambda_H), \log(\lambda_C) \sim \mathcal{N}(\mu, \tau^2) \tag{12}$$

The MAP prior is then the posterior predictive distribution

$$\pi(\lambda_C | T_{+1} \ldots T_{+H}, \mu, \tau) =: \pi(\lambda_C) \tag{13}$$

Using a full Bayesian approach, the distribution of the hyper-parameter (μ, τ) needs to be specified. Following the recommendations of Schmidli et al.,[5] these are modeled as independent from each other a priori. A non-informative normal prior is placed on μ with mean $\mu_0 = 0$ and large variance as the data should be sufficiently informative, e.g. $\sigma_0^2 = 1000$. A half-normal prior is placed on the standard deviation τ. Schmidli et al.[5] describe sensitivity analyses of the developed MAP prior for varying hyper-prior distributions, especially for the between trial standard deviation τ when only few historical sources are available. However, we found in our example (see Section 5.1 below) no relevant influence on the MAP prior (results not shown).

3.2 Remark on censoring mechanism

If the underlying censoring mechanism is not type II, classical large sample properties will be utilized instead of the finite property (11). For example, the asymptotic distribution of the maximum likelihood (ML) estimator for λ_h holds favorably for type I and type II censoring. Andersen et al.[26] (see chapter 6) discuss more general independent censoring schemes, defined in terms of preserving the structure of counting process intensities, under which these weak convergence results hold. Assuming exponentially distributed event times, the ML estimator for λ_j is given by

$$\hat{\lambda}_j = \frac{d_j}{\mathrm{T}_{+j}} \tag{14}$$

It holds that $\hat{\lambda}_j$ is approximately normal distributed with mean λ_j and variance $I^{-1}(\lambda_j) = \lambda_j^2 / d_j$, where $I^{-1}(\lambda_j)$ denotes the inverse of Fisher's information measure.

The asymptotic approximation is improved with a log-transformation to be already suitable for samples with $d_h = 10$ (see Aalen et al.,[27] p. 215). Thus, instead of equation (11), the sampling distribution of the hierarchical model can be alternatively formulated with

$$\log\left(\frac{d_h}{T_{+h}}\right) \sim \mathcal{N}\left(\log(\lambda_h), \frac{1}{d_h}\right), \quad h = 1, \ldots, H \tag{15}$$

The random effects assumption on the parameters (12) stays the same. In simulation studies with moderate to large sample sizes we experienced quite similar results, which is why the latter is not followed up further.

3.3 MCMC representation and parametric density estimate of the MAP prior

Based on the hierarchical model (11), (12), the prediction (13), and the hyper-priors, a sample $\log(\lambda_C^{(1)}), \ldots, \log(\lambda_C^{(M)})$ can be generated with Markov Chain Monte Carlo (MCMC) methods. The back-transformed sample $\lambda_C^{(1)}, \ldots, \lambda_C^{(M)}$ represents the predictive distribution of λ_C and hence the MAP prior for the future control group.

As the simulated MAP prior distribution is not a gamma distribution anymore, which we assumed in Section 2, we fit heuristically a mixture of gamma distributions with $K_C \geq 1$ components to preserve the conjugacy property:

$$\hat{\pi}(\lambda_C) = \sum_{k=1}^{K_C} w_k \cdot Gam(\alpha_{Ck}, \beta_{Ck}) \tag{16}$$

The parameters can be obtained with ML methods for a fixed number of components K_C. This approximate representation will be more efficient than the raw MCMC sample in the subsequent Section 5.2, when the posterior distribution is obtained through the combination with new exponentially distributed data on the control treatment C. The goodness-of-fit is assessed with the Kullback-Leibler (KL) divergence, which is a measurement of the discrepancy of the Rao-Blackwellized density estimate[28] to the approximated mixtures of gamma distributions.

4 Design optimization and operating characteristics

After the derivation of the prior distributions for the rate parameters, the remaining design parameters need to be specified during the planning stage. This concerns finding a well-suited sample size and the effect and probability thresholds of the Bayesian Go–NoGo criteria. The criteria are Bayesian in so far as all information that is available at the decision point will be addressed in the posterior distribution of the treatment effect. However, because we consider a multitude of future experimental treatments E and aim to further develop only the most promising compounds out of an existing Phase II portfolio, the trial design is optimized from a frequentist point of view in terms of its OC. In more detail, the OC of a particular design comprise of the probability of a Go or NoGo decision and the probability of an indeterminate result for a given assumed true treatment effect. For complex designs, the simulation of clinical trials is usually necessary. Figure 2 gives an overview of the methodological procedure. It illustrates the distinction of the optimization process at the planning stage and the subsequent actual conduct of the PoC trial to decide about the future development of the compound.

After the prior for λ_C is defined, it is combined with data from the PoC trial (see Figure 2). PoC trial data are simulated from an exponential distribution with constant new λ_C, and the Go–NoGo criterion is evaluated. This is done N_S times. The probability of continuing the development erroneously, can be estimated as follows: we simulate N_S trials with the assumption of no treatment effect, i.e. with a HR equal to 1. A meaningful value smaller than 1 is possible but not followed up now. The proportion of successful trials is then an estimate of the false-positive rate (FPR) of the decision procedure. This is analogous to the concept of type 1 error in the assessment of significance to an observed trial result.

Estimates of the true-positive rate (TPR), i.e. the probability of detecting an assumed fixed clinical relevant treatment effect when the effect actually exists, can be obtained in the same manner. Thus, further N_S trials are simulated with a fixed plausible borderline value of the treatment effect for which the novel treatment should be approved. The proportion of successful trials is then an estimate of the TPR analogous to the concept of power in

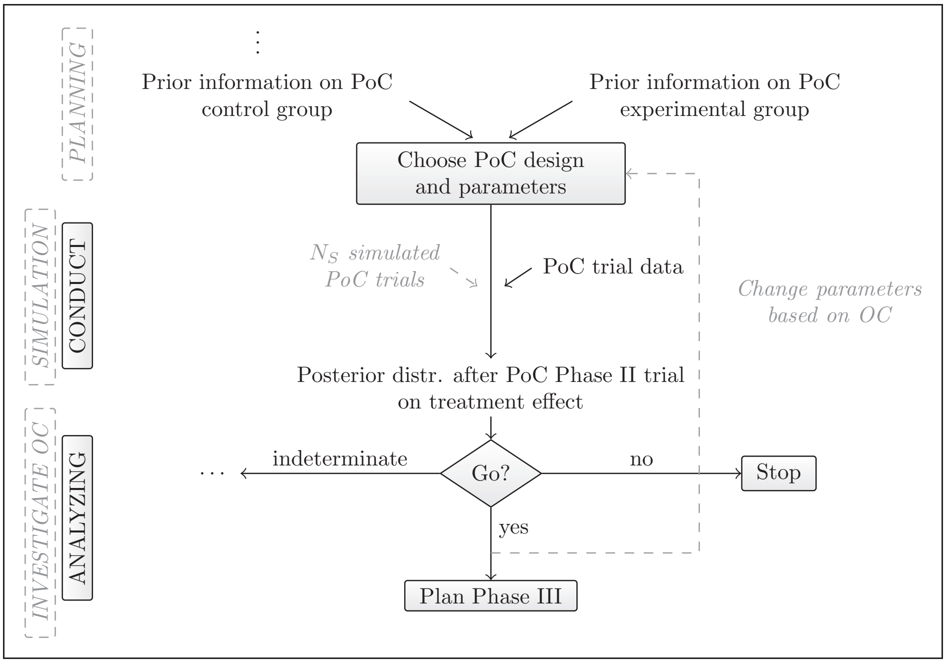

PLANNING
SIMULATION
CONDUCT
INVESTIGATE OC
ANALYZING
⋮
Prior information on PoC control group
Prior information on PoC experimental group
Choose PoC design and parameters
N_S simulated PoC trials
PoC trial data
Change parameters based on OC
Posterior distr. after PoC Phase II trial on treatment effect
…
indeterminate
Go?
no
Stop
yes
Plan Phase III

Figure 2. Methodological approach, simulation and optimization loops added (in gray, italics, dashed).

single clinical trial planning. Note that if we allow for indeterminate results, the FPR is defined as “1 minus TPR minus indeterminate rate”.

Often a whole success, failure, and indeterminate probability curve for varying assumed treatment effects are computed to compare the OC of different designs.[12,21] In the following, we do not consider an indeterminate region and therefore set $\theta_L = \theta_U$ and $\gamma_s = 1 - \gamma_f$.

For the design optimization, it has to be agreed on the maximal permitted error rates and the minimal relevant treatment effect that should be detected, as in classical sample size calculation. These agreements depend mainly on the indication, the medical need for the indication, the company’s portfolio, and the company’s risk tolerance. For example, a “reasonable” anti-tumor treatment is expected to achieve a HR smaller than or equal to 0.75. Also we target for 80% true-positive decisions for a meaningful treatment effect size, in order to improve the number of successful Phase III trials in comparison to the past.

Conditional on the maximal permitted error rates and the minimal relevant treatment effect, the following design parameters have to be optimized for a simple Go criterion (1):

1. the required number of events d,
2. the required number of patients $n = n_E + n_C$,
3. the allocation ratio to the treatment groups $E : C$,
4. the probability threshold $\gamma_s (= 1 - \gamma_f)$, and
5. the effect threshold $\theta_L (= \theta_U)$.

We assume that an appropriate treatment scheme including dose level and dosing schedule for the two groups are given. We note that the provided budget, a reasonable trial duration, and the incidence of the disease will contribute as an informal upper limit for the size of the trial.

For simplicity, we reduce the five-dimensional optimization task to two dimensions by pre-fixing some parameters. At first, the relation between the overall number of events and the number of patients is fixed, say $d = \frac{3}{4} n$. Note that the choice of the factor $\frac{3}{4}$ might be quite arbitrary but is based on historical experience regarding drop-out behavior and restrictions regarding timelines or costs. Furthermore, only reasonable allocation ratios $(E : C)$ are considered, e.g. 2:1 and 1:1. Finally, the probability threshold γ_s is fixed. This is sensible as the probability and efficacy thresholds of one criterion are highly correlated. Despite the simplification, this heuristic proceeding can result in an exhausting trial-and-error procedure. The optimal design has to be understood as the vector of design parameters that fulfills the constraints on the error levels and minimizes n_E and n_C. In particular, the critical boundary θ_L and the (minimal) sample sizes n_E and n_C are obtained. The detailed optimization steps are described in Web Appendix A.

5 Case study: Lung cancer trial

To show an application of the presented approach, a subset of the LUME-Lung 1 trial[29] will serve as the Phase II PoC trial with PFS as primary endpoint. The trial is actually a randomized, double-blinded Phase III trial that compared the combination of Nintedanib and Docetaxel versus Docetaxel monotherapy in patients with non-small-cell lung cancer after failure of first-line chemotherapy. The advantage in using a random subset of the Phase III data is that the sample size of the subset mimicking the Phase II trial is not fixed in advance and the determined optimal sample size / allocation ratio can be applied. In the final analysis, a simple Go criterion without an indeterminate region based on a relevant treatment effect size will be evaluated.

5.1 Historical data and corresponding MAP prior

The first task playing a big role at the planning stage is the collection of relevant former trials and the extraction of required key information. Because the control treatment is an already approved treatment for the target population, there exist published summary results, but not for the experimental treatment. No patient individual data was available. We observe that the published Kaplan-Meier (KM) estimates do not speak against exponential distributions.

Regarding the patient population, the administered dose-level, the number of cycles, and the endpoint PFS, we identified three relevant large Phase III trials with similar observation periods: INTEREST,[30] ZODIAC,[31] and REVEL.[32] The control groups in these three trials are deemed exchangeable with each other and with the control group of the future trial. As a matter of fact, the trial REVEL was not finished at the time of planning the original

LUME-Lung 1 trial (recruitment started in December 2008), but for the illustrative purpose of this retrospective data example the study can nevertheless be included.

From the three different sources, we have to determine the parameter of the exponential distribution for each historical control group, see Table 1.[33] With the three published trials, we derive the prior distribution of the rate parameter λ_C for the future control group with the MAP approach. The resulting MAP prior distribution (13) is visualized in Figure 3. In addition to the non-parametric density estimates (histogram of MCMC sample and Rao-Blackwellized density estimate), the ML estimates of a one- and two-component gamma distribution are displayed. The histogram was too difficult to handle computationally, in particular, for the prior-to-posterior calculation in equation (18). Whereas the simple gamma distribution $Gam(16.59, 5.46)$ does not lead to a satisfactory approximation, a mixture of two components

$$\hat{\pi}(\lambda_C) = 0.78 \cdot Gam(153.75, 51.86) + 0.22 \cdot Gam(3.12, 0.89) \tag{17}$$

provides an approximation that is already sufficiently close to the empirical distribution in terms of the KL divergence. Moreover, the simple component is more sensitive to the chosen seed for the MCMC simulation, especially for single large values in the MCMC sample, whereas the two component distribution is rather stable. The ML estimate of a three-component mixture appears to be numerically intractable. However, although the peak of the empirical distribution was not completely filled by the two-component mixture, the KL divergence to the Rao-Blackwellized density estimate is already quite small. The two-component mixture (17) is used in the subsequent sections as the MAP prior for the control group C. Moreover, the prior effective sample size (ESS)

Table 1. Extracted information from publications on patients treated with Docetaxel.

Study	Recruitment start date	Number of events d_h	Median PFS (years)	Total obs. time $\hat{T}_{+h}$ (years)	$\hat{\lambda}_h$ (1/years)
INTEREST[30]	Mar 2004	544	0.23	177	3.08
ZODIAC[31]	May 2006	~620[a]	0.27	205	3.03[b]
REVEL[32]	Dec 2010	429	0.25	155	2.77

[a] The Docetaxel arm of the ZODIAC trial contains 697 patients but the number of events is only given for the whole trial population. Then, the number of progressed patients has been estimated graphically based on the reported KM curve.[34]

[b] Estimate is based on reported PFS at 0.5 years (18%) instead of median PFS as it fits better to the complete KM curve.

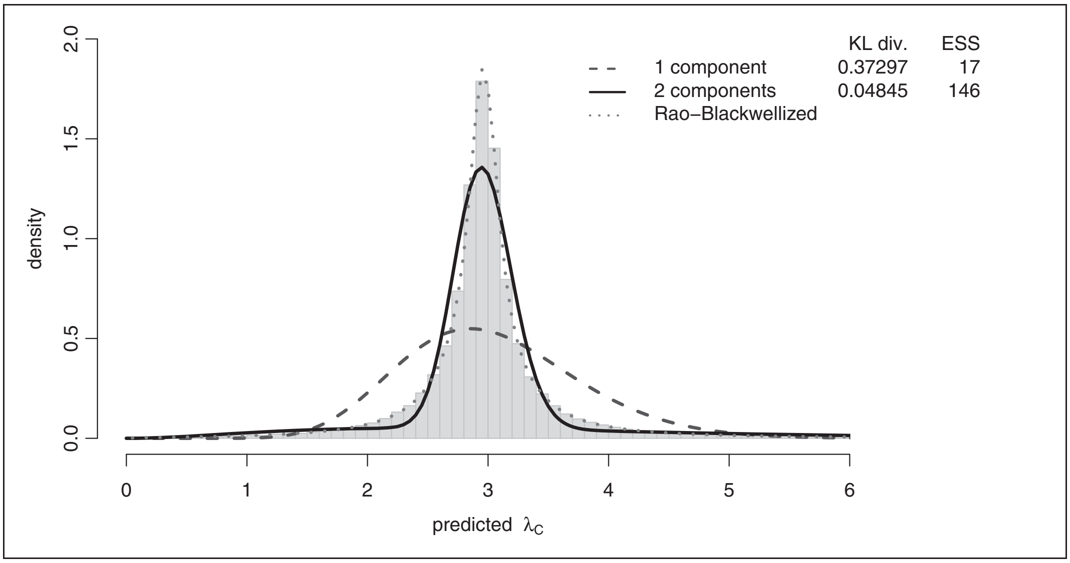


Figure 3. Histogram of the MCMC sample $\lambda_C^{(1)}, \ldots, \lambda_C^{(10\,000)}$ (MAP prior (13)) with Rao-Blackwellized density estimate (non-parametric) and ML density estimates for the one- and two-component mixture gamma distribution.

is calculated after Morita et al.[35] (see Figure 3). The ESS is in the time-to-event setting the equivalent number of events contained in the prior distribution. Note that the resulting ESS only refers to the control group but not to the whole 2-arm trial.

For the Nintedanib plus Docetaxel group E, we use the non-informative gamma prior $Gam(0.01, 0.01)$ since no relevant historical data is available.

5.2 Optimization

Further design parameters are optimized based on the OC as outlined in Section 4 and Web Appendix A. We apply a simple Go criterion and would proceed to Phase III, if

$$\mathbb{P}(\theta < \theta_L | \text{data and prior information}) > 0.5 \tag{18}$$

Otherwise the development will be stopped or at least further aspects have to be taken into account. The effect threshold θ_L has to be optimized, whereas the probability threshold is fixed ($\gamma_s = 0.5$). The latter is equivalent to requiring a posterior median HR smaller than θ_L. For the illustration, we choose balanced error rates of 20%. For the simulation of clinical trial data we assume exchangeability and therefore use a constant $\lambda_C = 3$ in line with Table 1 and the mode of the MAP prior (see Figure 3), and adapt the experimental parameter λ_E according to the assumed fixed HR. Further calculations with variable λ_C sampled from the MAP are shown in Web Appendix C. More precisely, for the FPR we require

$$Prob(\text{Go} | \theta = 1, \lambda_C = 3) \approx 0.2 \tag{19}$$

where "Go" is described in equation (18) and takes up the uncertainty described by the MAP (see equation (17) and Figure 3) whereas "*Prob*" denotes the frequentist interpretation of probability, meaning the long-run frequency. For the TPR, we require

$$Prob(\text{Go} | \theta = 0.75, \lambda_C = 3) \approx 0.8 \tag{20}$$

The minimal required treatment effect size is chosen as 0.75 leading to $\lambda_E = 2.25$. This is a realistic borderline treatment effect that a new oncology substance should show to be clinically relevant. The event times of the two groups for the future trial are simulated independently with rate parameters $\lambda_C = 3$ (similar to the historical controls), $\lambda_E = 3$ or $\lambda_E = 2.25$, and uniform accrual. The event times occurring after the d-th event are censored. The two-dimensional optimization is performed in a step-wise fashion:

1. Determination of the minimal sample size ($n_E + n_C$) that fulfills approximately the requirements on the error rates conditional on a 2:1 allocation ratio ($E : C$) and $d = \frac{3}{4}(n_E + n_C)$ equivalent to a study duration of 0.5620 years on the by-patient time scale (derivation see Web Appendix 1).
2. Determination of the effect threshold θ_L such that the error rates are balanced.

The initial values for the sample size are chosen in accordance with the typical size of an oncological Phase II trial, e.g. 120 patients in total. If the error rates are too large, i.e. FPR > 0.2 and TPR < 0.8, the sample size will have to be increased. On the other hand, if the error rates are smaller than required, the sample size can be decreased. The initial value for the effect threshold θ_L is expected to be slightly above $\theta = 0.75$ and is therefore chosen as 0.8. We tried out the range from 0.8 to 0.94 with step size 0.01 (see left panel of Figure 4). If the error rates are not balanced, the effect threshold will have to be adapted accordingly. This means for our example, if FPR > 0.2 and TPR > 0.8 the threshold θ_L will have to be chosen smaller. On the other hand, if FPR < 0.2 and TPR < 0.8, the threshold θ_L will have to be chosen larger. The step-wise tailoring of the design parameters leads finally to the following optimal design using the MAP given the constraints for simplification

$$n_E = 60, \quad n_C = 30, \quad d = 68, \quad \theta_L = 0.86 \tag{21}$$

The corresponding OC are visualized in Figure 4. In the left panel, the determination of the optimal effect threshold θ_L is shown for the final fixed sample size using the MAP prior or solely non-informative priors, respectively. Applying the MAP prior, for $\theta_L = 0.86$ the two error rates are approximately 20% as well as for

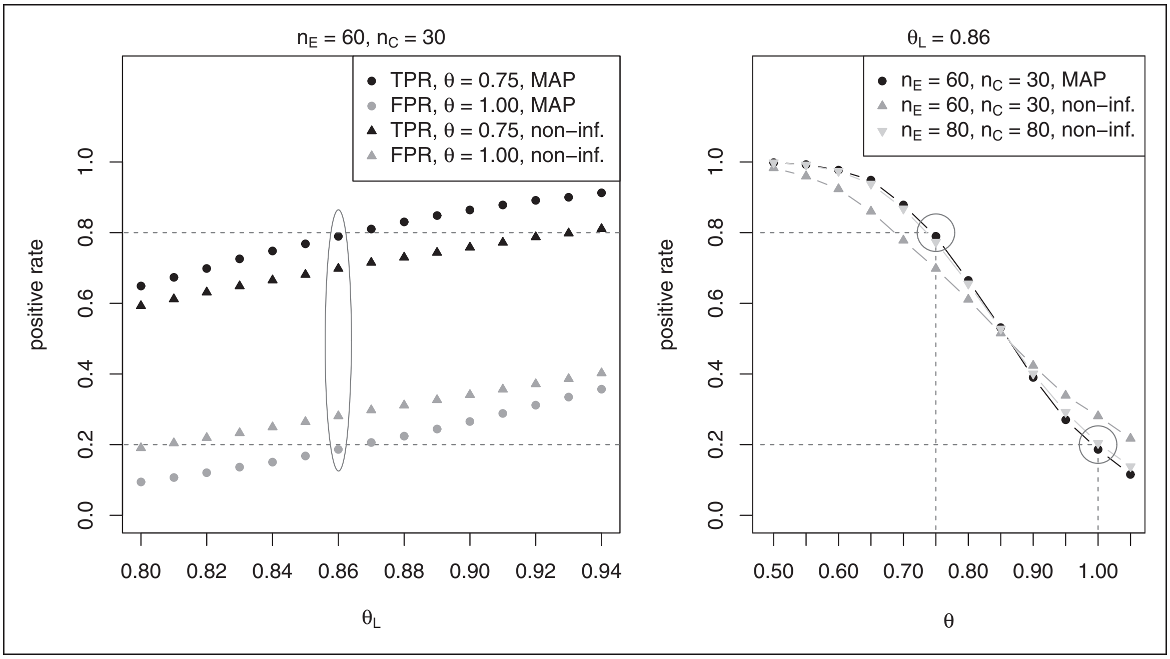


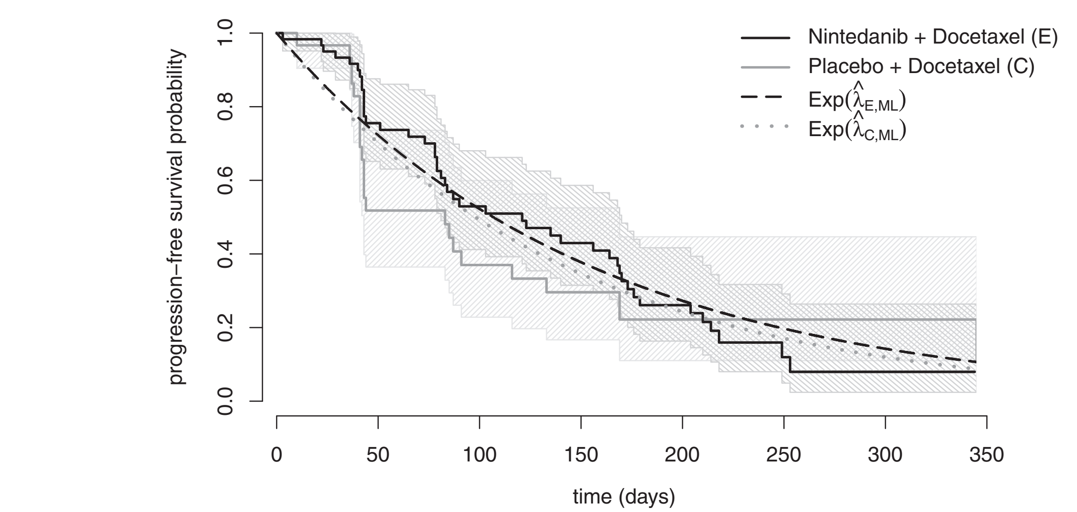
Figure 4. Left panel: OC for varying effect threshold θ_L, fixed $\lambda_C = 3$, fixed assumed true treatment effects ($\theta = 1$ or $\theta = 0.75$), and fixed sample size. Right panel: OC of three different designs for varying θ with fixed θ_L, and fixed $\lambda_C = 3$.

$\theta_L = 0.87$. We chose the smaller value for the final analysis. In the right panel, the corresponding success curve is shown in black. In addition, the success curves with non-informative priors on both groups are displayed for two different sample sizes representing the analysis without the historical information on the control. It can be seen that without incorporating the historical information, we would need almost twice as many patients to receive the same OC, namely 160 patients instead of 90. The possible sample size reduction with similar efficiency when applying the MAP might be better suited to quantify the influence of the historical information than the ESS.

Moreover, the increased error rates (Figure 4, left panel) or the flatter success curve (right panel) is shown for the smaller sample size with the non-informative priors. This setting should also be considered to provide later on a reference analysis solely based on the newly observed data. Although the informative priors represent best the state of knowledge of the decision makers, it is sensible to provide a reference analysis to evaluate the dependence of the evaluated criteria on the historical information. For the reference analysis, we obviously cannot increase the sample size to achieve the same OC, instead we have to assume that the error rates become larger than 20%. However, we can evaluate whether the optimal θ_L conditional on $n_E = 60$, $n_C = 30$, and $d = 68$ is far away from the MAP based optimum indicating potential non-robustness. But it actually turns out that we get the same optimal values for both designs with and without the MAP information.

Web Appendix C shows in addition the OC, depending on θ_L (Figure 4, left panel), when the λ_C of the new trial is drawn randomly from the two-component mixture MAP. The additional uncertainty has the consequence that $n_E = 60$ and $n_C = 30$ are no longer sufficient to reach a TPR of 0.8 together with an FPR of 0.2. Instead, $n_E = 80$ and $n_C = 40$ are necessary, which have to be compared with $n_E = 80$ and $n_C = 80$ in the non-informative case (Figure 4, right panel). Also, we investigated the TPR and FPR with λ_C sampled from the histogram and they are indistinguishable from those obtained from the two-component mixture MAP.

For comparison, two common significance-based criteria were also investigated instead of equation (18). The two groups of the Phase II study were compared with a one-sided test non-parametrically (Logrank test) as well as parametrically based on exponential distributions (see equation (11) and Lawless,[25] p. 156f.). Naturally, the FPR was identical with the alpha level. The TPR, i.e. the power for finding the ratio $\hat{\lambda}_E/\hat{\lambda}_C$ significantly <1 given $\lambda_C = 3$ and $\lambda_E = 2.25$, was in both cases slightly smaller than, but similar to the non-informative case in Figure 4. Note that a direct comparison is impossible as equation (18) is based on effect size whereas the test criteria are based on signal strength.



Number at risk								
Nintedanib + Docetaxel	60	41	27	21	12	3	2	0
Placebo + Docetaxel	30	14	10	8	5	4	4	0

Figure 5. KM estimates with pointwise 95% CIs of the random subset of the LUME-Lung 1 trial and fitted exponential survival curves. ML estimates of the rate parameters are $\hat{\lambda}_C \approx 2.58$ (95% CI: 1.53, 3.64) and $\hat{\lambda}_E \approx 2.37$ (95% CI: 1.68, 3.07) per year.

5.3 Verification on the actual data

We now want to provide a reality check of the assessed optimal trial design on the basis of the subset of the LUME-Lung 1 study[29] in size corresponding to the optimized sample size. As a surrogate for the true first 90 patients the randomization date of the first patient is selected randomly and the following 60 patients allocated to Nintedanib plus Docetaxel are selected. For the control group, we take randomly 30 patients out of the following 60 patients allocated to the real Placebo plus Docetaxel group such that the randomization dates in both groups are evenly spread. In practice, we would wait until 68 patients have observed the event of interest, i.e. progressed or died, whatever comes first. This would happen after 367 days since the first patient entered the study. At this time point, three patients in the control and one patient in the Nintedanib and Docetaxel group of the subset have not progressed and have not been censored, these four subjects are artificially censored for the reality check. The corresponding KM estimates and the estimated exponential survival curve with the ML method are shown in Figure 5. Median PFS at the final analysis in the Phase II population is 121 days (0.33 years) in the Nintedanib plus Docetaxel group versus 83 days (0.23 years) in the Placebo plus Docetaxel group. The 95% confidence interval (CI) is (81, 170) and (43, 169), respectively. Hence, the concurrent controls appear healthier than the historical controls, c.f. Table 1. But $\lambda_C = 3$ is not excluded as the 95% CI is (1.53, 3.64) which is derived from equation (14). Also, although there appears to be some interval censoring for the PFS endpoint through the regular six weekly measurements, the classical pointwise confidence intervals of the KM estimates do not contradict the exponential distribution overall.

The Go criterion (18) is fulfilled using the MAP prior whereas it is not fulfilled in the reference analysis. We want to mention that the results of the reference analysis coincide with the classical Log-rank test. The corresponding posterior distributions of λ_C and the HR θ are shown in Figure 6. With respect to the sharpness of the two distributions it becomes obvious that the treatment effect can be estimated more precisely by borrowing from the historical trials. We do even see that the MAP prior is considerably more sharply peaked and contains considerably more information than the new data.

As verification, the LUME-Lung 1 study showed a significant benefit from Nintedanib plus Docetaxel to Docetaxel alone ($\text{HR} = 0.79$ with 95% CI (0.68, 0.92), $p = 0.0019$) and Nintedanib has meanwhile been registered in the European Union and more than 20 additional countries for this indication.

6 Conclusions and outlook

We have shown that the inclusion of prior information in the sense of an informative MAP into the PoC decision will lead either to improved OC or to a relevantly reduced size of the actual PoC trial.

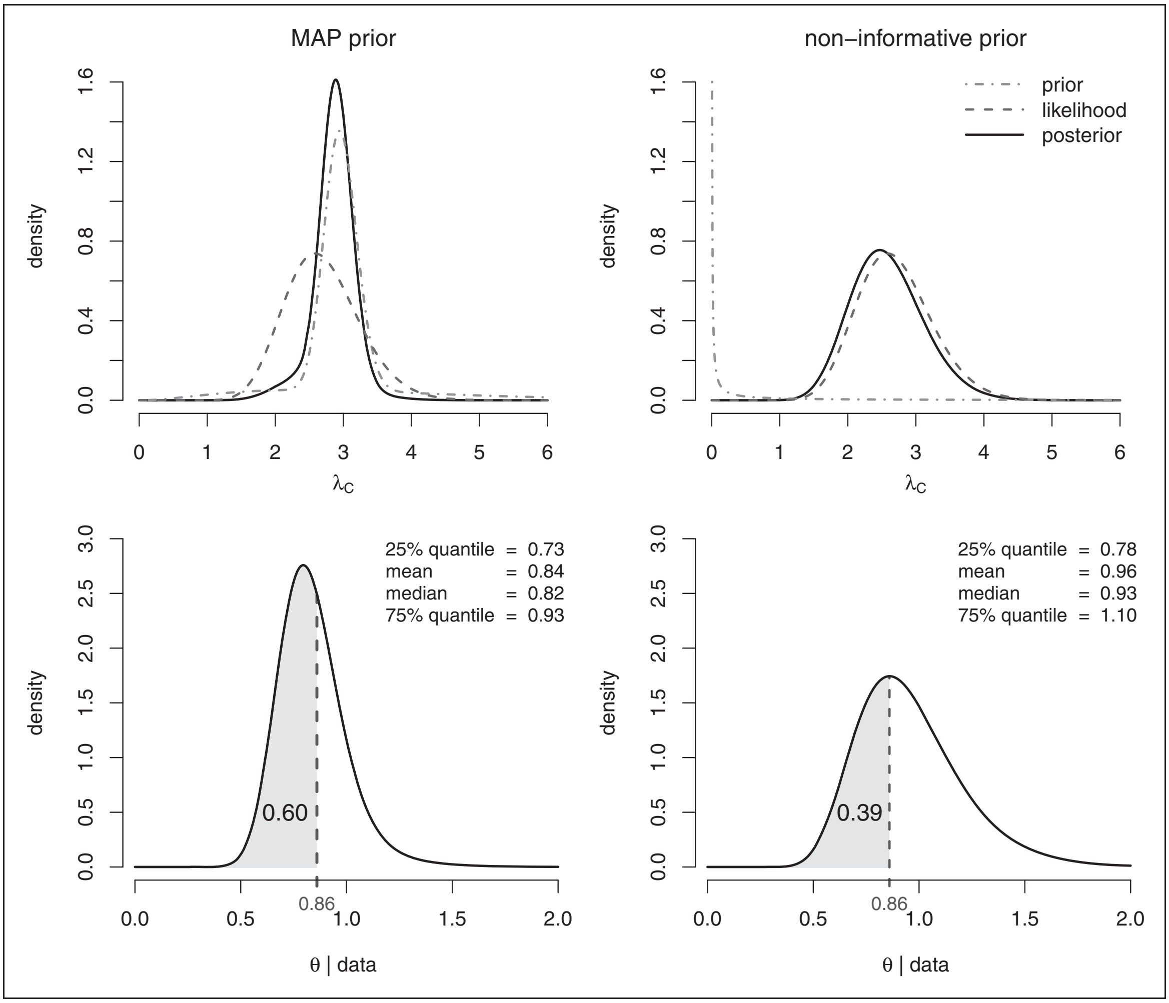


Figure 6. Posterior distribution of λ_C and the HR after the PoC trial using either the MAP prior (17) or a non-informative prior on the control group, and calculation of the degree of certainty for success based on the optimal effect threshold 0.86.

In this article, we have elucidated on the MAP/optimization approach in the time-to-event setting. PoC criteria and the design of the PoC trial have been optimized. The concept has been illustrated by a real trial example from second line non-small-cell lung cancer. In particular, the case study has confirmed that by a reasonable inclusion of historical data the precision of the treatment effect estimate and hence the probability for a correct PoC decision can be largely increased. The MAP distribution has relevantly influenced the final Go decision, as demonstrated by Figure 6 (upper left corner).

In the context of PoC decision making, it is important to clarify the basis for the final decision. This has often been based exclusively on the data coming from the actual PoC trial. Nevertheless, implicitly other sources have been taken into account by the experts. Here, a systematic approach to include historical data will improve the objectivity and transparency of the decision making and will allow quantification. Therefore, final evidence for PoC should not come from the PoC trial alone, but should always consider all available data at the decision time point. Later update of the prior or of the criterion or even of the targeted sample size is possible, but their disadvantage is that the quantitative framework will become arbitrary.[36] In our experience, however, the actual PoC decision will and should not rely on the Go criterion alone. There are usually other factors, most prominently safety, that will play a role in the final decision. So, any Go criterion should be seen as guidance for decision making but not as a strictly binding rule.

The PoC criterion itself was based on the actual treatment effect. It has been suggested to use the probability-of-success (PoS), regarding the prior probability of significance[37,38] in confirmatory phase as an alternative criterion. In the literature this is also called assurance.[39,40] The actual criterion would then be as follows. A Go will be achieved, if the PoS is at least $x\%$, where the PoS is based on the posterior after the PoC trial and the assumed design of the confirmatory trial where the number x needs to be specified. Usually, such a PoS-based approach assumes that all details of the confirmatory trial including its sample size are already fixed in order to be able to calculate the PoS based on the posterior of the PoC trial. Although this approach might be reasonable in some scenarios,[41,42] we have opted to base the decision on the actual treatment effect as we assume the confirmatory phase to be still flexible at the time-point of designing the PoC trial. It is worthwhile to remark here that it is very straightforward to transform our approach to a PoS-based approach if necessary.

We also investigated the case that PoC was considered to be equivalent to a significant difference between the groups C and E in Phase II. The OC was slightly inferior, but quite similar to the non-informative case in Figure 4.

The Bayesian approach used in this article is to model the two treatment groups separately. Thereby an inclusion of historical information on a treatment group basis is enabled. In our experience, this is the more frequent case compared to having historical information regarding the treatment effect (in our case the HR). In particular, this holds for early phase/PoC trials whereas the situation might be different in later phases. If indeed the information is available for the treatment effect it might make more sense to turn to a normal-normal model for the HR instead of modeling the treatment groups separately. Note that if both are available, i.e. information for the single treatment groups as well as for the HR, it is also possible to have this included into our approach but that is beyond the scope of this article.

In our example, one could speculate whether a slight trend in time regarding the hazard rates λ_h for the control treatment in Table 1 is observable and the assumed exchangeability is actually not present. Whereas in the earlier trials (INTEREST and ZODIAC) the hazard rate has been slightly above 3, the results from REVEL as well as the data from the PoC trial show a hazard rate of 2.77 and 2.58, respectively. In particular, in such a case it is important to consider a range of possible values for the assumed true future hazard rate λ_C and to investigate the subsequent optimization. This has been performed for the example given (see Web Appendix B) and has shown robustness across a certain range of λ_C. Web Appendix C shows OC curves when λ_C has been simulated from the mixture MAP. The additional uncertainty is reflected in higher Phase II sample sizes that are necessary to achieve a similar TPR and FPR, but there is still a considerable benefit relative to the non-informative case ($80 + 40$ vs. $80 + 80$).

Generally, when using any historical information, a crucial question is whether the historical data is representative for the scenario and the new trial considered. More specifically, for the MAP approach, exchangeability is assumed between the old and the new data, i.e. in our case the underlying overall hazard rate is assumed to be equal across all data considered including the new one. This does not mean that the underlying expected hazards for specific trials need to be the same or even the observed ones. Still this assumption of exchangeability should be carefully evaluated and discussed with non-statistical experts beforehand to assure that it is indeed fulfilled. Thorlund et al.[43] discuss effects that may arise if this evaluation is not done to the extent needed. It may also make sense to check in-between the trial or afterwards for major deviations from these assumptions, e.g., based on the prior and the actual observed hazard rate for the control group. If that is done in an interim analysis during the course of the trial this may lead to an adaptation of the sample size needed in the PoC trial (see also Schmidli et al.[5] for similar considerations). Note that this technique could be added to the approach considered but is not investigated further in this article.

How robust is our method against deviations from the exponential distribution? We investigated with our exponential model data that were actually Weibull distributed with shape parameters 3/2, 5/4, 10/9, 9/10, 4/5, and 2/3. The results are shown in Web Appendix D. The TPRs and FPRs differ from those obtained with truly exponential data substantially. Further research is necessary here. In particular, piecewise constant hazards allow for quite flexible modeling of survival data. At present, non-robustness within the Weibull family has to be stated, if the shape parameter moves away from 1.

With respect to the inclusion of the historical information, instead of a MAP approach a meta-analytic-combined (MAC) approach may be alternatively implemented.[20] The advantage of the MAC approach would be that, starting with a prior distribution, new external results are combined with the historical ones in a meta-analysis in a step-wise manner. While this sounds attractive there are also some downsides connected. Most prominently the organizational requirements increase massively. In this context, we would also like to stress that the results of an analysis based on non-informative priors should always be reported in addition to the primary analysis in order to allow for unbiased future meta-analyses (see also Sung et al.[44]).

In our approach, we have chosen to use the MAP prior derived from the historical data directly without any further possible discount. This is leading to a relatively large weight of the historical information compared to the actual data generated. The weight or more precisely the difference in the weights of the historical information will also influence the optimal allocation ratio. Obviously, when having different amount of prior information for the different treatment groups a 1:1 ratio is no longer optimal. Within our optimization one could either define some discrete ratios to be considered (e.g. 2:1) or investigate different allocation ratios on a more or less continuous basis. A good starting point here might be an allocation ratio that on average may lead to approximately balanced posterior information in the two treatment groups. Note, however, that the amount of information generated in the PoC trial should at least allow for a proper assessment whether exchangeability has been correctly assumed. Therefore, in particular, a too small number of patients in the control group of the PoC trial should be avoided. Hence, too extreme allocation ratios should not be considered even though they might be technically optimal if exchangeability is fulfilled.

A potential downside of the systematic integration of historical data is an increased effort in statistical planning beforehand. The performance of a systematic review regarding available data on its own is usually a demanding exercise. After that, the data needs to be checked for relevance and for similarity to the planned setting of the PoC trial. The MAP prior needs to be generated and potentially approximated by a mixture of conjugate priors. The PoC criterion needs to be optimized via back-and-forth iteration. Finally, the actual analysis might be more complicated and results might be not-so-easy to communicate to non-statisticians. Nevertheless, one may argue that the enhanced decision making or alternatively the smaller sample size needed for the PoC should easily excel these additional efforts with respect to capacities and resources involved. Even in a frequentist or non-informative setting a systematic review of the available data in the considered setting is necessary. Regarding the aspect of communication of the approach and the results, it is always important to provide upfront explanation to the non-statistical team members.

We have assumed the target error rates to be given and fixed at 20%. The 20% have been chosen as a good compromise between an efficient resulting sample size and a still tolerable error rate. It is also obvious that there might be examples where other (smaller) error rates are better suited. In particular when thinking about therapeutic areas and indications where patients are easier to recruit and project values are high. There is also literature available that aim at optimizing these error rates in the context of the whole development chain.[16] Note that we deliberately opted to not follow such an approach. Overall optimization of the full development usually demands for a very detailed prior knowledge regarding the planned design and size of the confirmatory trials to follow. For many applications this might be an unrealistic assumption as these factors should depend on the actual results observed within Phase II.

How can the described optimized PoC decision procedure improve Phase III success rates? Historically, success rates of 60% and in oncology even 45% have been reported.[45] As our correct-guess rates of 80% are analogous to sensitivity and specificity of a diagnostic procedure, the positive and negative predictive value (PPV and NPV) of an observed Go signal or NoGo signal, respectively, are of interest. The PPV and NPV depend from the unknown "prevalence" of effective compounds in the portfolio described at the beginning of Section 4. Elementary calculation (see Web Appendix E) shows that the PPV will exceed 60% already if the prevalence is greater than 27.3%. Should the prevalence be 50%, the PPV will rise to 80%, a relevant improvement to 60% or 45%.

Another criticism when incorporating historical data is a potential inflation of the false-positive error. In general, two situations have to be distinguished here. In the first situation, the exchangeability assumption might be actually violated by having a different underlying overall hazard for the historical data compared to the new data. In the second situation, exchangeability may still hold but randomness would lead to a difference in the random mean effect and/or the observed effect between the historical data and the data obtained from the PoC trial. Note that even within the historical data exchangeability might be violated but that is not subject of our considerations here. When exchangeability is incorrectly assumed there is indeed a potential inflation of the error rates both for the false-positive and the false-negative case. It is important to stress that information is only induced for the control group in our example. Therefore there is no initial tendency regarding the direction of the change in error rates unless differences are expected without considering them regarding the exchangeability assumption. In our case, if the actually observed hazard in the control group is large compared to the prior, the treatment effect based on the posterior will become smaller and vice versa. This will finally lead to two effects, an increase in the FPR for the one and a decrease for the other direction compared to the case when using non-informative priors. In summary and as discussed for our example, it is crucial to evaluate potential error rates for different scenarios of interest. It might be a worthwhile approach to aim for the control of error rates under a range of scenarios (see Web Appendix B). In contrast to this situation, if the difference with respect to the actual or the observed treatment effect in the study is driven by randomness only, it holds that there will be no error inflation here and that the Bayesian approach including an informative prior provides a more informative and more concise overall impression regarding the data and inference to be drawn from them. In this sense, an error rate purely based on assumptions for the hazards in the upcoming PoC trial might not reflect the actual error rates and hence should not be considered. This is well reflected by the fact that the MAP prior approach will dampen extreme observations in the PoC trial, both negative and positive. This facilitates a reasonable way of dealing with unexpected observations in the control group that otherwise might negatively influence the correct decision process. These considerations are complicated by the fact that it is usually hard to decide solely based on the data whether exchangeability indeed holds, i.e., whether deviations are caused by randomness or indeed non-exchangeability. It might be checked based on, e.g., Figure 6 how good the alignment between the prior and posterior is. Additionally it should be investigated after the PoC trial whether there are significant deviations with respect to major covariables. So in our opinion, the best and most robust approach is to have both situations incorporated by using the informative prior but still controlling for a broader range of initial hazards.

There are several possible extensions to the approach presented. First of all, the underlying time-to-event distribution might be extended, e.g., to a Weibull distribution or a piece-wise constant hazard. This will in particular be necessary if there is evidence that the event times are Weibull distributed with shape parameter away from 1. Also non-parametric approaches similar to the one in Cotterill and Whitehead[18] may be considered. The inclusion of covariates might also be an attractive option. However, one should note that more complicated models may also have their downsides regarding the derivation of the prior as well as the precision of parameter estimates. A plausible next step would be to proceed to piecewise-exponential distributions.

Acknowledgements

This article is based on the first author's master thesis at the University of Ulm. The authors thank two referees for valuable suggestions that improved the argumentation.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The first, second and fourth authors are employees of Boehringer Ingelheim Pharma GmbH & Co. KG.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material is available online for this article.

References

1. Lendrem D, Senn SJ, Lendrem BC, et al. R&D productivity rides again? *Pharm Stat* 2015; **14**: 1–3.
2. Kieser M and Hauschke D. Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharm Stat* 2005; **4**: 101–107.
3. Strom BL, Buyse ME, Hughes J, et al. Data sharing – is the juice worth the squeeze? *N Engl J Med* 2016; **375**: 1608–1609.
4. Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. *Clin Trials* 2010; **7**: 5–18.
5. Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; **70**: 1023–1032.
6. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis* 1976; **29**: 175–188.
7. Lecoutre B, Mabika B and Derzko G. Assessment and monitoring in clinical trials when survival curves have distinct shapes: a Bayesian approach with Weibull modelling. *Stat Med* 2002; **21**: 663–674.
8. Thall PF. Monitoring event times in early phase clinical trials: some practical issues. *Clin Trials* 2005; **2**: 467–478.
9. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat* 2014; **13**: 41–54.
10. Grieve AP. Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharm Stat* 2016; **15**: 96–108.
11. Frewer P, Mitchell P, Watkins C, et al. Decision-making in early clinical drug development. *Pharm Stat* 2016; **15**: 255–263.
12. Fisch R, Jones I, Jones J, et al. Bayesian design of proof-of-concept trials. *Ther Innov Regul Sci* 2015; **49**: 155–162.

13. Walley RJ, Smith CL, Gale JD, et al. Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study. *Pharm Stat* 2015; **14**: 205–215.
14. Musser B, Bolognese J, Fayad GN, et al. Early clinical development planning via biomarkers, clinical endpoints, and simulation: a case study to optimize for phase 3 dose selection. *Ther Innov Regul Sci* 2015; **49**: 405–414.
15. Nikolakopoulou A, Mavridis D and Salanti G. Planning future studies based on the precision of network meta-analysis results. *Stat Med* 2016; **35**: 978–1000.
16. Kirchner M, Kieser M, Götte H, et al. Utility-based optimization of phase II/III programs. *Stat Med* 2016; **35**: 978–1000.
17. Götte H, Schüler A, Kirchner M, et al. Sample size planning for phase II trials based on success probabilities for phase III. *Pharm Stat* 2015; **14**: 515–524.
18. Cotterill A and Whitehead J. Bayesian methods for setting sample sizes and choosing allocation ratios in phase II clinical trials with time-to-event endpoint. *Stat Med* 2015; **34**: 1889–1903.
19. Cellamare M and Sambucini V. A randomized two-stage design for phase ii clinical trials based on a bayesian predictive approach. *Stat Med* 2015; **34**: 1059–1078.
20. Neuenschwander B, Roychoudhury S and Schmidli H. On the use of co-data in clinical trials. *Stat Biopharm Res* 2016; **8**: 345–354.
21. Gsponer T, Gerber F, Bornkamp B, et al. A practical guide to Bayesian group sequential designs. *Pharm Stat* 2014; **13**: 71–80.
22. Yuan Y, Guo B, Munsell M, et al. MIDAS: a practical bayesian design for platform trials with molecularly targeted agents. *Stat Med* 2016; **35**: 3892–3906.
23. Götte H, Donica M and Mordenti G. Improving probabilities of correct interim decision in population enrichment designs. *J Biopharm Stat* 2015; **25**: 1020–1038.
24. Ohwada S and Morita S. Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharm Stat* 2016; **15**: 420–429.
25. Lawless JF. *Statistical models and methods for lifetime data*. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2002.
26. Andersen PK, Borgan Ø, Gill RD, et al. *Statistical models based on counting processes*. New York, NY: Springer, 1993.
27. Aalen OO, Borgan Ø and Gjessing H. *Survival and event history analysis: a process point of view*. New York, NY: Springer, 2008.
28. Gelfand AE and Smith AFM. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 1990; **85**: 398–409.
29. Reck M, Kaiser R, Mellemgaard A, et al. Docetaxel plus nintedanib versus docetaxel plus placebo in patients with previously treated non-small-cell lung cancer (LUME-Lung 1): a phase 3, double-blind, randomised controlled trial. *Lancet Oncol* 2014; **15**: 143–155.
30. Kim ES, Hirsh V, Mok T, et al. Gefitinib versus docetaxel in previously treated non-small-cell-lung-cancer (INTEREST): a randomized phase III trial. *Lancet* 2008; **372**: 1809–1818.
31. Herbst RS, Sun Y, Eberhardt WEE, et al. Vandetanib plus docetaxel versus docetaxel as second-line treatment for patients with advanced non-small-cell lung cancer (ZODIAC): a double-blind, randomised, phase 3 trial. *Lancet Oncol* 2010; **11**: 619–626.
32. Garon EB, Ciuleanu TE, Arrieta O, et al. Ramucirumab plus docetaxel versus placebo plus docetaxel for second-line treatment of stage iv non-small-cell lung cancer after disease progression on platinum-based therapy (REVEL): a multicentre, double-blind, randomised phase 3 trial. *Lancet* 2014; **384**: 665–673.
33. Parmar MKB, Torri V and Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998; **17**: 2815–2834.
34. Tierney JF, Stewart LA, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007; **8**: 16.
35. Morita S, Thall PF and Müller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008; **64**: 595–602.
36. Carroll KJ. Decision making from phase II to phase III and the probability of success: reassured by "assurance"? *J Biopharm Stat* 2013; **23**: 1188–1200.
37. Spiegelhalter DJ, Freedman LS and Blackburn PR. Monitoring clinical trials: Conditional or predictive power? *Control Clin Trials* 1986; **7**: 8–17.
38. O'Hagan A, Stevens JW and Campbell MJ. Assurance in clinical trial design. *Pharm Stat* 2005; **4**: 187–201.
39. Gasparini M, Di Scala L, Bretz F, et al. Predictive probability of success in clinical drug development. *Epidemiol Biostat Public Health* 2013; **10**: e8760–1-e8760–14.
40. Gamalo MA, Wu R and Tiwari RC. Bayesian approach to non-inferiority trials for normal means. *Stat Methods Med Res* 2016; **25**: 221–240.
41. Nehmiz G. *An optimization approach for the early phases of drug development*, www.biometrische-gesellschaft.de/arbeitsgruppen/bayes-methodik/workshops/2007-reisensburg.html (2007, accessed 15 November 2016).
42. Whitehead J. Designing phase II studies in the context of a programme of clinical research. *Biometrics* 1985; **41**: 373–383.

43. Thorlund K, Druyts E, Aviña-Zubieta JA, et al. Why the findings of published multiple treatment comparison meta-analyses of biologic treatments for rheumatoid arthritis are different: an overview of recurrent methodological shortcomings. *Ann Rheum Dis* 2013; **72**: 1524–1535.
44. Sung l, Hayden J, Greenberg ML, et al. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *J Clin Epidemiol* 2005; **58**: 261–268.
45. Hay M, Thomas DW, Craighead JL, et al. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014; **32**: 40–51.
46. Heitjan DF, Ge Z and Ying GS. Real-time prediction of clinical trial enrollment and event counts: a review. *Contemp Clin Trials* 2015; **45**: 26–33.

Appendix 1

The relation of the required number of observed events across both groups and the study duration is given as follows. We have mentioned in Section 5.2 that $n_E = 2n_C$ and $d := d_E + d_C = \frac{3}{4}(n_C + n_E) = \frac{9}{4}n_C$. If we now assume that $\lambda_E = 2.25$ and $\lambda_C = 3$, the fraction of patients without an event up to time t will be given by $2n_C(1 - \exp(-2.25t)) + n_C(1 - \exp(-3t))$ which is equated to $\frac{9}{4}n_C$. Without loss of generality, we set $n_C = 1$ and obtain $t \approx 0.5620$. For uniform accrual see Heitjan et al.[46]