gsbDesign: An R Package for Evaluating the Operating Characteristics of a Group Sequential Bayesian Design

Florian Gerber
University of Zurich

Thomas Gsponer
University of Berne

Abstract

The R package **gsbDesign** provides functions to evaluate the operating characteristics of Bayesian group sequential clinical trial designs. More specifically, we consider clinical trials with interim analyses, which compare a treatment with a control, and where the endpoint is normally distributed. Prior information can either be specified for the difference of treatment and control, or separately for the effects in the treatment and the control groups. At each interim analysis, the decision to stop or continue the trial is based on the posterior distribution of the difference between treatment and control. The decision at the final analysis is also based on this posterior distribution. Multiple success and/or futility criteria can be specified to reflect adequately medical decision-making. We describe methods to evaluate the operating characteristics of such designs for scenarios corresponding to different true treatment and control effects. The characteristics of main interest are the probabilities of success and futility at each interim analysis, and the expected sample size. We illustrate the use of **gsbDesign** with a detailed case study.

Keywords: adaptive design, Bayesian inference, clinical trials, clinical trial monitoring, expected sample size, group sequential design, interim analyses, operating characteristics, stopping criteria.

This document was published in a similar form in

Gerber F, Gsponer T (2016). “gsbDesign: An R Package for Evaluating the Operating Characteristics of a Group Sequential Bayesian Design.” Journal of Statistical Software, 69(11), 1-23. http://doi.org/10.18637/jss.v069.i11.

The R code of this vignette is accessible via

```
R> library("gsbDesign")
R> demo("usage")           # code of section 4
R> demo("PoC")             # code of section 5
```

1. Introduction

In traditional clinical trials, patients are randomized to a treatment or a control (e.g., placebo) arm. At the end of the trial the data are analyzed, comparing the two arms. Extensions of this setting are group sequential clinical trials (Jennison and Turnbull 1999). These adaptive designs have one or more interim analyses, where decisions are made on whether to stop or continue the trial. The advantage of a group sequential design is that futile trials can be stopped early, if the treatment is ineffective. In that case, useless treatment is avoided and money saved. On the other hand, studies can be stopped early for success, which may result in faster access to the new treatment.

The critical aspect of a group sequential design is the decision at each interim analysis on whether to stop or continue the trial. Because Bayesian approaches are particularly well suited to support decision-making, several authors proposed these for the monitoring of group sequential clinical trials; for a review and references see for example the books by Spiegelhalter, Abrams, and Myles (2004) and by Berry, Carlin, Lee, and Müller (2010).

The Bayesian framework also facilitates the incorporation of external information through informative priors. For example, historical trials often contain relevant information on the control arm that can be quantified by priors. Hence, fewer patients may then be randomized to the control arm, which reduces the cost and duration of the clinical trial (Pocock 1976; Neuenschwander, Capkun-Niggli, Branson, and Spiegelhalter 2010; Schmidli, Gsteiger, Roychoudhury, O'Hagan, Spiegelhalter, and Neuenschwander 2014). Furthermore, meta-analytic approaches can be used to include information on both the treatment and the control arms (Spiegelhalter *et al.* 2004; Schmidli, Wandel, and Neuenschwander 2013).

We consider group sequential Bayesian trial designs that incorporate decision making based on the posterior distribution of the difference between the treatment and the control arms. The posterior distribution contains the information from the ongoing clinical trial and the external information captured in the prior distribution.

In order to reflect medical decision-making, several stopping criteria based on this posterior distribution may be combined. Such a combination of multiple criteria goes beyond the significance testing framework of classical group sequential designs, and can, for example, include requirements on the observed effect size. The traditional sole focus on significance testing has also been criticized from a frequentist perspective (Armitage 1989; Kieser and Hauschke 2005; Carroll 2009; Chuang-Stein, Kirby, Hirsch, and Atkinson 2011b; Chuang-Stein, Kirby, French, Kowalski, Marshall, Smith, Bycott, and Beltangady 2011a).

The described approach is well suited to clinical trials conducted in the learning phases of drug development. Here, quantitative decision criteria based on the probability of achieving a clinically meaningful treatment effect may justify further investment in a novel compound (Cartwright, Cohen, Fleishaker, Madani, McLeod, Musser, and Williams 2010; Gsteiger, Neuenschwander, Mercier, and Schmidli 2013; Fisch, Jones, Jones, Kerman, Rosenkranz, and Schmidli 2015).

Once stopping criteria have been defined to correspond with clinical decision-making, it is important to evaluate the operating characteristics of the Bayesian group sequential design. To do so, some true effects of the treatment and control arms are assumed and the probability of stopping for success or futility, as well as the expected sample size, are calculated (Emerson, Kittelson, and Gillen 2007a,b).

In this article, we propose the R (R Core Team 2016) package **gsbDesign** (Gerber and Gsponer 2016) to evaluate the operating characteristics of such group sequential Bayesian designs available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=gsbDesign`. It supports designs with two arms, normal endpoints, and known standard deviations of the effects in the treatment and control arms. As shown by Spiegelhalter *et al.* (2004), this setting can be extended to many other types of outcome data by forming an appropriate approximate normalized likelihood; see Gsponer, Gerber, Bornkamp, Ohlssen, Vandemeulebroecke, and Schmidli (2014) for some examples. We consider the case where the number of patients per interim analysis is known at the beginning of the trial, although a more flexible approach has been proposed in a classical framework by Burington and Emerson (2003).

Several software packages exist that evaluate group sequential clinical designs, e.g., **S+SeqTrial** (Insightful Corporation 2002), **PEST** (The MPS Research Unit 2000), **ADDPLAN** (Wassmer and Eisebitt 2005), **East** (Cytel Software Corporation 2014), and **FACTS** (LLC Consultants 2014). However, to the best of our knowledge, **gsbDesign** is the only package that can both incorporate prior information and also allows the user to specify multiple decision criteria.

This article provides a detailed description on how to compute the operating characteristics for Bayesian group sequential designs, and how to use the **gsbDesign** package in practice. We structured the paper as follows: In Section 2, the general setup of Bayesian group sequential designs is presented. In Section 3, the evaluation of their operating characteristics is described, first for the case where prior information on the difference between treatment and control is available, and then for the case where prior information on the treatment and control arms is available. In Section 4, the use of the R package **gsbDesign** is explained in detail. Finally, in Section 5, a case study is presented, followed by a short conclusion.

2. General setup of the Bayesian design

Two-arm trials with zero, one, or more interim analyses are considered. At each analysis, the success and futility criteria are evaluated to decide if the trial should be stopped. The modeling framework assumes continuous outcome data with normally distributed errors. However, as shown by Spiegelhalter *et al.* (2004), by forming an appropriate approximate normalized likelihood, many other types of outcome data with corresponding sampling models (e.g., count data with an assumed Poisson distribution) can be approximated with this setup.

The criteria are based on the posterior distribution of the treatment effect δ, where δ is specified in terms of improvement over the control treatment (i.e., positive values are used to express the benefit of the experimental treatment over the control).

An arbitrary number of success and futility criteria can be specified at each analysis. The success criteria have the form $\mathsf{P}\{\delta > s \mid \text{data}\} \geq p$ and the futility criteria have the form $\mathsf{P}\{\delta < f \mid \text{data}\} \geq q$. Here, s and f are user-specified effect thresholds, p and q are user-specified probability thresholds.

Prior information can either be available for the treatment difference δ, or for the effect in the control arm, μ_1, and the treatment arm, μ_2. **gsbDesign** supports only normally distributed prior information. The variances of the prior distributions for control and treatment arms have the form σ_1^2/n_{10} and σ_2^2/n_{20}, respectively. Here, n_{10} and n_{20} correspond to the prior information in terms of number of patients in the control and the treatment arms, respectively.

It is assumed that single observations in the two arms have variances σ_1^2 and σ_2^2.
Thus, the full specification of the design requires:

- σ_k, $k = 1, 2$: standard deviations for control arm ($k = 1$) and treatment arm ($k = 2$);
- n_{k0}, $k = 1, 2$: quantification of prior information per arm;
- η_{k0}, $k = 1, 2$: prior expected response per arm;
- I: the number of interim analyses including final analysis;
- n_{ki}, $i = 1, \dots, I$: the added number of patients per arm and interim. Hence, the total number of patients in arm k at interim i is $N_{ki} = \sum_{j=1}^{i} n_{kj}$;
- s_{ir}, $i = 1, \dots, I$, $r = 1, \dots, R_{si}$: effect thresholds for each success criterion at each interim. R_{si} being the number of success criteria at interim i;
- p_{ir}, $i = 1, \dots, I$, $r = 1, \dots, R_{si}$: probability thresholds for each success criterion at each interim;
- f_{ir}, $i = 1, \dots, I$, $r = 1, \dots, R_{fi}$: effect thresholds for each futility criterion at each interim. R_{fi} being the number of futility criteria at interim i;
- q_{ir}, $i = 1, \dots, I$, $r = 1, \dots, R_{fi}$: probability thresholds for each futility criterion at each interim.

All R_{fi} futility criteria have to be fulfilled to stop for futility at an interim or the final analysis i. Similarly, all R_{si} success criteria have to be fulfilled to stop for success. If the trial is neither stopped for success nor for futility, the trial continues (unless the last analysis $i = I$ has been reached).

3. Operating characteristics

Given a set of scenarios for the true value of δ and a set of design parameters, the operating characteristics of main interest are the probabilities of success and futility at each interim analysis, and the expected sample size. For example, if the true treatment effect was small, we could examine whether the design would lead to a high probability of early stopping for futility. On the other hand, if the true treatment effect was large, we could examine whether the design would lead to a high probability of early stopping for success.

We consider simulation and numerical integration for computing the operating characteristics. The former simulates a large number of trials given some true treatment effects of interest. At each interim analysis, we compute the posterior distribution of the treatment effect given the data and evaluate the stopping criteria based on the trials not stopped at the previous interim analysis. The latter translates the criteria from the posterior distribution of the treatment effect to the distribution of the observed treatment effect. The precision related to the treatment effect estimates at each interim analysis can be calculated analytically (which is a consequence of assuming a known standard deviation associated with each treatment group in the design setup). Under the assumption of non-informative priors ($n_{k0} = 0$, $k = 1, 2$), this approach yields boundaries on the treatment effect scale that can be translated into a number of standard frequentist criteria such as conditional probabilities and p values. We then use the R package **gsDesign** (Anderson 2016) for numerical integration as described by Jennison and Turnbull (1999).

3.1. Prior information on the treatment effect

Let $Y_{kij} \sim N(\mu_k, \sigma_k^2)$ denote the observations for treatment $k = 1, 2$ at interim $i = 1, \ldots, I$ for subject $j = 1, \ldots, n_{ki}$.

The aggregated treatment effect at interim i is $D_i = \bar{Y}_{2i} - \bar{Y}_{1i}$ with $\bar{Y}_{ki} = (N_{ki})^{-1} \sum_{l=1}^{i} \sum_{j=1}^{n_l} Y_{klj}$ and $N_{ki} = n_{k1} + \cdots + n_{ki}$. Thus, $D_i \sim N(\delta, \sigma_1^2/N_{1i} + \sigma_2^2/N_{2i})$ with $\delta = \mu_2 - \mu_1$.

Assume prior information is available for the treatment effect: $\delta \sim N(\alpha_0, \sigma_1^2/n_{10} + \sigma_2^2/n_{20})$. This prior reflects information on the treatment effect as if n_{10} and n_{20} patients had been treated with the control and the test treatment, respectively.

For Bayesian updating, it is convenient to parametrize the normal distribution not in terms of variances but in terms of precisions. The precision is the inverse of the variance. The prior precision is denoted by $\beta_0 = n_{10}n_{20}/(n_{10}\sigma_2^2 + n_{20}\sigma_1^2)$ and the precision of the observed treatment effect at interim i is denoted by $B_i = N_{1i}N_{2i}/(N_{1i}\sigma_2^2 + N_{2i}\sigma_1^2)$. Normal distributions that are parametrized with the precision are denoted by $N_P(\cdot\,, \cdot)$.

The posterior is proportional to the likelihood times the prior. Here, the likelihood and the prior are $D_i \mid \delta \sim N_P(\delta, B_i)$ and $\delta \sim N_P(\alpha_0, \beta_0)$, respectively.

A normal likelihood with a normal prior leads to a conjugate analysis, and hence a normally distributed posterior. More precisely, the posterior expectation is a weighted average of the prior expectation and the sample mean, and the posterior precision is the sum of the prior and sample precisions. Thus, a sequential update yields the normal posterior distribution at interim i with expectation $\alpha_i = \omega_i\alpha_0 + (1 - \omega_i)D_i$ with $\omega_i = \beta_0/\beta_i$ and precision $\beta_i = \beta_0 + B_i$.

To characterize the distribution of D_i, we use the fact that the sequence $Z_1 = D_1\sqrt{B_1}, \ldots, Z_I = D_I\sqrt{B_I}$ is multivariate normal with $\mathsf{E}\{Z_i\} = \delta\sqrt{B_i}$, $i = 1, \ldots, I$ and $\mathsf{COV}\{Z_i, Z_j\} = \sqrt{B_i/B_j}$, $1 \leq i \leq j \leq I$. Jennison and Turnbull (1999) call this the canonical joint distribution for the parameter δ with information levels $B_1, \ldots, B_I$.

This formulation is convenient for both approaches, simulation and numerical integration.

Simulation

When evaluating the operating characteristics of a design, a range of true treatment effect values or scenarios, denoted by δ_u, $u = 1, \ldots, U$, is considered. In the case of the simulation approach, a complete set of interim treatment effects, D_i, $i = 1, \ldots, I$, is generated for a large number of trials (T_0) and each of the scenarios. To simulate the D_i, we use the canonical joint distribution for δ.

At each interim analysis, the posterior distribution is updated and the decision criteria are applied. The operating characteristics are then derived by computing the proportion of trials for which the success and/or futility criteria are fulfilled. Note that the denominator for the computation of the proportion is not the same at each interim, because, at interim $i + 1$, we only have to consider the trials that continued from the previous analysis i. Therefore, T_0 must be large enough to ensure that enough simulated trials are continued to the final analysis.

The simulation is summarized in the following pseudo-algorithm.

```
For a large $T_0$ and each $\delta_u$ do
    For each $i = 1, \dots, I$ do
        1. Simulate $D_i^{(t)}$, $t = 1, \dots, T_{i-1}$ with $T_{i-1}$ the number of trials not stopped at interim $i - 1$.
        2. Recursively compute the Bayesian update of the posterior distribution:
           $\beta_i = \beta_0 + B_i, \qquad \alpha_i^{(t)} = w_i \alpha_0 + (1 - w_i) D_i^{(t)}.$
        3. Compute $T_i^S$, the number of trials fulfilling all success criteria at interim $i$.
        4. Compute probability of success at stage $i$ as $T_i^S / T_{i-1}$.
        5. Compute $T_i^F$, the number of trials fulfilling all futility criteria at interim $i$.
        6. Compute probability of futility at stage $i$ as $T_i^F / T_{i-1}$.
        7. Set $T_i = T_{i-1} - T_i^S - T_i^F$.
    End loop for $i$.
End loop for $\delta_u$.
```

Numerical integration

The decision criteria are formulated in terms of the posterior distribution of the treatment effect δ. These criteria can be transformed and formulated in terms of the distribution of the observed treatment effects D_i.

The success criteria are fulfilled if $(s_{ir} - \alpha_i)\sqrt{\beta_i} \leq Q_N(1 - p_{ir})$, where $Q_N(\varepsilon)$ denotes the $\varepsilon \times 100\%$ quantile of a standard normal distribution. Solving for D_i yields the success criterion that r is fulfilled if $D_i \geq S_{ir} = \{s_{ir} - \omega_i \alpha_0 - \beta_i^{-1/2} Q_N(1 - p_{ir})\}/(1 - \omega_i)$. The trial will be stopped if all success criteria at interim analysis i are fulfilled, i.e., if $D_i \geq \max_r S_{ir} = S_i$.

Similarly, all futility criteria at interim analysis i are fulfilled if $D_i \leq \min_r F_{ir} = F_i$, where $F_{ir} = \{f_{ir} - \omega_i \alpha_0 - \beta_i^{-1/2} Q_N(q_{ir})\}/(1 - \omega_i)$.

We now have for each interim analysis a lower and an upper bound. If the observed treatment effect D_i at interim i is beyond these bounds, the trial is stopped. This situation is similar to the setting of classical group sequential designs, where at each interim a decision is taken to stop the trial if a certain standardized test statistic exceeds some threshold.

Thus, our group sequential Bayesian design yields a sequence of test statistics $\{D_1, \dots, D_I\}$, which is the same as for a classical group sequential design with different variances and unequal numbers of patients in the two treatment arms. Jennison and Turnbull (1999) provide efficient numerical integration techniques for computing probabilities of crossing thresholds based on the canonical joint distribution of δ with information levels $\{B_1, \dots, B_I\}$.

To derive the operating characteristics, we therefore need to compute $\mathsf{P}[\{(Z_i \geq u_i) \text{ or } (Z_i \leq l_i)\} \text{ and } l_j < Z_j < u_j \ \forall j < i]$, where $u_i = S_i\sqrt{B_i}$ and $l_i = F_i\sqrt{B_i}$. The function `gsProbability` from the R package **gsDesign** (Anderson 2016) implements the numerical integration techniques for computing these probabilities.

3.2. Prior information on both treatment arms

We consider the aggregated arm-wise treatment response at interim i, which is given by $\bar{Y}_{ki} = (N_{ki})^{-1} \sum_{l=1}^{i} \sum_{j=1}^{n_l} Y_{klj}$ and $N_{ki} = n_{k1} + \cdots + n_{ki}$.

Assume that there is prior information available for both the control and treatment arms: $\mu_k \sim N_P(\eta_{k0}, \gamma_{k0})$, with $\gamma_{k0} = n_{k0}/\sigma_k^2$. In this case, the prior to posterior updating is done per arm: $\mu_k \mid \bar{Y}_{ki} \sim N_P(\eta_{ki}, \gamma_{ki})$ with $\eta_{ki} = \omega_{ki}\eta_{k0} + (1 - \omega_{ki})\bar{Y}_{ik}$ and $\gamma_{ki} = \gamma_{k0} + N_{ki}/\sigma_k^2$.

The posterior for the treatment effect is then $\delta \mid \bar{Y}_{1i}, \bar{Y}_{2i} \sim N_P(\tilde{\alpha}_i, \tilde{\beta}_i)$, where $\tilde{\alpha}_i = \eta_{2i} - \eta_{1i}$ and $\tilde{\beta}_i = (1/\gamma_{1i} + 1/\gamma_{2i})^{-1}$.

When conducting the simulation approach, we generate the observed stagewise average treatment response, i.e., $\tilde{Y}_{ki} = (n_{ki})^{-1} \sum_{j=1}^{n_{ki}} Y_{kij}$, $i = 1, \ldots, I$, for a large number of trials (T_0), under a series of different true average treatment responses μ_{k0}, $k = 1, 2$. The aggregated arm-wise treatment response is then $(n_{ki}\tilde{Y}_{ki} + N_{k,i-1}\bar{Y}_{k,i-1})/(n_{ki} + N_{k,i-1})$.

At each interim analysis the posterior distribution is updated arm-wise and transformed to the treatment effect. The decision criteria are then applied to the posterior distribution of the treatment effect. The operating characteristics are derived by computing the proportion of trials that fulfill the success and/or futility criteria. Note that the denominator for the computation of the proportion is not the same at each interim, because at interim $i + 1$ we have to consider only the trials not stopped at interim i.

The simulation is summarized in the following pseudo-algorithm.

```
For a large $T_0$ and each plausible $\mu_{10}$ and $\mu_{20}$ do

    For each $i = 1, \ldots, I$ do
        1. Simulate $\tilde{Y}_{ki}^{(t)}$, $t = 1, \ldots, T_{i-1}$, $k = 1, 2$, with $T_{i-1}$ the number of trials not
           stopped at interim $i - 1$.
        2. Compute $\bar{Y}_{ki} = (n_{ki}\tilde{Y}_{ki} + N_{k,i-1}\bar{Y}_{k,i-1})/(n_{ki} + N_{k,i-1})$.
        3. Recursively compute the Bayesian update for the posterior distribution per
           arm:
               $\gamma_{ki} = \gamma_{k0} + N_{ki}/\sigma_k^2, \qquad \eta_{ki}^{(t)} = w_{ki}\eta_{k0} + (1 - w_{ki})\bar{Y}_{ki}^{(t)}.$
        4. Transform arm-wise posterior distributions to posterior distribution of treat-
           ment effect:
               $\tilde{\alpha}_i^{(t)} = \eta_{2i}^{(t)} - \eta_{1i}^{(t)}, \qquad \tilde{\beta}_i^{(t)} = (1/\gamma_{1i} + 1/\gamma_{2i})^{-1}.$
        5. Compute $T_i^S$, the number of trials fulfilling all success criteria at interim $i$.
        6. Compute probability of success at stage $i$ as $T_i^S/T_{i-1}$.
        7. Compute $T_i^F$, the number of trials fulfilling all futility criteria at interim $i$.
        8. Compute probability of futility at stage $i$ as $T_i^F/T_{i-1}$.
        9. Set $T_i = T_{i-1} - T_i^S - T_i^F$.
    End loop for $i$.

End loop for $\mu_{10}$ and $\mu_{20}$.
```

3.3. Expected sample size

The expected sample size in a group sequential design is computed as $\sum_{i=1}^{I}(n_{1i} + n_{2i})\pi_i$, where π_i denotes the probability of stopping at interim i. Once the probabilities of stopping for futility and stopping for success are available, the expected sample size is straightforward to compute.

4. Using gsbDesign

Here, we illustrate how to use the R package **gsbDesign**. After installation, the package can be loaded by

```
R> library("gsbDesign")
```

There are three main functions needed for the computation of the operating characteristics:

- `gsbDesign` fully specifies the design, i.e., all required parameters described in Section 2. The function returns an object of class '`gsbDesign`'.
- `gsbSimulation` specifies the methods for computing the operating characteristics, i.e., whether to use simulation or numerical integration, whether to update per arm or on the treatment effect. The function returns an object of class '`gsbSimulation`'.
- `gsb` calculates the operating characteristics and takes as arguments an object of class '`gsbDesign`' and an object of class '`gsbSimulation`'. The function returns an object of class '`gsbMainOut`'.

For objects of class '`gsbDesign`', '`gsbSimulation`', and '`gsbMainOut`', there exist `print` methods. For the class '`gsbMainOut`', there further exist `summary` and `plot` methods. More information on specific functions and methods are given in the reference manual of the package.

4.1. Specifying the design

The full specification of a group sequential Bayesian design requires the number of interim analyses (including final analysis), the standard deviation of individual observations per arm (σ_k), prior specification potentially per arm (n_{k0}), number of patients per arm and stage (n_{ki}), and success and futility criteria per stage (s_{ir}, p_{ir}, f_{ir}, q_{ir}).

The minimum requirement for the function `gsbDesign` is the specification for one stage. In this situation the specification will be the same in all later stages and the final analysis.

Prior information on treatment effect

The following code shows how to specify such a design with `nr.stages` $= 4$ analyses (interim plus final) when prior information is available for the treatment effect.

```
R> design1 <- gsbDesign(nr.stages = 4, patients = c(10, 20),
+    sigma = c(7, 7), criteria.success = c(0, 0.8, 7, 0.5),
+    criteria.futility = c(2, 0.8), prior.difference = c(3, 5, 2))
```

The argument `patients` specifies the number of patients (n_{ki}) per arm and stage. If `patients` is a single number, the same number of patients is used for all stages and both arms. If it is a vector of length 2, the first element of the vector gives the number of patients for the control arm in each stage and the second element gives the number of patients for the treatment arm in each stage. Finally, if the number of patients changes across stages, the argument `patients` must be a matrix with `nr.stages` rows and 2 columns.

The argument `sigma` specifies the standard deviations (σ_k) per arm. If `sigma` is a single number, the standard deviation is the same for both arms. If it is a vector of length 2 the first element of the vector gives the standard deviation for the control arm and the second element gives the standard deviation for the treatment arm.

In the example above, there are $n_{11} = 10$ patients in each of the stages in the control arm with standard deviation $\sigma_1 = 7$. In the treatment arm there are $n_{21} = 20$ patients in each stage with standard deviation $\sigma_2 = 7$.

The argument `criteria.success` specifies the success criteria in terms of the posterior distribution. The first two elements of the vector correspond to effect and probability thresholds for the first success criterion, and the second two elements to effect and probability thresholds for the second success criterion. In the example, the specification corresponds to $\mathsf{P}\{\delta > 0 \,|\, \text{data}\} \geq 0.8$ and $\mathsf{P}\{\delta > 7 \,|\, \text{data}\} \geq 0.5$. The success criteria are the same for all analyses.

Similarly, the argument `criteria.futility` specifies the futility criteria. In the example, there is only one futility criterion corresponding to $\mathsf{P}\{\delta < 2 \,|\, \text{data}\} \geq 0.8$. The futility criterion is the same for all analyses.

If success and/or futility criteria change with stages, the corresponding arguments must be matrices that have the same number of rows as there are stages in the design.

The argument `prior.difference` specifies the prior distribution and must be a vector of length 3. The first element gives the prior treatment effect. The second and third elements indicate the number of hypothetical patients in the control (n_{10}) and treatment (n_{20}) arms, respectively. The default is no prior information corresponding to $n_{10} = n_{20} = 0$, i.e., zero precision and is specified as `"non-informative"`.

The prior in the example can be interpreted as if $n_{10} = 5$ patients in the control and $n_{20} = 2$ patients in the treatment arm were added to the new trial, with an observed treatment difference of 3.

The object `design1` is of class '`gsbDesign`' and contains the following information.

```
R> names(design1)

[1] "nr.stages"        "patients"         "sigma"
[4] "criteria"         "prior.difference" "prior.control"
[7] "prior.treatment"
```

Prior information on both arms

The following code shows how to specify such a design with `nr.stages`$= 4$ analyses (interim plus final) when prior information is available for the treatment response in both arms.

In this case, the arguments `prior.control` and `prior.treatment` must be specified. Both arguments are vectors of length 2, where the first element is the arm specific effect and the second element is the number of hypothetical patients in each arm.

The other design specifications are identical to the previous design.

```
R> design2 <- gsbDesign(nr.stages = 4, patients = c(10, 20), sigma = c(7, 7),
+    criteria.success = c(0, 0.8, 7, 0.5), criteria.futility = c(2, 0.8),
+    prior.control = c(3, 5), prior.treatment = c(6, 2))
```

4.2. Specifying methods for the computation of operating characteristics

The methods for computing the operating characteristics depend on the availability of prior information. If prior information is available on the treatment effect, the numerical integration approach described in Section 3.1 is used by default. If prior information is available on both, the control and the treatment arm, the simulation approach described in Section 3.2 is used.

Prior information on treatment effect

For the `design1` defined above with prior information on the treatment effect, we can choose between the numerical integration and the simulation method to derive the operating characteristics. The numerical integration method, as well as an appropriate set of true treatment effects, are specified with the following command.

```
R> simulation1 <- gsbSimulation(truth = c(-10, 20, 60),
+    type.update = "treatment effect", method = "numerical integration")
```

The argument `truth` is a vector of length 3. The first two elements of this vector define the range of true treatment effects (δ_0) over which the operating characteristics are to be evaluated. The third element of the vector specifies the number of distinct values to consider.

The argument `type.update` indicates that the posterior updating is performed on the treatment effect.

The argument `method` indicates that numerical integration is used.

Alternatively, the operating characteristics can be computed based on the simulation approach described in Section 3.1. In this case, the function `gsbSimulation` takes a few additional arguments.

```
R> simulation1a <- gsbSimulation(truth = c(-10, 20, 60),
+    type.update = "treatment effect", method = "simulation",
+    nr.sim = 50000, warnings.sensitivity = 100, seed = "generate")
```

The argument `nr.sim` indicates the number of simulated trials to run.

With the simulation approach, the operating characteristics are computed as the number of successful/futile trials at a given stage divided by the number of trials not stopped yet. Hence, if `nr.sim` is not large enough, this leads to unstable results. Therefore, the argument `warnings.sensitivity` forces R to print a warning when the number of trials not stopped prior to a given stage is less than 100 in this example.

The argument `seed` ensures reproducibility. In the example, the argument specifies that the random seed is automatically generated.

To compare the results of numerical integration and simulation, the argument `method` can be set to `"both"`, in which case the subsequent call to `gsb` will generate results for both methods.

Prior information on both arms

For the `design2` defined above, an appropriate specification of the methods is achieved with the following command.

```
R> simulation2 <- gsbSimulation(truth = list(seq(-5, 5, 3), seq(0, 5, 3)),
+    type.update = "per arm", method = "simulation", grid.type = "table",
+    nr.sim = 10000, warnings.sensitivity = 500, seed = "generate")
```

Depending on the value of the argument `grid.type`, the argument `truth` has to be specified differently. If `grid.type = "table"` as in the example above, the argument `truth` must be a list with two elements, the first containing the true responses for the control arm (μ_{10}) and the second containing the true responses for the treatment arm (μ_{20}). Alternatively, if `grid.type = "manually"`, the argument `truth` must be a matrix with two columns and as many rows as true values. If `grid.type = "plot"`, the argument `truth` must be a vector of length 5. The first two elements give the range of true values for the control arm and the second two elements give the range of true values for the treatment arm. The fifth element indicates the number of grid points that are used to produce the graphic.

The argument `type.update` indicates that the posterior updating is performed per arm.

The argument `method` indicates that simulation is used to compute the operating characteristics.

4.3. Computing operating characteristics

Prior on treatment effect

The following command is used to compute the operating characteristics for `design1` via numerical integration.

```
R> oc1 <- gsb(design1, simulation1)
```

The object `oc1` is of class ‘`gsbMainOut`’ and contains the following elements.

```
R> names(oc1)
```

```
[1] "OC"          "boundary"    "design"      "simulation"
[5] "system.time"
```

The element `OC` contains the operating characteristics, i.e., probabilities of stopping for success and/or futility, and expected sample size. The other elements contain the decision boundaries on the D-scale, design, computational methods, and computing time.

The `summary`-method produces a concise summary of the operating characteristics.

```
R> summary(oc1, atDelta = c(0, 2, 7))

*** Group Sequential Bayesian Design ***

 Analysis N1 N2    S      F std.S  std.F
    Prior  5  2   NA     NA    NA     NA
        1 10 20 7.86 -0.729  2.90 -0.269
        2 10 20 7.43  0.195  3.88  0.102
        3 10 20 7.29  0.565  4.65  0.361
        4 10 20 7.21  0.775  5.32  0.572

sigma treatment: 7         sigma control: 7

stopping for success:
 delta stage 1 stage 2 stage 3 stage 4  total E{N}
     0  0.0019  0.0000  0.0000  0.0000 0.0020 68.1
     2  0.0157  0.0012  0.0001  0.0000 0.0170 97.6
     7  0.3764  0.1384  0.0745  0.0479 0.6372 75.4

stopping for futility:
 delta stage 1 stage 2 stage 3 stage 4  total
     0  0.3944  0.2095  0.1218  0.0777 0.8035
     2  0.1581  0.0848  0.0537  0.0383 0.3349
     7  0.0023  0.0001  0.0000  0.0000 0.0024
```

The argument `atDelta` allows the specification of values for δ_0, at which the operating characteristics are summarized. If the operating characteristics are not evaluated at the specified values, they are approximated by a linear interpolation of the nearest evaluated characteristics.

The first part of the summary output above summarizes the design. The columns `N1` and `N2` give the sample sizes of the prior and each stage, respectively. Columns `S` and `F` give success and futility boundaries for the observed treatment effect, respectively. Similarly, `std.S` and `std.F` give the standardized boundaries. If and only if the observed treatment effect is within `F` and `S` (or equivalently, the standardized treatment effect is within `std.S` and `std.F`), is the trial continued.

For example, if the observed treatment effect at the third interim analysis is between 7.29 and 0.56, the trial is continued. The corresponding standardized effect must be between 4.65 and 0.36 in order to continue the trial.

The second part summarizes the operating characteristics by providing the probabilities of success and futility, as well as the expected sample size, for each interim analysis and different true treatment effects (`delta`).

In the example above, the total probability of stopping for futility, if there is no treatment effect (`delta` $= 0$), is 80%. If the true treatment effect is 7, this probability is 0.24%. Similarly, the total probability of stopping for success if there is no treatment effect is 0.2% and if the true treatment effect is 7 this probability is 63.7%. The expected sample size is between 69 and 98.

The `plot` method allows one to display the operating characteristics graphically. The following code produces Figure 1, which shows the cumulative probabilities of success, futility, and an indeterminate decision. An indeterminate decision means that further information is needed to decide in favor of success or futility.

```
R> plot(oc1, what = "cumulative all")
```

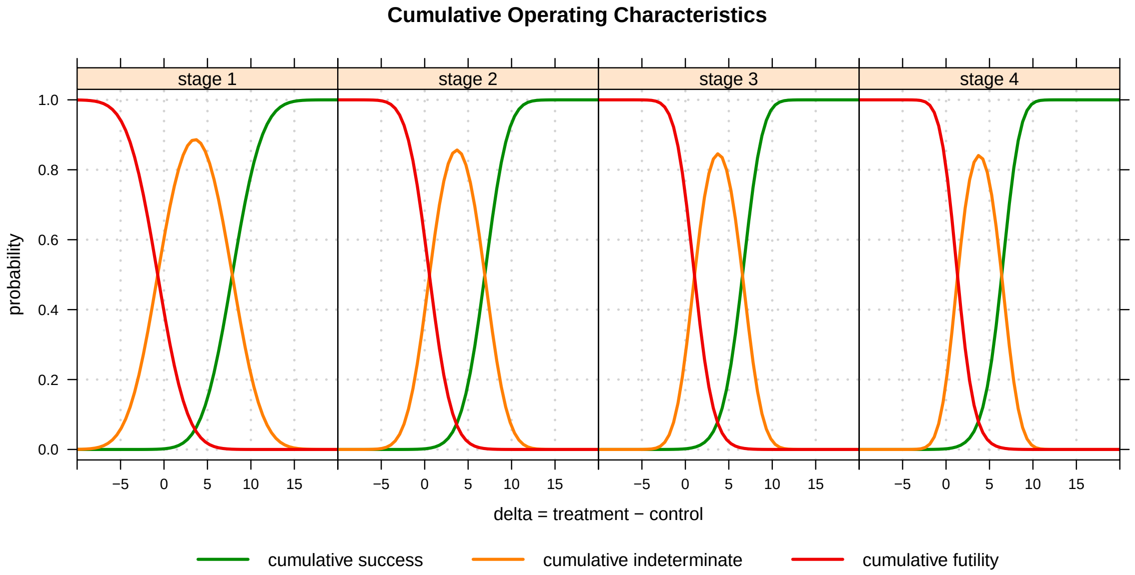

Figure 1: The operating characteristics are shown as cumulative probabilities of success, futility, and an indeterminate decision.

The argument `what` in the plot function in the previous code chunk can take the following values:

```
R> c("all", "cumulative all", "both", "cumulative both", "sample size",
+    "success", "futility", "success or futility", "indeterminate",
+    "cumulative success", "cumulative futility",
+    "cumulative success or futility", "cumulative indeterminate",
+    "boundary", "std.boundary", "delta.grid", "patients")
```

The following code produces the expected sample size and the decision boundaries that are depicted in Figure 2.

```
R> plot(oc1, what = "sample size")
R> plot(oc1, what = "boundary")
```

The function `tab` allows the extraction of operating characteristics from the `gsbMainOut` object in spreadsheet form.

```
R> tab(oc1, what = "cumulative success", atDelta = c(0, 2, 7), digits = 4,
+    export = FALSE)
```

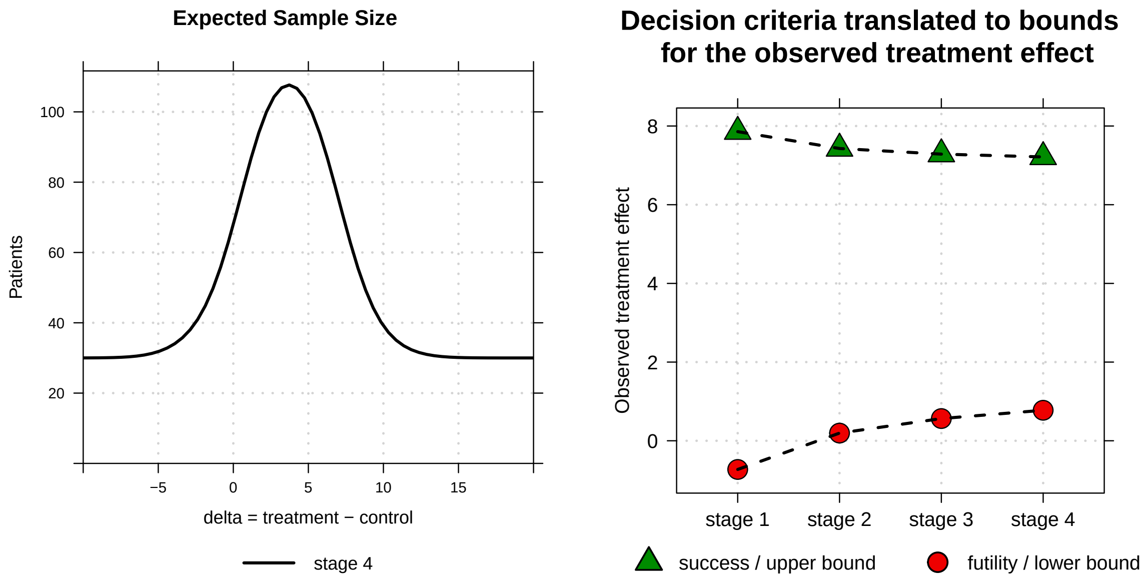

Figure 2: Illustration of the expected sample size (left panel) and the decision boundaries (right panel).

```
    stage 1 stage 2 stage 3 stage 4
1 0  0.0019  0.0020  0.0020  0.0020
2 2  0.0157  0.0169  0.0170  0.0170
3 7  0.3764  0.5148  0.5893  0.6372
```

The listing above shows the cumulative success probabilities at each interim analysis for true treatment effects of 0, 2, and 7. For example, for a true treatment effect of 7, the cumulative probability of success at interim analysis 4 is 63.72%.

The argument `what` of the `tab` function can take the following values:

```
R> c("all", "cumulative all", "success", "futility", "indeterminate",
+    "success or futility", "cumulative success", "cumulative futility",
+    "cumulative indeterminate", "cumulative success or futility",
+    "sample size")
```

With the argument `atDelta`, we can specify at which values of the true treatment effect the operating characteristics are reported. If the selected values are not part of the values specified for the computation, we interpolate linearly.

The argument `export` allows one to export the tables into a CSV file.

Prior on both treatment arms

For our second example, the code for computing the operating characteristics is the same.

```
R> oc2 <- gsb(design2, simulation2)
```

The same utilities as presented above can be used to summarize and display the operating characteristics in this situation. However, the complete output is somewhat long and not presented here.

To extract the expected sample size, we can use the following command.

```
R> tab(oc2, what = "sample size", digits = 0)

  control treatment delta stage1 stage2 stage3 stage4
1      -5         0     5     30     57     83    109
2      -2         0     2     30     51     68     83
3       1         0    -1     30     40     43     45
4       4         0    -4     30     32     33     33
5      -5         3     8     30     53     69     82
6      -2         3     5     30     58     84    110
7       1         3     2     30     54     74     93
8       4         3    -1     30     43     48     51
```

In this table, each row corresponds to one combination of true responses in the control and treatment arms. For example, the expected sample sizes for a true effect difference of `delta` = 2 is calculated for two combinations of true control and treatment arm values, see rows 2 and 7 in the table above. The corresponding expected numbers of patients in stage 4 are 83 and 93.

It is also possible to create graphical output in this situation. Because the operating characteristics depend now on both the true response in the treatment and the control arms, they are presented as contour plots. Figure 3 shows the cumulative probabilities of success or futility.

Further graphics and summaries can be produced when working directly on the data frame `oc2$OC`.

5. Case Study: Design of a PoC trial in Crohn's disease

Crohn's disease is an inflammatory bowel disease with diverse symptoms, mainly in the gastrointestinal tract. We consider the case where a new test treatment is believed to be potentially beneficial to patients with Crohn's disease; for more details see Gsponer *et al.* (2014). To investigate this, an initial small clinical trial with patients is planned. Such clinical trials are often called proof-of-concept (PoC) trials or pilot studies. If the PoC trial is successful, it is followed by larger clinical trials to explore more fully the efficacy and safety of the test treatment. The PoC study should be designed in such a way that, at its end, a decision on whether to continue or abandon further exploration of the test treatment can be made.

In the Crohn's disease case study, a parallel-group, double-blind, randomized clinical trial was planned. In such a design, patients are randomly allocated to two groups: one group receiving control treatment, and the other receiving the test treatment. Neither the patients nor their doctors know which of the two treatments they receive. After six weeks of treatment, the change in the disease activity from baseline (i.e., start of treatment) would be evaluated. To measure disease activity, the Crohn's disease activity index (CDAI) can be used; the

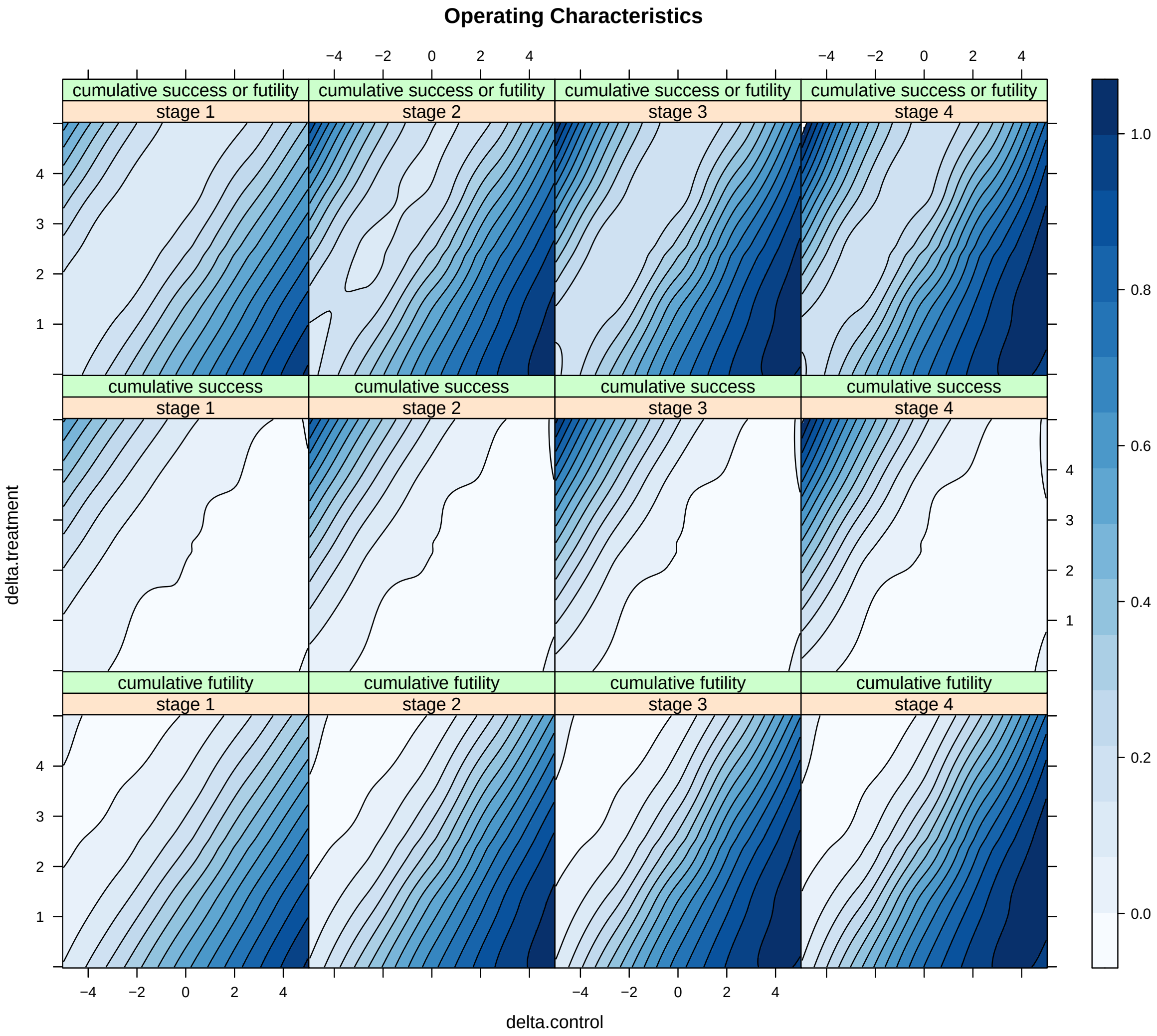

Figure 3: Contour plots of the operating characteristics when prior information is specified per arm. The cumulative probabilities of success or futility are shown.

CDAI is a score for which low values correspond to low activity. As the efficacy measure, the negative of the change from baseline to week six in the CDAI score is used, such that large values of this measure correspond to an improvement. The efficacy measure is approximately normally distributed with a standard deviation of about $\sigma = 88$, based on information from past studies with Crohn's disease patients. If n_1 patients have been allocated to the control, and n_2 patients to the test treatment, then the average efficacy measure in the two treatment groups is $\bar{Y}_1 \sim N(\mu_1, \sigma^2/n_1)$ and $\bar{Y}_2 \sim N(\mu_2, \sigma^2/n_2)$, respectively. The true treatment effect is then $\delta = \mu_2 - \mu_1$.

At this stage of the planning, one can quantify when a PoC trial can be considered a success. The PoC trial should first provide clear evidence that the test treatment is better than the control, and then give some indication that the test treatment is at least similarly effective as available treatments for Crohn's disease.

Available treatments for Crohn's disease have shown a difference to placebo in the efficacy measure of about 50 units. Hence, the results from the PoC trial would be considered positive by the clinician if the observed treatment effect is 50 units or more, and the treatment is very likely to be better than the placebo. To be competitive, the test treatment would actually have to be better by considerably more than 50 units.

In the following, we consider how the design of such a PoC trial may be developed, starting from a simple design with no interim analysis and no prior information, moving then to a design with one interim analysis (and no prior information), and finally also including prior information.

Simple design with no prior information and no interim analysis

The trial statistician, asked to come up quickly with a sample size for the PoC trial, may be tempted to formulate the requirements by the clinician as a conventional testing problem. For example, a one-sided test $H_0 : \delta \leq 0$ vs. $H_1 : \delta > 0$ with type I error of 5%. Then, to get a sample size, he might interpret the treatment difference of 50 units as the alternative and require a power of 80% at this alternative. For a design with no interim analysis, approximately 40 patients per treatment arm would be needed. In **gsbDesign** this design is specified with the following code.

```
R> desPoC1 <- gsbDesign(nr.stages = 1, patients = c(40, 40),
+    sigma = c(88, 88), criteria.success = c(0, 0.95),
+    criteria.futility = c(NA, NA))
R> simPoC1 <- gsbSimulation(truth = c(-50, 100, 60),
+    type.update = "treatment effect", method = "numerical integration")
R> ocPoC1 <- gsb(desPoC1, simPoC1)
R> summary(ocPoC1, atDelta = c(0, 50))
```

```
*** Group Sequential Bayesian Design ***

 Analysis N1 N2    S  F std.S std.F
    Prior  0  0   NA NA    NA    NA
        1 40 40 32.4 NA  1.64    NA

sigma treatment: 88         sigma control: 88

stopping for success:
 delta stage 1 E{N}
     0  0.0503   80
    50  0.8145   80

stopping for futility:
 delta stage 1
     0       0
    50       0
```

However, with this design, the result will be statistically significant when we observe a treatment effect of 32.4 units, which is much smaller than the 50 units observed in other studies. A

PoC study with a treatment effect of only 32.4 units, although statistically significantly better than the placebo at a one-sided significance level of 5%, would not be considered a success by clinicians, and the decision would probably be to discontinue further development. Hence, there is a need to formulate success criteria that better match the actual decision-making on whether to stop or continue further development.

Formally, these two success criteria can be expressed as:

$$\begin{aligned} &\text{(S1)} \quad \mathsf{P}\{\, \delta > 0 \mid \text{data} \,\} \geq 0.95, \\ &\text{(S2)} \quad \mathsf{P}\{\, \delta > 50 \mid \text{data} \,\} \geq 0.50. \end{aligned}$$

Here, $s_1 = 0$ and $s_2 = 50$, are the effect thresholds for the two success criteria, and the corresponding probability thresholds are $p_1 = 0.95$ and $p_2 = 0.50$.

Translated to a frequentist framework (Kieser and Hauschke 2005), these two criteria correspond approximately to the requirement that the test treatment is statistically significantly better than the placebo, and that the observed treatment difference is at least 50 units (when non-informative priors are used).

A futility criterion may also be specified to indicate when the test treatment is clearly inadequate. In this case, the futility criterion is expressed as:

$$\text{(F1)} \quad \mathsf{P}\{\, \delta < 40 \mid \text{data} \,\} \geq 0.90.$$

Here, $f_1 = 40$ and $q_1 = 0.9$ are the effect threshold and the probability threshold for the futility criterion. Again, these should reflect medical decision-making.

With these changed success and futility criteria, the trial statistician may reconsider the choice of an appropriate sample size. For example, he may choose the sample size such that the success criteria (S1) and (S2) are equivalent. Translating this to a frequentist framework, we may like to choose the sample size such that if the test treatment is statistically significantly better than the placebo, then the observed treatment difference is at least 50 units, and vice versa. Technically, this is achieved when the power is 50% at the observed difference of 50 units. Hence, a sample size of about 20 patients per group may be appropriate. It should be noted that the difference of 50 units is not the alternative hypothesis; see Neuenschwander, Rouyrre, Hollaender, Zuber, and Branson (2011) for a related discussion.

Using **gsbDesign**, the operating characteristics of the design with these success and futility criteria can now be evaluated.

```
R> desPoC2 <- gsbDesign(nr.stages = 1, patients = c(20, 20),
+    sigma = c(88, 88), criteria.success = c(0, 0.975, 50, 0.5),
+    criteria.futility = c(40, 0.9))
R> simPoC2 <- gsbSimulation(truth = c(0, 70, 60),
+    type.update = "treatment effect", method = "numerical integration")
R> ocPoC2 <- gsb(desPoC2, simPoC2)
R> summary(ocPoC2, atDelta = c(0, 40, 50, 60))
```

```
*** Group Sequential Bayesian Design ***

 Analysis N1 N2    S    F std.S std.F
```

```
    Prior  0  0   NA   NA    NA    NA
        1 20 20 54.5 4.34  1.96 0.156

sigma treatment: 88         sigma control: 88

stopping for success:
 delta stage 1 E{N}
     0  0.0250   40
    40  0.3007   40
    50  0.4352   40
    60  0.5777   40

stopping for futility:
 delta stage 1
     0  0.5619
    40  0.1000
    50  0.0504
    60  0.0228
```

The clinical team may think that the probability of success when the true treatment difference is 60 (a promising test treatment) is somewhat too low. Rather than now change the success criteria, the better option seems to be to expand the trial design to a two-stage design.

Design with no prior information and one interim analysis

Now consider a group sequential design with no prior information and one interim analysis. In both stages, we assign 20 patients to each treatment arm.

```
R> desPoC4 <- gsbDesign(nr.stages = 2, patients = c(20, 20),
+    sigma = c(88, 88), criteria.success = c(0, 0.975, 50, 0.5),
+    criteria.futility = c(40, 0.9))
R> simPoC4 <- gsbSimulation(truth = c(0, 70, 60),
+    type.update = "treatment effect", method = "numerical integration")
R> ocPoC4 <- gsb(desPoC4, simPoC4)
R> summary(ocPoC4, atDelta = c(0, 40, 50, 60, 70))
```

```
*** Group Sequential Bayesian Design ***

 Analysis N1 N2    S     F std.S std.F
    Prior  0  0   NA    NA    NA    NA
        1 20 20 54.5  4.34  1.96 0.156
        2 20 20 50.0 14.78  2.54 0.751

sigma treatment: 88         sigma control: 88

stopping for success:
 delta stage 1 stage 2  total E{N}
```

```
     0  0.0250  0.0026 0.0276 56.5
    40  0.3007  0.1102 0.4109 64.0
    50  0.4352  0.1582 0.5934 60.6
    60  0.5777  0.1828 0.7605 56.0
    70  0.7107  0.1718 0.8825 51.2

stopping for futility:
 delta stage 1 stage 2  total
     0  0.5619  0.2447 0.8066
    40  0.1000  0.0517 0.1518
    50  0.0504  0.0200 0.0704
    60  0.0228  0.0061 0.0288
    70  0.0091  0.0015 0.0106
```

With this design, if the treatment is placebo-like, there is a 2.8% probability of declaring the PoC successful and an 80.7% probability of declaring it futile. If the true treatment effect is $\delta = 60$, the success and futility probabilities are 76% and 2.9% respectively. The expected sample size varies between 52 and 64 patients.

Design with prior information for placebo and one interim analysis

An informative prior for the true treatment effect (η_{10}) in the placebo group was derived from six historical trials in patients with Crohn's disease, using a meta-analytic-predictive approach (Neuenschwander *et al.* 2010). More precisely, $\mu_1 \sim N(49, \sigma^2/20)$ was used as prior information; for details, see Gsponer *et al.* (2014). Thus, prior information on the placebo is worth 20 patients.

```
R> desPoC5 <- gsbDesign(nr.stages = 2, patients = c(10, 20),
+    sigma = c(88, 88), criteria.success = c(0, 0.975, 50, 0.5),
+    criteria.futility = c(40, 0.9), prior.control = c(49, 20))
R> simPoC5 <- gsbSimulation(truth = cbind(rep(c(30, 50, 70), each = 5),
+    c(30, 70, 80, 90, 100, 50, 90, 100, 110, 120, 70, 110, 120, 130, 140)),
+    nr.sim = 20000, type.update = "per arm", method = "simulation",
+    grid.type = "manually")
R> ocPoC5 <- gsb(desPoC5, simPoC5)
```

The resulting operating characteristics of this simulation are summarized in Table 1. If the test treatment is placebo-like, then the PoC will be declared to be successful in only 1.3% of the cases, i.e., the type I error is low. If the experimental treatment is borderline effective ($\delta = 50$) or similar to competitors ($\delta = 60$), then a successful PoC is expected in 65% and 82% of cases, respectively. The expected sample size is typically between 36 and 49, depending on the true effect size.

6. Conclusion

In this paper, we have presented the R package **gsbDesign**. The package provides utilities to study operating characteristics of group sequential Bayesian designs. When prior information

δ	Interim success	Interim futility	Final success	Final futility	Expected N
0	0.012	0.627	0.013	0.840	41
40	0.333	0.062	0.423	0.100	49
50	0.514	0.024	0.646	0.036	44
60	0.690	0.007	0.820	0.009	40
70	0.830	0.002	0.931	0.002	36

Table 1: Operating characteristics of the two stage design.

is available on the treatment effect, the package uses the efficient numerical integration methods from Jennison and Turnbull (1999) that are implemented in the R package **gsDesign**. If the amount of information is not the same for the treatment and control groups, **gsbDesign** uses simulation to compute the operating characteristics.

Acknowledgments

We would like to thank Björn Bornkamp, David Ohlssen, and Heinz Schmidli for their helpful and constructive comments and suggestions on the draft versions of the manuscript. We would also like to thank the two reviewers for their thoughtful comments that helped to improve the manuscript. We acknowledge support from the Swiss National Science Foundation (SNSF-144973).

References

Anderson K (2016). ***gsDesign**: Group Sequential Design.* R package version 3.0-1, URL http://CRAN.R-project.org/package=gsDesign.

Armitage P (1989). "Inference and Decision in Clinical Trials." *Journal of Clinical Epidemiology*, **42**(4), 293–299. doi:10.1016/0895-4356(89)90033-4.

Berry S, Carlin B, Lee J, Müller P (2010). *Bayesian Adaptive Methods for Clinical Trials.* Chapman & Hall/CRC. doi:10.1201/ebk1439825488.

Burington B, Emerson S (2003). "Flexible Implementations of Group Sequential Stopping Rules Using Constrained Boundaries." *Biometrics*, **59**(4), 770–777. doi:10.1111/j.0006-341X.2003.00090.x.

Carroll K (2009). "Back to Basics: Explaining Sample Size in Outcome Trials, are Statisticians Doing a Thorough Job?" *Pharmaceutical Statistics*, **8**(4), 333–345. doi:10.1002/pst.362.

Cartwright M, Cohen S, Fleishaker J, Madani S, McLeod J, Musser B, Williams S (2010). "Proof of Concept: A PhRMA Position Paper With Recommendations for Best Practice." *Clinical Pharmacology & Therapeutics*, **87**(3), 278–285. doi:10.1038/clpt.2009.286.

Chuang-Stein C, Kirby S, French J, Kowalski K, Marshall S, Smith M, Bycott P, Beltangady M (2011a). "A Quantitative Approach for Making Go/No-Go Decisions in Drug Development." *Drug Information Journal*, **45**(2), 187–202. doi:10.1177/009286151104500213.

Chuang-Stein C, Kirby S, Hirsch I, Atkinson G (2011b). "The Role of the Minimum Clinically Important Difference and Its Impact on Designing a Trial." *Pharmaceutical Statistics*, **10**(3), 250–256. doi:10.1002/pst.459.

Cytel Software Corporation (2014). ***East**. Software for Design Simulation and Interim Monitoring of Flexible Clinical Trials.* Cambridge, MA. URL http://www.cytel.com/software-solutions/east/.

Emerson S, Kittelson J, Gillen D (2007a). "Bayesian Evaluation of Group Sequential Clinical Trial Designs." *Statistics in Medicine*, **26**(7), 1431–1449. doi:10.1002/sim.2640.

Emerson S, Kittelson J, Gillen D (2007b). "Frequentist Evaluation of Group Sequential Clinical Trial Designs." *Statistics in Medicine*, **26**(28), 5047–5080. doi:10.1002/sim.2901.

Fisch R, Jones I, Jones J, Kerman J, Rosenkranz G, Schmidli H (2015). "Bayesian Design of Proof-of-Concept Trials." *Therapeutic Innovation & Regulatory Science*, **49**(1), 155–162. doi:10.1177/2168479014533970.

Gerber F, Gsponer T (2016). ***gsbDesign**: Group Sequential Bayes Design.* R package version 1.00, URL http://CRAN.R-project.org/package=gsbDesign.

Gsponer T, Gerber F, Bornkamp B, Ohlssen D, Vandemeulebroecke M, Schmidli H (2014). "A Practical Guide to Bayesian Group Sequential Designs." *Pharmaceutical Statistics*, **13**(1), 71–80. doi:10.1002/pst.1593.

Gsteiger S, Neuenschwander B, Mercier F, Schmidli H (2013). "Using Historical Control Information for the Design and Analysis of Clinical Trials with Overdispersed Count Data." *Statistics in Medicine*, **32**(21), 3609–3622. doi:10.1002/sim.5851.

Insightful Corporation (2002). ***S+SeqTrial 2**: User's Manual.* Seattle, WA. URL http://www.rctdesign.org/Software.html/.

Jennison C, Turnbull B (1999). *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC. doi:10.1201/9781584888581.

Kieser M, Hauschke D (2005). "Assessment of Clinical Relevance by Considering Point Estimates and Associated Confidence Intervals." *Pharmaceutical Statistics*, **4**(2), 101–107. doi:10.1002/pst.161.

LLC Consultants (2014). ***FACTS**: Fixed and Adaptive Clinical Trial Simulator.* Austin, TC. URL http://www.berryconsultants.com/software/.

Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter D (2010). "Summarizing Historical Information on Controls in Clinical Trials." *Clinical Trials*, **7**(1), 5–18. doi:10.1177/1740774509356002.

Neuenschwander B, Rouyrre N, Hollaender N, Zuber E, Branson M (2011). "A Proof of Concept Phase II Non-Inferiority Criterion." *Statistics in Medicine*, **30**(13), 1618–1627. doi:10.1002/sim.3997.

Pocock S (1976). "The Combination of Randomized and Historical Controls in Clinical Trials." *Journal of Chronic Diseases*, **29**(3), 175–188. doi:10.1016/0021-9681(76)90044-8.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B (2014). "Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information." *Biometrics*, **70**(4), 1023–1032. doi:10.1111/biom.12242.

Schmidli H, Wandel S, Neuenschwander B (2013). "The Network Meta-Analytic-Predictive Approach to Non-Inferiority Trials." *Statistical Methods in Medical Research*, **22**(2), 219–240. doi:10.1177/0962280211432512.

Spiegelhalter D, Abrams K, Myles J (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Statistics in Practice. John Wiley & Sons. doi:10.1002/0470092602.

The MPS Research Unit (2000). ***PEST**: Planning and Evaluation of Sequential Trials*. The University of Reading. URL http://www.mps-research.com/PEST/.

Wassmer G, Eisebitt R (2005). ***ADDPLAN**: Adaptive Designs – Plans and Analyses*. Reston, VA. URL http://www.addplan.com/.

Affiliation:

Florian Gerber
Institute of Mathematics
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich, Switzerland
E-mail: florian.gerber@math.uzh.ch